Medication Adherence Classification for Non-Communicable Disease Patients through Machine Learning Approaches

by

Wellington Kanyongo

Student Number: 222056902

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy *in* Computer Science

at the

School of Mathematics, Statistics, and Computer Science

University of KwaZulu-Natal

Durban, South Africa

June 2024

UNIVERSITY OF KWAZULU-NATAL

COLLEGE OF AGRICULTURE, ENGINEERING AND SCIENCE

DECLARATION

The research described in this thesis was performed at the University of KwaZulu-Natal under the supervision of Professor Absalom El-Shamir Ezugwu (main supervisor) and Dr. Jean Vincent Fonou Dombeu (co-supervisor). I hereby declare that all materials incorporated in this thesis are my original work except where acknowledgment is made by name or in the form of a reference. The work contained herein has not been submitted in part or whole for a degree at any other university.

Signed: [redacted]

Wellington Kanyongo

Date: May 2024

This is to confirm that this work was done under my supervision and that it is the candidate's original work. As the supervisor of this candidate, I have approved this thesis for submission.

Signed: [redacted]

Prof. Absalom Ezugwu

Date: May 2024

Signed: [redacted]

Dr. Jean Vincent Fonou Dombeu

Date: 27 June 2024

UNIVERSITY OF KWAZULU-NATAL

COLLEGE OF AGRICULTURE, ENGINEERING AND SCIENCE

DECLARATION – PLAGIARISM

I, _____ Wellington Kanyongo _____declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other University.
3. This thesis does not contain other persons' data, pictures, graphs, or other information unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain other persons' writing unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
    a. Their words have been re-written but the general information attributed to them has been referenced
    b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks and referenced.
5. This thesis does not contain text, graphics, or tables copied and pasted from the Internet, unless specifically acknowledged, and the source is detailed in the thesis and the References sections.

Signed:

.....

Wellington Kanyongo

Date: May 2024

DEDICATION

First and foremost, I devote this thesis to the most-high, omniscient, all-powerful, omnipresent, and everlasting God, the source of my life, salvation, strength, wisdom, sustenance, security, and success. I am dedicating my research to my wife, daughters, sons, family, and parents. I am highly indebted to them for their physical, financial, emotional, spiritual, and moral support.

ACKNOWLEDGEMENT

Above all, I express profound appreciation to the omnipotent and everlasting God for bestowing upon me His redemptive favour, the gift of life, and the invaluable opportunity to complete my Ph.D. studies. I am filled with reverence for you, Lord, because of your boundless wisdom and supreme deeds over the years.

The culmination of this thesis required top-notch input from a number of individuals. First and foremost, I am grateful to my supervisors, Professor Absalom El-Shamir Ezugwu and Dr. J.V.F. Dombeu, whose experience and wise guidance, motivation, and direction made a significant difference in the achievement of this thesis. I appreciate the sacrifices you have made to ensure the effective culmination of this programme. Prof. Absalom Ezugwu, I appreciate your persistent support, care, mentorship, prompt feedback, constructive criticism, and feedback. Your efforts have contributed to the exceptional quality of this research and have helped assure the successful completion of my PhD programme. I couldn't have asked for a better supervisor. I would like to express my gratitude, sir, for the privilege of being your student. You have shown faith in me and helped me reach my highest potential.

Secondly, I am also grateful to the Cimas Medical Aid Society and Dr. Tsitsi Moyo for orchestrating and providing me with the necessary research data. Despite their arduous work schedules, they created an inviting atmosphere to facilitate the data extraction process. Again, I am particularly grateful to the School of Mathematics, Statistics, and Computer Science, along with the administrator, Christel Barnard, for their invaluable contributions in achieving this significant milestone. I express my heartfelt thanks to the College of Engineering, Agriculture, and Science at the University of KwaZulu-Natal for granting me fee remission during my academic journey and creating a favourable atmosphere for conducting research. Additionally, I would like to express my sincere appreciation to the IT support staff, administrative staff at the UKZN, notably Jothimala Manickum and her coworkers, and higher-degree academic administrative officers for their tremendous support.

My wife Faith, daughters Maka and Aleena, boys Behold and Nolan, siblings, family, and parents have all provided me with joy, strength, moral, emotional, and spiritual support. I will remain eternally grateful to them for standing with me, encouraging me, cheering me on, and for their unwavering company and support that has always added value, purpose, and success to my academic journey.

ABSTRACT

Medication non-adherence is a significant public health issue that results in poor treatment outcomes, unnecessary hospitalisations, increased healthcare expenses, and increased risks of morbidity, disability, and mortality. According to the World Health Organisation, only 50% of people with chronic diseases adhere to treatment recommendations, despite receiving guidance on the need for medication adherence (MA). Notwithstanding considerable endeavours to tackle this concern, traditional methods have demonstrated low efficacy in predicting and intervening in non-adherent behaviours. In recent years, the advent of Fourth Industrial Revolution (4IR) technologies, such as artificial intelligence (AI) and machine learning (ML) has brought about promising solutions to MA problems. This study developed seven ML models (SVM, KNN, DT, Naïve Bayes, DNN, LR, and RF) to classify MA in a dataset of diabetes and hypertension patients in Zimbabwe, a developing country. The thesis's full structure includes numerous studies chronologically arrayed as follows: (1) an expansive systematic literature review on ML approaches to MA analytics among NCD (NCD) patients; (2) feature selection and the importance of predictors of NCDs MA from ML research perspectives; (3) the generation of a novel state-of-the-art dataset entitled "Data Wrangling and Generation for ML Models in MA Analytics: A Practical Standpoint Using Patient-Level and Medical Claims Data"; (4) development and validation of seven ML models, and (5) development of twinned machine learning algorithms for MA analytics in patients with diabetes and hypertension. The relevant data for diabetics and hypertensives were filtered from the initial raw dataset based on the International Classification of Diseases 10th Revision Code (ICD-10). Since adherence metrics measure the percentage of patients covered by medication refill claims for the same medication in the same therapeutic class, the created dataset only included medication refill data and patient-level data for 2022, from January 1 to December 31, 2022. Non-adherence was defined as medication refills that accounted for less than 75% of the projected 12-month claims, whereas adherence required at least 75% of the refills. This study acknowledges the absence of a universally accepted compendium of data wrangling (DW) steps while developing the novel dataset. However, it accepts certain building blocks and functionalities with significant overlaps as typical DW elements. In light of this, the study employs a pragmatic approach to dataset generation, encompassing nine essential DW building blocks and tasks that are employed iteratively and incrementally, comparable to an agile approach to dataset generation. An ML model was then developed using the generated dataset to evaluate the novel dataset's soundness, and its 81% accuracy demonstrated a generally good level of performance. During the ML model development process, the study used an RF feature importance mechanism to identify the important variables needed for building MA prediction models. The study also produced Shapley Additive Explanations (SHAP) beeswarm plots to aid in the interpretation of the model outputs and evaluate the contribution of the features to the prediction process. These factors included the annual quantity of medical supplies, the annual claim amount, the patient's

age, their membership in a wellness programme, the medical aid cover, their contribution to the cover, comorbidity, diagnosis, the type of hospital cover, the occurrence of complications, gender, and the type of medical aid scheme. The total number of medical supplies dispensed annually was revealed as the most important predictor of MA. The ML classifiers had a classification accuracy of 84.7% to 87.6%, with AUC values ranging from 0.857 to 0.934. The most robust classifier was the RF, an ensemble learning technique with 87.6% accuracy, an AUC of 0.9351, and superior precision, recall, and F1-score. The hybridization of classical SVM with tree-based ML algorithms (RF and DT) demonstrated improved results, with the newly developed TWINNED-SVM-RF and TWINNED-SVM-DT classifiers outperforming the standard SVM. Among these, the TWINNED-SVM-RF model showed superior performance. Additionally, the application of LIME (Local Interpretable Model-Agnostic Explanations) enhanced the explainability of the models, providing valuable insights into their decision-making processes. The ensemble learner (RF) and the TWINNED-SVM-RF show promise as prognostic tools for improving MA and determining patient adherence levels. These findings contribute to reducing discrepancies in medication refills and adherence rates among NCD patients. The ML model serves as a foundation for the creation of intelligent MA and intervention apps to promote patient MA among chronic disease patients. The outcomes, techniques, and insights proffered in novel dataset generation inform future researchers, data scientists, and analysts on how to conduct data mining, dataset enrichment, and DW. The dataset generated can spark additional research in areas such as evaluating MA and performing ML tasks such as classifying, predicting, and clustering MA behaviours.

LIST OF INCLUDED ARTICLES

1. Article in Peer-reviewed Journal (ISI)

a. Kanyongo, W., Ezugwu, A. E. S., Moyo, T., & Dombeu, J. V. F. (2024). Data Wrangling and Generation for Machine Learning Models in Medication Adherence Analytics: A practical Standpoint using Patient-Level and Medical Claims *Data. Data Intelligence*, 1-42

b. Kanyongo, W., & Ezugwu, A. E. (2023). Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives. *Informatics in Medicine Unlocked*, 101232.

c. Kanyongo, W., & Ezugwu, A. E. (2023). Machine learning approaches to medication adherence amongst NCD patients: A systematic literature review. *Informatics in Medicine Unlocked*, *38*, 101210.

2. Articles under Review for Peer-reviewed Journals (ISI)

a. Kanyongo, Moyo, T., W., Ezugwu, A.E., Dombeu, J.V.F. (2024.) Machine Learning Classification of Medication Adherence in NCD Patients: A Retrospective Study. *Scientific Reports*.

PUBLISHED DATASET

3. Dataset published in Mendeley Data

a. Kanyongo, W., Moyo, T., Ezugwu, A.E., Dombeu, J.V.F. (2024). Dataset for analysing medication adherence among diabetes and hypertension patients: Patient-level and medication refill data. *Mendeley Data.* https://doi.org/10.17632/ZKP7SBBX64.2

4. Articles under Review in Peer-reviewed Conference Proceedings

a. Kanyongo, W. and Ezugwu, A.E. (2024). On a twin machine learning algorithms for medication adherence analytics in patients with diabetes and hypertension. In: 12th International Conference on Big Data and Artificial Intelligence (BDA 2024) December 17-20, 2024, Hyderabad, India.

TABLE OF CONTENTS

DECLARATION – PLAGIARISM ii
DEDICATION iii
ACKNOWLEDGEMENT iv
ABSTRACT vi
LIST OF INCLUDED ARTICLES viii
PUBLISHED DATASET viii
GLOSSARY xi
CHAPTER 1: 1
INTRODUCTION 1
1.1 Introduction 1
1.2 Background to the Study 2
1.3 Research Motivation 8
1.4 Aim and Objectives 11
1.5 Scope of the Thesis 12
1.6 Research Contributions 13
1.7 Thesis Layout 15
CHAPTER 2 18
Systematic literature review on ML approaches to medication adherence analytics (Paper 1) 18
2.1 Introduction 18
2.2 Chapter Summary 52
CHAPTER 3 53
Systematic literature review on feature selection and importance of predictors of medication adherence from ML research perspectives 53
(Paper 2) 53
3.1 Introduction 53
3.2 Chapter Summary 87
CHAPTER 4 88
Data wrangling and generation for ML models in medication adherence analytics 88
(Paper 3) 88
4.1 Introduction 88
4.2 Chapter Summary 130
CHAPTER 5 131
Development and validation of ML models for medication adherence analytics (Paper 4) 131

5.1 Introduction 131
5.2 Chapter Summary 167
CHAPTER 6 168
Development and validation of a twin machine learning algorithm for medication adherence analytics in patients with diabetes and hypertension 168
(Paper 5) 168
6.1 Introduction 168
6.2 Chapter Summary 182
CHAPTER 7: 183
SUMMARY, CONCLUSIONS AND RECOMMENDATIONS 183
7.1 Introduction 183
7.2 General Summary 183
7.2.1 Summary of findings related to Review on ML Approaches to Medication Adherence Analytics 185
7.2.2 Summary of findings related to feature selection and importance of predictors of medication adherence from ML research perspectives 185
7.2.3 Summary of results related to data wrangling and generation for ML models in medication adherence analytics 186
7.2.4 Summary of results related to development and validation of ML models for medication adherence analytics 187
7.3 Implications for Practice 188
7.4 Implications for Policy 190
7.5 Theoretical Implications 190
7.6 Future Research Directions 192
References 194
Appendices 200
Appendix A: Letter of Approval for Data Collection 200

GLOSSARY

A

Accuracy: It quantifies the proportion of accurately predicted instances (including both true positives and true negatives) relative to the total number of instances in the dataset.

AI (Artificial Intelligence): The technology that allows computers and machines to mimic human intelligence and problem-solving abilities.

Algorithm: An algorithm is a systematic and well-defined set of instructions used to solve a problem or carry out a computation.

AUC (Area under the ROC Curve): A prescribed sequence of instructions that a computer must adhere to to carry out calculations or solve problems.

B

BPNN (Back-Propagation Neural Networks): Backpropagation is the method of modifying the weights of a neural network by evaluating the error rate from the previous iteration.

C

Cimas Medical Aid Society: A Medical Service Provider that offers modern health solutions for members, including Medical Aid, International Travel Insurance, Primary Healthcare, and Laboratory Services.

Confusion Matrix: A visualization method used to assess classification model performance in machine learning by comparing predicted values against actual values for a dataset.

Cross-validation: Cross-validation is a method used to assess the effectiveness of a model by training it on a portion of the input data and then testing it on a previously unseen subset of the input data.

CVD (Cardiovascular disease): A range of problems affecting the heart and blood arteries, such as coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other related conditions.

D

Data: Observations or measurements that are utilised to train a machine-learning model.

Dataset: A collection of data related to a particular topic, theme, or industry that is used to train and test algorithms and models.

Diabetes: A chronic, metabolic disorder marked by high levels of blood glucose (or blood sugar), resulting in long-term harm to the heart, blood vessels, eyes, kidneys, and nerves.

DL (Deep Learning): A branch of machine learning that employs deep neural networks, which are multi-layered neural networks, to replicate the intricate decision-making capabilities of the human brain.

DNN (Deep Neural Network): A subset of machine learning that uses multi-layered neural networks, also known as deep neural networks with a specific level of complexity, or neural networks with more than two layers, to mimic the human brain's intricate decision-making power.

DT (Decision Tree): A supervised learning algorithm that models decisions based on input features. It is a tree-structured classifier, with internal nodes representing dataset attributes, branches representing decision rules, and leaf nodes representing outcomes.

DW (Data Wrangling): The process of transforming and mapping data from one "raw" data type to another in order to make it more relevant and valuable for downstream applications such as analytics.

E

EHR (Electronic Health Records): An electronic version of a patient's medical history, that is maintained by the provider over time, and may include all of the key administrative clinical data relevant to that person's care under a particular provider, including

demographics, progress notes, problems, medications, vital signs, and any past medical history.

EMR (Electronic Medical Records): An electronic (digital) collection of medical information on a person saved on a computer. An electronic medical record contains information about a patient's health history, including diagnosis, medications, tests, and treatment plans.

F

F1-score: The harmonic mean of the precision and recall of a classification model.

G

GB (Gradient boosting): An ensemble machine learning technique that builds a strong predictive model by iteratively combining numerous weak models to enhance the overall prediction accuracy. During training, it employs a gradient descent-like method to reduce errors.

H

Hypertension: is also high blood pressure, is a medical disorder characterised by consistently elevated pressure in the blood arteries.

K

K-NN (K-Nearest Neighbor): A non-parametric, supervised learning classifier that uses proximity to classify or predict the grouping of an individual data point.

L

LR (Logistic Regression): A supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation.

M

MA (Medication Adherence): The extent to which a patient's actions regarding medications align with the prescribed medical instructions or guidance provided by a healthcare professional.

Medical Claim: An invoice or bill that your doctor's office submits to your health insurance company following your receipt of medical services.

Medication Refill: A subsequent dispensing of the medicine ordered by the practitioner in the original prescription order, based upon the practitioner's authorization for the subsequent dispensing in that original prescription order.

ML (Machine Leaning): A branch of AI that specifically deals with the development of computer systems capable of learning from data.

Machine Leaning Algorithm: A collection of rules or procedures employed by an artificial intelligence system to perform tasks, typically including the recognition of new data patterns and insights, or the prediction of output values based on a given set of input variables.

Machine Leaning Model: Computational algorithms utilised to identify patterns within data or generate predictions.

MRP (Medication Possession Ratio): A metric computed by dividing the number of days a patient has enough medication in stock by the number of days they should have enough medication in stock.

N

Naïve Bayes: A probabilistic machine learning technique that relies on the Bayes Theorem, commonly employed in a diverse range of classification tasks.

NCD (Non-Communicable Disease): A medical condition that cannot be transmitted directly from one individual to another.

P

PDC (Proportion of Days Covered): The proportion (or percentage) of days that an individual has access to medication within a particular observation period, based on the fill dates and days' supply for each dispensing.

Precision: A metric of a machine learning model's performance—the quality of a positive prediction generated by the model.

R

Recall: A metric that indicates how often a machine learning model accurately detects positive examples (true positives) among all of the actual positive samples in the dataset.

RFE (Recursive Feature Elimination): Recursive Feature Elimination is an iterative technique for selecting features in a dataset, which operates by systematically deleting features from the dataset and assessing the effectiveness of a machine learning model at each iteration.

RF (Random Forest): An ensemble learning technique used for classification, regression, and other tasks involving creating several decision trees during the training process.

S

SHAP (Shapley Additive Explanations): a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions

Support Vector Machines: A supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximises the separation between different classes in a space with multiple dimensions.

T

TDM (Therapeutic Drug Monitoring): the medical procedure of regularly assessing certain medications in a patient's circulation to ensure a consistent concentration, therefore maximising the effectiveness of a unique dosing regimen.

TPS (Transactional Processing System): a software application designed to efficiently handle the gathering and retrieval of customer and business data during a commercial transaction.

T2DM (Type 2 diabetes mellitus): A medical condition characterised by elevated blood glucose levels due to insufficient production or use of insulin by the body.

W

WHO (World Health Organisation): is the United Nations agency that connects nations, partners, and people to promote health, keep the world safe, and serve the vulnerable – so everyone, everywhere can attain the highest level of health.

CHAPTER 1:

INTRODUCTION

1.1 Introduction

Given that medication non-adherence exacerbates disease progression in NCD patients, resulting in poor health outcomes and significant economic burdens due to increased hospitalisations, emergency care visits, and complications, this thesis employed several ML techniques to develop predictive models for MA analytics among NCD patients, with a particular focus on diabetic and hypertensive patients. The framework of this thesis approaches the research subject systematically and presents the findings coherently. Chapter one (1) presents a detailed overview of NCDs, the importance of MA, and the promise of ML to address the adherence dilemma. Chapter two (2) provides a thorough overview of existing ML approaches to predicting MA, noting gaps and potential for future study. Chapter three (3) focuses on feature selection and the importance of various predictors, with a detailed analysis of the features that impact MA using data from a Zimbabwean medical aid firm. Chapter four (4) delves into the practical components of data preparation, including data wrangling, cleaning, and generation processes required for building accurate and robust ML models. Chapter five (5) entails the development and validation of the MA classification models and discusses their interpretability and explainability. In Chapter 6, a new hybrid model is developed by twinning machine learning algorithms for MA analytics in patients with diabetes and hypertension. Finally, Chapter seven (7) provides summaries of the key research findings, makes recommendations to healthcare practitioners and governments, and suggests areas for future research. The thesis's organised approach intends to contribute to the field of healthcare management by developing robust models and solutions to enhance MA among NCD patients, thereby improving health outcomes and lowering economic burdens in Zimbabwe and other comparable contexts.

1.2 Background to the Study

Non-communicable diseases (NCDs) cause more than two-thirds of global mortality (Benziger et al., 2016; World Health Organization, 2018). The NCDs claim 41 million lives each year, accounting for almost 74% of all worldwide deaths (WHO, 2022). These conditions result from a complex interaction of genetic, physiological, environmental, and behavioural factors (Swed et al., 2023; WHO, 2022). Thus, the focus of this study is on a population with NCDs, not only because they account for the majority of total global deaths but also because they impose a significant economic burden on family and national budgets by making NCD patients less productive, working for fewer years, and dying prematurely (Odunyemi et al., 2023; Thanh & Duong, 2022; Yang & Yan, 2023). It is critical to focus on hypertensives and diabetics, as the number of people with hypertension has increased to 1.28 billion in the last two decades (WHO, 2021). Hypertension is the leading cause of cardiovascular disease (CVD) and premature death worldwide (Mills et al., 2020).

Diabetes affects at least 422 million people globally, with the vast majority living in low- and middle-income nations. The condition causes 1.5 million deaths per year, according to the World Health Organisation (2022). Diabetes, namely type 2 diabetes mellitus (T2DM), has emerged as the most common chronic disease and a leading cause of death worldwide (Chen et al., 2024). A recent epidemiological study predicts that by 2030, the global prevalence of type T2DM will rise to 7079 per 100,000 people (Abdul Basith Khan et al., 2020). Type 2 diabetes (T2D), a highly prevalent global disorder (Ogurtsova et al., 2017) is a chronic condition that requires ongoing treatment to improve quality of life and reduce the risk of associated disability and mortality (Jacobson, 2004; Stolar, 2010). Research has demonstrated the importance of MA for patients with T2D to improve glycemic control, minimise complications, and reduce overall healthcare expenditures (Comer et al., 2015; Lin et al., 2017).

Researchers have found that not following through with treatment for NCDs, such as diabetes and hypertension is a major cause of these conditions' bad outcomes (Algabbani & Algabbani, 2020; Hamrahian, 2020). Medication adherence (MA) is the extent to which a patient's medication-taking behaviour corresponds to the recommendations prescribed by

a healthcare expert. It is critical for achieving the desired therapeutic outcomes in NCD patients (Babel et al., 2021). Medication adherence (MA) is administering medication according to a dosing schedule's suggested intervals and doses. Medical adherence refers to the extent to which a patient's actions align with the recommendations of healthcare professionals, specifically in relation to taking medications as prescribed (Lam & Fresco, 2015).

Effective adherence to prescribed medication is critical for successful medication management and achieving clinical goals. Medication non-adherence is a common issue in chronic disease management (Náfrádi et al., 2017). Maintaining consistent adherence to prescribed medications is a serious challenge in healthcare, with far-reaching consequences (Pradeep, 2023). Despite the repercussions of not adhering to medical advice, over 50% of patients do not take their prescriptions as directed. Failure to comply with medication regimens has an immediate impact on clinical outcomes (Huber et al., 2019; von Wyl et al., 2020). According to WHO research on MA for chronic diseases, people in developed countries barely take half of their prescribed medications (Ahmed et al., 2018). In developing countries, this adherence statistic is comparably lower. The majority of medications provided for chronic conditions, such as diabetes and hypertension, have adherence rates of 50–60 percent. This holds even for those with comprehensive insurance coverage and medication benefits, according to the American Heart Association (2021). Although diabetes demands long-term care, the adherence rate to injectable medicines such as insulin is around 60%, which is lower than the acceptable threshold of 80% (Chen et al., 2024).

Medication adherence assessment using management data is critical for evaluating drug effectiveness, identifying variables contributing to poor adherence, and developing appropriate interventions. The World Health Organisation states that various factors such as economic, social, healthcare, patient-related, provider-related, and therapy-related aspects can cause non-adherence. A multiplicity of variables affect MA levels, including patient-related factors (such as low health literacy, multiple health conditions, and lack of involvement in treatment decisions), physician-related factors (such as communication

difficulties or receiving care from multiple doctors), and healthcare system-related factors (such as limited access to care and a lack of health information technologies) (Brown & Bussell, 2011). Effective data analysis on administrative databases containing pharmacy and claims information can optimise prescription processes, improve healthcare staff productivity, and enhance patient satisfaction.

The rapidly evolving science of medical informatics has brought about a period of data-rich medical services (Berger & Doban, 2014). Because of the prevalence of big electronic health records (EHR) datasets and the scarcity of medical specialists in the healthcare sector, analysing a patient's medical history and evaluating outcomes would be a time-consuming process (Berwick & Gaines, 2018). Claims databases and EHRs are increasingly consolidating patient MA data (Xu et al., 2023). Administrative reimbursement records and EHRs contain significant medical data because of their large volume, frequent updates, and diversity. This data provides a thorough insight into patient behaviour and health trends (Yang et al., 2020). In a healthcare setting with abundant data, it is important to consider the use of indirect methods to assess MA (Al-Noumani et al., 2023).

Scientists have developed a variety of advanced metrics to precisely assess how well people adhere to their medication regimen. The assessments cover a wide range of actions, from simple tallies of distributed medications to complex analyses of unmet procurements or periods without medication access (Galozy et al., 2020). Every method provides a unique viewpoint on the patterns of medicine use. Administrative data, such as the medicine possession ratio (MPR) and the proportion of days covered (PDC), can be used to measure MA (Galozy & Nowaczyk, 2020; Malo et al., 2017). The data used in this thesis originated from the transactional processing system (TPS), which collects and manages medical claims and patient-level data. Modern systems such as electronic prescription services and complex pharmacy insurance claims have built secondary databases, which have become critical tools for precisely monitoring drug adherence using intricate refill metrics (Lam and Fresco, 2015).

Traditional risk evaluation and reduction techniques in MA frequently rely on manual processes and retrospective analysis, which may be insufficient for recognizing and resolving developing risks promptly (Guerra, 2024). Given its importance, exact adherence evaluation has become a top priority for policymakers, researchers, and healthcare professionals (Franchi et al., 2022). Artificial intelligence (AI) and ML deliver accurate estimates and results. The complexity and variety of MA barriers necessitate the implementation of various methods to improve adherence (Brown & Bussell, 2011). Artificial intelligence (AI) technology is emerging as a useful tool in these interventions. AI systems have demonstrated promising results in reliably detecting patients who do not adhere to their medication, providing an innovative approach to addressing the challenges of medication non-adherence (Pulice & Coustasse, 2024). This study employed ML algorithms to construct MA prediction models.

Machine learning (ML) is a subset of AI that enables computers to learn and make decisions without explicit programming (Christopoulou, 2024). Machine learning (ML) uses advanced algorithms and techniques to analyse large and complicated datasets, extracting key insights and discovering patterns (Ai et al., 2021; Almutiri et al., 2022). Data analytics requires developing algorithms and models to study and comprehend data, identify trends, and make informed predictions (Christopoulou, 2024). Healthcare systems are complex and demanding for all parties involved, but AI has revolutionised various sectors, including healthcare, by improving patient care and quality of life (Alowais et al., 2023). Despite challenges in integration, AI has considerable potential for improving MA and patient outcomes. As a result, substantial training for healthcare staff is required to effectively use AI technology in clinical contexts (Pulice & Coustasse, 2024). According to (Warren et al., 2022), AI and ML are excellent techniques for predicting drug adherence in diseases like diabetes and hypertension. These devices have an accuracy of 70% to 80% in detecting those who do not take their medication as prescribed.

Machine learning (ML) is becoming an important part of the healthcare industry, ranging from biomedical research to healthcare delivery (Masiero et al., 2023). Healthcare ML is a powerful tool that can help with illness diagnosis, medical image analysis, forecasting

patient outcomes, optimising processes, detecting high-risk patients, and delivering personalised treatment (Christopoulou, 2024). Technological breakthroughs are continuously improving patient care, enhancing diagnostic capabilities, and optimizing healthcare procedures. Hospitals can improve their ability to identify and address possible risks, allocate resources more effectively, and improve clinical outcomes by implementing advanced technologies like AI-driven predictive analytics, natural language processing, and ML algorithms (Guerra, 2024).

Researchers (Li et al., 2022; Aziz et al., 2020; Warren et al., 2022) have successfully used ML to predict the development of medical conditions such as type 2 diabetes, hypertension, and other NCDs. Machine learning (ML) provides various advantages over traditional statistical methods, including increased flexibility, higher prediction accuracy, automation opportunities, and the ability to handle large datasets (Masiero et al., 2023). Furthermore, scientists have used ML approaches to predict the likelihood of nonadherence in people with type 2 diabetes, demonstrating its ability to identify patients who are prone to poor MA (Walsh et al., 2019). Using predictive models and ML to identify non-adherent patients at an early stage could improve outcomes for people with NCDs (Zakeri et al., 2022). Because of the increased availability of data and the pressing need to convert it into practical insights, ML has emerged as a critical tool in data mining (Xu et al., 2023). These modern technologies have improved prediction accuracy by applying variable selection algorithms, examining non-linear correlations between predictors and outcomes, and investigating complicated interconnections among predictors (Deo, 2015).

There is still a gap that requires further investigation, as Beam & Kohane (2018) claim that there is no clear distinction in the performance between ML models and classical statistical models. Once again, the prevalence of inconsistent results in the available literature regarding the efficacy of various ML models in predicting MA highlights the need for further research. A study by Owusu-Adjei et al. (2023) looked at how well different ML models worked and found that LR had the best predicted score of 0.90, followed by GB with 0.87 and LR with 0.87. RF, on the other hand, had the worst predicted score of 0.77. Wang et al. (2020) demonstrated that support vector machines (SVM) performed better in

classification accuracy than LR and back-propagation neural networks (BPNN). Lee et al. (2013) found that SVM outperformed LR in terms of classification accuracy; SVM achieved a classification accuracy of 97.3%, while LR reached just 71.1%. Other studies (Gottlieb et al., 2022; Owusu-Adjei et al., 2023) found that the RF outperformed the LR in predicting the probability of patients leaving a rehabilitation programme.

Upon further examination of previous research, it emerged that ML methods do not exhibit superior predictive capabilities when compared to traditional statistical methodologies. Franklin et al. (2016), Hu et al. (2020) and Zullig et al. (2019) found no discernible variation in the models' performance when they included similar predictive factors. Another study found that the ML model had much greater accuracy (0.97) than conventional approaches (0.71), but had an internal validation issue. Lee et al. (2013) evaluated the ML model using the dataset they had previously used to train it. Their evaluation results revealed a high level of accuracy, albeit with bias. Two other studies (Bourdès et al., 2011; Hu et al., 2020) incorporated multiple variables about the patient's medical history into the ML models. These models exhibited greater accuracy compared to the LR models. Upon initial examination, it appeared that only ML models offered the opportunity to incorporate multiple predictive variables into the model. Another study also found that incorporating multiple plus a history of MA into the LR model improves its predictive accuracy even without the use of ML approaches (Franklin et al., 2016).

Previous research used various regression methods to uncover features linked with MA (Al-Ganmi et al., 2020; Fernandez-Lazaro et al., 2019). However, this thesis takes an alternative approach, viewing the MA problem as a classification problem and employing ML-based classification techniques.

An alternative viewpoint with conflicting research outcomes found that race, ethnicity, and gender do not significantly predict treatment discontinuation (Stafford et al., 2022), whereas Li et al. (2019) discovered that race or ethnicity is a significant predictor of non-adherence to antihypertensive drugs. Furthermore, (Son et al., 2010), as well as Hasan et al. (2021), found that gender can predict MA in patients with high fever and opioid use

disorder, respectively. There was inconsistent data regarding gender's ability to predict major depressive disorder in patients with the same chronic condition. This demonstrates that variations in the research environment, study duration, and participant characteristics have a major impact on the research results. Gottlieb et al. (2022) found that age had no significant effect on model prediction accuracy. Other studies, however, have found that younger people are less likely to adhere to specified treatment plans (Lappan et al., 2020).

It is critical to realise that the factors influencing MA differ across different study scenarios. Several factors influence susceptibility to the various types of NCDs being studied. These factors include the techniques used for feature selection, the study setting, cultural and regional differences, the level of economic development, education level, the presence of social desirability bias, differences in living habits, urban or rural residence, variations in social and healthcare systems, and other relevant factors. (Comer et al., 2015; Kanyongo & Ezugwu, 2023; Kassahun et al., 2016). These difficulties impede the development of accurate models for predicting patient adherence and treatment plans. This study uses various ML approaches to classify microalbuminuria (MA) in diabetic and hypertensive patients. It also incorporates unique supplementary attributes and makes use of a distinct Zimbabwe dataset.

1.3 Research Motivation

This study is motivated by the urgent need to address the prevalent and costly issue of medication non-adherence in the management of NCDs. NCDs, including diabetes, hypertension, and CVD, are among the leading causes of death and disability worldwide, affecting millions of people and exerting a significant strain on healthcare systems. The successful management of many chronic illnesses is strongly dependent on constant MA. However, non-adherence rates are unnervingly high, with up to 50% non-compliance with long-term therapy (Unni & Bae, 2022), resulting in poor health outcomes, more frequent hospitalisations, and large healthcare expenses. The study refers to this challenge as 'the MA problem'. Indeed, the problem of non-adherence to medication among NCD patients has persisted ever since the WHO reported about two decades ago that poor adherence to medicines is a global phenomenon, particularly for long-term medication, with estimated

non-adherence rates of 50% (World Health Organisation, 2003). Chronic disease is typically associated with increased non-adherence, with research demonstrating that medication non-adherence complicates diseases such as hypertension or diabetes that have been around for several years (Kleinsinger, 2018). If not addressed, the consequences of poor MA can range from patient-specific negative health effects to significant demands on healthcare system resources.

This study focuses on the NCD population, not only because they account for the majority of global deaths, but also because they significantly burden family and national budgets due to reduced productivity, shorter work lifespans, and premature death (Odunyemi et al., 2023; Thanh & Duong, 2022; Yang et al., 2023). Medication nonadherence is a significant healthcare barrier, particularly for asymptomatic conditions like diabetes and hypertension (Gardezi et al., 2023). According to a WHO study on drug adherence for chronic illnesses, citizens of developed economies take only half of their prescribed drugs (Ahmed et al., 2018); in underdeveloped countries, the statistic is even worse in underprivileged communities and developing nations.

It is critical to focus on hypertensives and diabetics, as the number of people with hypertension has increased to 1.28 billion in the last two decades (World Health Organisation, 2021). Hypertension is the leading cause of CVDs and premature death worldwide (Mills et al., 2020). Diabetes affects at least 422 million people worldwide, the vast majority of whom live in low- and middle-income nations. The disease causes 1.5 million deaths per year (World Health Organisation, 2022). Even though the numbers on NCDs are very low, researchers have found that not taking medications as prescribed for diseases like diabetes and high blood pressure is a major cause of their effects (Algabbani & Algabbani, 2020). The medication non-adherence problem is highly pronounced globally, yet in developing countries, the majority of traditional methods for monitoring and improving MA, such as patient self-reports, patient interviews, pill counts, prescription refill surveys, diaries, and electronic monitoring devices, have significant limitations. These methods are frequently obtrusive, erroneous, and unscalable, failing to deliver timely and useful insights. There is an urgent need for more advanced, precise, and non-invasive

techniques for classifying and improving MA. Machine learning (ML) has transformational potential in this particular scenario. The ability of ML algorithms to analyse large and complex datasets such as EHRs, pharmacy refill data, patient demographics, and behavioural patterns provides a promising avenue for identifying predictive factors of MA as well as classifying MA or non-adherence. ML-based outcome-oriented approaches outperform human-driven or traditional statistical methods in computing the appropriate thresholds for distinguishing patients depending on outcomes of interest. (Gu et al., 2021) emphasise the exceptional performance of using ML to extract relevant amounts of data from medical care data for medication management.

Wang et al. (2020) argue that ML models have the advantages of including nonlinear connections, less biassed auto-learning, and greater flexibility to avoid overfitting. The primary difference between ML approaches and conventional methods is that ML trains a model by observing real-world data rather than preprogramming it with preconceived instructions. AI and ML have ushered in a new era of chronic disease treatment and management, shifting away from traditional management strategies and towards individualised data-driven precision care (Sahu et al., 2022). The goal of this study is to use ML to overcome the limitations of traditional methods in MA classification, specifically by developing several ML models for diabetics and hypertensive patients. This study intends to give healthcare clinicians significant insights into tailoring interventions and improving patient outcomes in a developing nation by generating various robust, interpretable, and generalizable ML models for MA classification. Accurately diagnosing non-adherence early on is critical for assigning resources to individuals who require additional care and tailoring the intervention to address the specific causes of medication nonadherence.

1.4 Aim and Objectives

The primary aim of this study entails harnessing the potential of ML techniques to build models for classifying MA among patients with NCDs, with a particular emphasis on diabetics and hypertensives. By leveraging a multidimensional approach and incorporating patient-level and medical claims data, the study endeavours to ascertain significant features influencing MA. The development and comparison of multiple ML models for MA classification among diabetes and hypertension patients is central to this aim, which employs varied ML approaches encompassing supervised learning algorithms, a deep neural network, an ensemble method, and a newly developed hybrid ML algorithm. The objectives listed below provide direction for achieving this goal:

i. Perform an extensive systematic literature review of the ML approaches to MA analytics among NCD patients.
ii. Conduct a comprehensive analysis of the existing body of literature on feature selection and the significance of MA predictors in patients with NCDs from the standpoint of ML research.
iii. Using patient-level and medical claims data, to wrangle and generate data for ML models in MA analytics from a practical perspective.
iv. Determine the features that significantly predict MA in patients diagnosed with both diabetes and hypertension.
v. Leveraging multiple ML algorithms, encompassing K-NN, SVM, RF, DNN, DT, LR, and Nave Bayes, to develop models for classifying MA among NCD patients.
vi. Compare and assess the performance of K-NN, SVM, RF, DNN, DT, LR, and Nave Bayes models for classifying MA in patients with diabetes and hypertension.
vii. Develop twinned machine learning algorithms for medication adherence analytics in patients with diabetes and hypertension.

1.5 Scope of the Thesis

The scope of this study encompasses the classification of MA among diabetic and hypertensive patients using ML techniques, with a particular emphasis on data retrieved from a Zimbabwean medical aid organisation in 2022. Medication adherence was the primary variable of interest, and it was predicted using ML classification modelling, with the prescription refill threshold serving as a proxy. Medical claims and patient-level data were used to create the classification models for MA. The study created classification models using ML techniques such as SVM, random forests (RF), decision trees (DT), naive bayes, logistic regression (LR), k-nearest neighbor (k-NN), deep neural networks (DNN), newly developed hybrid ML technique, as well as identifying patterns in MA and features influencing MA.

The study limited its geographic scope to data taken from a medical aid organisation operating in Zimbabwe, offering a more localised context that takes into account the region's specific healthcare environment and patient demographics. This localised focus enables a thorough examination of the unique healthcare dynamics, socioeconomic factors, and patient behaviours within the Zimbabwean context, providing insights that are specifically tailored to this setting while also having room for generalizability in similar settings across the region and beyond.

The study's time frame is set for 2022. This encompasses the data collection, model creation, validation, and analysis stages. By focusing on this timeframe, the study gathers a complete snapshot of MA patterns and affecting factors throughout a full year, assuring the robustness and applicability of the predictive models developed. Because adherence metrics measure the percentage of patients covered by prescription claims for the same medication in the same therapeutic class within a given year, the developed and used dataset only included data for 2022, which ran from January 1 to December 31, 2022. The ultimate goal is to offer healthcare clinicians actionable information to improve MA among diabetic and hypertensive patients, ultimately improving health outcomes and maximising resource utilisation within Zimbabwe's healthcare system. This study focuses on a population with diabetes and hypertension, not only because they are among the leading

causes of comorbidities and deaths worldwide, but also because they impose a significant economic burden on family and national budgets as patients become less productive, work for fewer years, and die prematurely.

1.6 Research Contributions

This study aims to make major contributions to the field of healthcare management, specifically in the prediction and classification of MA among patients with NCDs using ML methods. The research has made significant contributions in the following areas:

- The study is a direct response to Bohlmann et al. (2021) who revealed the dearth of studies that incorporate ML to assess how well NCD patients adhere to medication.
- This thesis provides a vast and well-collated literature on MA analytics integrated with ML approaches. The study provides a complete, well-informed, systematic review of the literature on NCD MA employing ML approaches such as deep learning, RL, and ensemble learning. Contemporary research using ML techniques and data acquired from EMRs, self-reported MA surveys, or remote real-time assessments of medication administration and therapeutic drug monitoring (TDM) in NCD patients has been compiled and retrospectively reviewed.
- Due to the increasing number of scientific papers, it is currently nearly impossible for a scientist to keep up with all of the new study findings in the scientific literature about ML-based analytics of MA in NCD patients. As a result, the scientific literature has grown fragmented, with different experts focusing on areas relevant to their specific fields of interest. Interestingly, this study collects, unearths, and summarises a large body of literature on ML-based MA analytics in NCD patients, with a focus on MA predictors. The study's findings help researchers develop and test new research hypotheses and theories. Scientists will be inspired to explore new research areas as a result of the time and effort savings afforded by our study.
- In the context of developing countries, the study generates a state-of-the-art dataset for MA analysis among diabetic and hypertensive patients. The data can examine variations in MA levels among NCD patients based on individual features. This then warns the patients who exhibit the related characteristics and require specialised care to tackle the epidemic of medication non-adherence.

- This study provides significant tools for researchers, data scientists, and analysts interested in data preparation, mining, and enrichment. This, in turn, may lead to the development of additional datasets for assessing MA and undertaking research on NCDs, which currently have few datasets available for MA evaluation.
- This study created multiple robust machine-learning models that can effectively predict MA in diabetic and hypertensive patients. These models combine multidimensional data such as patient-level data, medical claims, pharmaceutical refill data, and linked EHRs to provide healthcare practitioners with a comprehensive tool for identifying at-risk individuals. The established models considerably improve clinical decision-making by accurately identifying cases of non-adherence. This allows healthcare practitioners to prioritise patients who require extra attention and personalise interventions to the individual factors that contribute to medication non-compliance.
- In particular, the study informs NCD patients, health carers, and health practitioners about the accuracy of ML algorithms in measuring NCD MA, allowing them to make informed decisions. In doing so, the study encourages an open discussion about whether the value proposition of ML-based analytics for monitoring MA and implementing treatments is true or flawed.
- This study examines a large dataset from a medical aid provider to identify crucial features impacting MA. These findings contribute to a better understanding of adherence behaviors in the Zimbabwean context, as well as in possibly similar situations.
- The research work focuses on the development of interpretable ML models using SHAP beeswarm plots, guaranteeing that the outputs are practical and understandable for healthcare providers. This makes it easier to put the models into practice in clinical settings, allowing providers to make more informed patient care decisions. Improving interpretability is critical in healthcare, as model predictions have a direct impact on patient care and outcomes.
- This study helps diabetic and hypertensive patients have better health outcomes by giving healthcare practitioners additional tools to anticipate and manage drug

adherence. This study's findings lead to more customised and effective interventions, decreasing the burden of NCDs in Zimbabwe and abroad.

- Overall, this study enhances the use of ML in healthcare by proposing fresh techniques for addressing the recurrent problem of medication non-adherence in the management of NCDs. The study highlights the significance of ML in MA analytics, which enhances patient health monitoring outcomes, reinforces patient safety measures and standards, and raises the overall level of care.

1.7 Thesis Layout

To effectively address the intricate problem of MA in NCD patients using ML techniques, it is vital to approach the study in a structured and systematic manner. The study's chapters were designed to gradually lay the foundation, present the methodology, and give the study's findings, conclusions, and recommendations. Each of the seven chapters serves a specific purpose and helps to achieve the overall goal of building robust ML-based classification models for MA analytics. This thesis follows a specified structure:

Chapter 1: Introduction

This chapter provides a detailed explanation of the background and rationale behind the study while placing the research problem within its appropriate perspective. Therefore, the chapter provided a thorough understanding of the MA phenomenon in the context of NCDs and introduced the essential concepts of MA and ML. This laid a solid foundation for the upcoming research components and chapters. The research motivation offered allows for the conceptualization of the research problem. The chapter underscored the significance of addressing MA in the context of NCDs highlighting the value proposition of employing various machine-learning methods to address this perennial problem. The chapter additionally delineates the scope of the study, with a specific emphasis on the conceptual scope, geographical delimitation, and time scope covered by the extracted data. The study articulates its research contributions in the latter part of the first chapter.

Chapter 2: Systematic literature review on ML approaches to medication adherence analytics (Paper 1)

This chapter provides a systematic review of the current body of literature on ML methods used to predict MA in patients with NCDs. The paper scrutinizes various algorithms, approaches, and datasets employed in previous research. The paper specifically identifies areas that require further investigation and potential avenues for future study. The review establishes a robust theoretical basis and situates the current research within a wider scope of relevant domain knowledge.

Chapter 3: Systematic literature review on feature selection and importance of predictors of medication adherence from ML research perspectives (Paper 2)

This chapter focuses primarily on identifying and reviewing the most relevant features that impact MA in patients with NCDs. The chapter examines several feature selection strategies and emphasises the importance of different predictors from the viewpoint of ML research. By offering a comprehensive understanding of the key variables that influence MA among NCD patients, this chapter establishes the foundation for the later creation of predictive and classification ML models.

Chapter 4: Data wrangling and generation for ML models in medication adherence analytics (Paper 3)

This chapter delves into the pragmatic aspects of data preparation for ML models. The chapter discusses the procedures and building blocks involved in manipulating, refining, and creating data using patient-level and medical claims data obtained from a medical aid firm in Zimbabwe. Additionally, it delineates the approach for data transformation, enrichment, and merging diverse data sources while guaranteeing data quality.

Chapter 5: Development and validation of ML models for medication adherence analytics (Paper 4)

This chapter focuses on the creation and validation of ML models that were used to classify MA. The process encompasses the identification of suitable algorithms, the model training phase, and the assessment of model performance. The chapter also discusses model

interpretability, with a particular emphasis on the use of SHAP values to elucidate the model's predictions. A comparison was made between the performance of seven built ML models, and the ML model that achieved the highest performance in classifying MA was identified and reported.

Chapter 6: Development and validation of a twin machine learning algorithm for medication adherence analytics in patients with diabetes and hypertension (Paper 5)

The paper in this chapter aimed to improve model performance in classifying medication adherence by developing two hybrid machine learning models: TWINNED-SVM-RF and TWINNED-SVM-DT. This innovation involved integrating Support Vector Machines (SVM) with tree-based algorithms—Decision Trees (DT) and Random Forests (RF)—to enhance the classification of medication adherence in patients with diabetes and hypertension.

Chapter 7: Conclusions and Recommendations

The last chapter provides a concise overview of the study's major findings, concluding the examination and outcomes reported in the preceding chapters. The document provides implications for healthcare practitioners, policymakers, and future research endeavours. The chapter also discusses the limitations of the study and suggests prospects for future research.

CHAPTER 2

Systematic literature review on ML approaches to medication adherence analytics (Paper 1)

2.1 Introduction

Chapter two (2) presents a detailed and organised review of the current ML methods for predicting MA in patients with NCDs. The chapter begins by delineating the scope and methodology of the systematic literature review, explicating the techniques employed for conducting the search, the criteria for including or excluding literature sources, and the electronic literature databases used for retrieving the required research material. The chapter then examines the implementation of different ML algorithms, including supervised learning approaches such as LR, DT, RF, SVM, and deep learning techniques, to address this specific issue. Each method is analysed based on its application, performance, and appropriateness for predicting MA. The chapter also examines unsupervised learning methods, including clustering and association rule mining, to reveal patterns and groupings within adherence data. By and large, the chapter collates diverse literature pertaining to recent work on ML application to MA analytics across different NCD patients, encompassing diabetics, hypertensives, those living with chronic respiratory diseases, CVDs, cancer, and other NCDs.

The review examines various datasets used in previous research, including EHRs, pharmacy refill data, patient-reported MA survey data, continuous glucose monitoring data, and wearable device data. The multiplicity of data sources underscores the intricate and diverse nature of MA, underscoring the necessity for integrated data solutions. The chapter evaluates the various data sources and methodologies for generating data that then allow the ML algorithm to analyse MA. In addition, the chapter examines the evaluation metrics and validation techniques used in previous research to measure the effectiveness of ML models in MA analytics and modeling. These include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic (ROC) curve. The chapter also discusses the importance of ML model evaluation, interpretability, and explainability, particularly in the healthcare domain, where comprehending the reasoning

behind model predictions is essential for clinical acceptance and confidence. Ultimately, the chapter identifies areas that require further investigation and potential avenues for future research. It highlights the importance of future researchers concentrating on empirically studying the use of ML in assessing MA among patients with NCDs, using real-world data from Africa. The paper also highlights the potential advantages of integrating sophisticated methodologies, such as ensemble learning, to improve the performance of the model.

Paper status: Published

Paper link: https://doi.org/10.1016/j.imu.2023.101210

Journal Name: Informatics in Medicine Unlocked

Journal impact factor: NA (Q2)

Student's contribution: Conceptualisation, data collation, methodology, results and original manuscript writing

Other author's contributions: Supervision, editing, and revision

Percentage of student's contribution: 90%

Informatics in Medicine Unlocked 38 (2023) 101210

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Machine learning approaches to medication adherence amongst NCD patients: A systematic literature review

Wellington Kanyongo [a,b,**], Absalom E. Ezugwu [c,*]

[a] Department of Computer Science, Faculty of Science Engineering, Bindura University of Science Education, Bindura, Zimbabwe
[b] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg Campus, Pietermaritzburg, 3201, South Africa
[c] Unit for Data Science and Computing, North-West University, 11 Hoffman Street, Potchefstroom, 2520, South Africa

ARTICLE INFO

Keywords:
Medication adherence
Machine learning
Deep learning
Non-communicable disease
Chronic disease
NCD patients
Diabetes
Hypertension
Cardiovascular disease (CVD)
Cancer
Respiratory disease

ABSTRACT

Non-adherence to prescribed medication is a major public health concern that escalates the risk of morbidity and death as well as incurring extra expenses associated with hospitalisation. According to the World Health Organization (WHO), only 50% of people suffering from chronic diseases follow the treatment recommendations despite the counsel provided to patients on the importance of medication adherence (MA). Early detection of non-communicable disease (NCD) patients poorly adhering to recommended medications using analytics based on machine learning (ML) may improve the outcomes of NCD patients positively. This paper presents a systematic review of literature involving the application of ML in evaluating MA amongst NCD patients. The articles considered in this study were extracted from Web of Science, Google Scholar, PubMed, and IEEE Explore. Twenty-five articles in total met the criteria for inclusion. These were articles that utilised ML techniques to analyse MA in NCDs, with patients suffering from diabetes (n = 8), hypertension (n = 3), cardiovascular disease (CVD) and statin adherence (n = 6), cancer (n = 3), respiratory diseases (n = 2), and other NCD conditions (n = 3). The proportion of days covered (PDC) was typically used to evaluate MA. It emerged from the study that for MA to be considered high, the adherence threshold should be at least 75% of the PDC, a universally accepted threshold. In MA analytics research and practice, a PDC ≥80% threshold is typically regarded as a high level of adherence to prescription medication. Logistic regression (LR) (n = 12), random forest (RF) (n = 11), support vector machine (SVM) (n = 7), neural net (n = 6), ensemble learning (n = 6), MLPs (n = 4), XGBoost (n = 3), Bayesian network (BN) (n = 3), and gradient boosting (n = 3) were the most frequently applied ML techniques in the analytics of MA amongst NCD patients. It should be underscored that leveraging standard ML, deep learning (DL), and ensemble learning has enormous potential for measuring MA amongst NCD patients based on various analytics such as prediction, regression, classification, and clustering. Moreover, a further study could be conducted to comprehend how the application of alternative ML-based techniques can be used to measure MA among patients with chronic infectious diseases.

1. Introduction

Non-adherence to medication regimens that medical specialists have prescribed has been pinpointed as a major problem globally (Sabat et al., 2003; [1–4]. Medication non-adherence is associated with several adverse consequences, notably increased patient health risks, an additional financial burden for clinicians and the patients themselves, healthcare providers, and the healthcare sector, among other stakeholders [5], as well as increased morbidity, mortality, and hospital admissions [6]. Medication adherence (MA) is defined by Ref. [7] as the degree to which a patient's behaviour regarding medications is consistent with the recommended medical instruction or guidance from a healthcare expert. MA is essential for non-communicable disease (NCD) patients to achieve their expected clinical results [8]. The current systematic review focused on NCDs such as diabetes, hypertension, CVDs, cancer, and respiratory diseases which claimed approximately 33.2 million people worldwide in 2019, a 28% increase compared to year-2000 [1]. Consequently, identifying patients predisposed to poor adherence can lower the risk of hospitalisation and adverse health consequences.

The level of MA must always be assessed to detect non-adherent

* Corresponding author.
** Corresponding author. Department of Computer Science, Faculty of Science Engineering, Bindura University of Science Education, Bindura, Zimbabwe.
E-mail addresses: wkanyongo@gmail.com (W. Kanyongo), Absalom.Ezugwu@nwu.ac.za (A.E. Ezugwu).

https://doi.org/10.1016/j.imu.2023.101210
Received 3 January 2023; Received in revised form 3 March 2023; Accepted 4 March 2023
Available online 8 March 2023
2352-9148/© 2023 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Abbreviations and acronyms

The abbreviations and acronyms used in this article are listed below.

Nomenclature	Full Expression
AET	Adjuvant Endocrine Therapy
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
CAP	Cumulative Accuracy Profile
CBR	Case Based Reasoning
CART	Classification And Regression Trees
CGM	Continuous Glucose Monitor
CHAID	Chi-squared Automatic Interaction Detection
CI	Confidence Interval
CNN	Convolutional Neural Network
COPD	Chronic Obstructive Pulmonary Disease
DL	Deep Learning
DNN	Deep Neural Network
EDA	Exploratory Data Analysis
T2D	Type 2 Diabetes
EHR	Electronic Health Record
EMR	Electronic Medical Record
GPU	Graphical Procesing Unit
HF	Heart Failure
K-NN	K-Nearest Neighbour
LASSO	Least Absolute Shrinkage and Selection Operator
LOOCV	Leave-One-Out Cross-Validation
LMTs	Logistic Model Trees
LR	Logistic Regression
LSVM	Linear Support Vector Machine
MA	Medication Adherence
MAE	Mean Absolute Error
MEMS	Medication Event Monitoring System
META	Mean Ensemble Test Accuracy
ML	Machine Learning
MLP	Multi-Layer Perceptron
MVP	Medtronic Virtual Patient
NCD	Non-Communicable Diseases
NPV	Negative Predictive Value
NYHA	New York Heart Association
OHE	One Hot Encoding
PDC	Proportion of Days Covered
PPV	Positive Predictive Value
QUEST	Quick Unbiased Efficient Statistical Tree Algorithms
RF	Random Forest
RL	Reinforcement Learning
ROC	Receiver Operating Characteristic Curve
RMSE	Root Mean Square Error
RSV	Random Survival Forest
SOM	Self-Organizing Map
SMOTE	Synthetic Minority Oversampling Technique
SPSS	Statistical Package for Social Sciences
SSB	Smart Sharps Bin
ST	Survival Trees
SVM	Support Vector Machine
SVR	Support Vector Regression
TDM	Therapeutic Drug Monitoring
WHO	World Health Organisation
XGBoost	Extreme Gradient Boosting

patients. However, precise and cost-effective adherence assessment is a significant hurdle. Within NCDs, a broad range of data sources must be available to aid the analysis of MA or non-adherence; these are diverse kinds of data accessible in large sets, allowing ML analytics in hindsight or in real time. To cluster, classify, predict, and analyse MA, ML approaches require input data generated through diverse health and medication management procedures. Health data provide artificial intelligence (AI) applications and ML models with valuable information for individual healthcare evaluation, analysis, and disease prognosis. Traditionally, less precise and rigorous approaches comprise clinician evaluation via patient interviews and self-reports [9]. Arising from patient interactions with the healthcare system, a vast number of electronic medical records (EMRs) are maintained by healthcare institutions, government agencies, and medical health insurance companies. An EMR enables data recording consistently and uniformly and greater data accessibility. EMRs store data that can be used to train and improve algorithms which can predict and manage NCDs [10].

Measuring MA is common practice and can be based on pharmacy dispensing data, patient medication refill data, insurance claims data, and EMRs [11,12] or based on secondary databases. The most common metrics are the proportion of days covered (PDC) and medication possession ratio [13]. [14] further point out that pharmacy refills and claims data are some of the most often used variables for assessing MA in everyday practice and clinical intervention research. These metrics describe the number of medications acquired over a predetermined period as a percentage, typically between 0% and 100%. The 80% cutoff point is commonly used to define acceptable adherence; at or above this level, patients are regarded as adhering to a certain prescription and non-adherent otherwise. The historical roots of the 80% threshold may be traced back to a 1975 blood pressure therapy trial involving 230 steelworkers [15]. The 80% threshold is the most frequently employed for arbitrary or historical reasons in adherence and prediction research. If this level is exceeded, it is presumed that medication treatment has been successful [16,17]. Studies in the past have shown that preliminary and historical trends of prescription refilling can significantly improve model performance, making the model predictions more accurate [18–20]. Moreover, the growing digitization of healthcare is creating a wealth of new tools that can be used to look at refilling habits and the timing of drug exposure validly and cost-effectively [12].

In real-time medication dosing, technology improvements that automate real-time medication dose monitoring present an opportunity to leverage continuous data sources to analyse MA dynamically and take timely and appropriate action [21]. claim that real-time dose assessment can dynamically and accurately predict MA, enabling early clinical intervention to improve patient outcomes. Data on missed once-daily basal insulin doses, for example, in diabetes patients, can be utilised to give patients and medical professionals feedback, thus, improving their health management efforts [22]. According to Ref. [23]; technological advancements have resulted in the introduction of wearable devices and smart computing devices, among other digital gadgets, which aid in the remote surveillance and tracking of patients' symptoms, biomarkers, and disease status in a continuous manner. Diabetes research can benefit incrementally from deep learning (DL) when significant continuous glucose monitoring (CGM) data sets become available as connected medical devices, including CGM and injectable data, become more available [22]. Indeed, numerous AI and ML approaches have been used to automate NCD care, including the automation of insulin infusion rates based on CGM data and the recommendation of insulin bolus doses [10]. The development of technologies that digitalize real-time monitoring of prescribed medication doses in real-time medication dosing presents an opportunity to use continuous data sources to analyse MA and dynamically take proactive action as necessary. According to Ref. [21]; real-time dose assessment can be utilised to dynamically predict drug adherence with high accuracy, enabling

proactive clinical intervention to improve health outcomes. Missed once-daily basal insulin doses, for example, in diabetes patients, can be utilised to provide feedback to patients, health caretakers, and clinicians, thus improving their health management [22]. Relatedly [23], indicates that technological advancements have resulted in the introduction of wearable devices and smart computing devices, among other digital gadgets, which aid in the continuous and burden-free remote monitoring and tracking of patients' symptoms, biomarkers and disease status.

Statistics are frequently employed to determine how successfully patients take their medications for chronic diseases such as hypertension [24]. But due to the massive amounts of data being regularly collected in the medical and pharmaceutical industries, predictive models created using ML techniques are being utilised to extract knowledge and generate patterns in the data. Outcome-oriented approaches based on ML are more suited for determining suitable thresholds that best discriminate patients regarding outcomes of interest than human expert-driven or traditional statistical methods [25]. Compared to conventional logistic regression [26], claim that ML models have the advantages of adding nonlinear connections, less biased auto-learning, and better flexibility to prevent over-fitting. The primary distinction between ML approaches and conventional methods is that, in ML, a model is taught by observing real-world data rather than being pre-programmed with predetermined instructions. Computers then learn how to map features to labels using ML algorithms that learn from observations to create a model that generalises the knowledge so that a task may be completed correctly with new, never-seen-before inputs [27].

AI and ML have introduced a paradigm shift in treating chronic diseases, moving away from traditional management tactics and developing tailored data-driven precision care. Data mining, ML, and DL are three branches of AI techniques that have shown promise in many healthcare applications [28]. [14] remarked that AI comprises ML and DL and has shown promising outcomes in evaluating prescribed medicine adherence and enhancing adherence levels, which aligns with their findings [7]. also emphasise the excellent outcomes of applying ML in extracting relevant data from massive amounts of healthcare data for tracking and prediction. Artificial neural networks (ANN) and support vector machines (SVMs), both ML techniques, have shown considerable promise in developing predictive models to enhance medical decision-making. This is especially true when treating diabetes, heart failure, and kidney disease [29–31]. Using SVM [32], assessed patients' adherence to chronic disease therapy [33]. used statistical and ML techniques, such as SVM, LR, and K-nearest neighbour (K-NN), to evaluate hypertension risk successfully. Throughout the development of medical diagnostics, random forest RF has been utilised to create models for predicting the likelihood of developing chronic conditions such as hypertension. According to Ref. [34]; an ML model can be trained using millions of patient charts saved in EMRs with hundreds of billions of data points and no breaks in concentration. In contrast, a human doctor can fail to serve less than 20% of several such patients in a lifetime [27].

Predictive models created using ML techniques are being used to extract knowledge and discover patterns in the available medical dataset. However, this is only possible due to the vast amount of data collected continuously in the healthcare industry. AI areas such as data mining, ML, and DL have shown promise [28] in many healthcare applications [14]. revealed that AI, including ML and DL, has demonstrated promising outcomes in monitoring prescribed medicine adherence and improving adherence levels [7]. further highlight the outstanding performance of applying ML to extract meaningful amounts of data from medical care data for medication management purposes.

Recent evidence has revealed AI and ML's efficacy in the identification of chronic disease status [35–37]; prediction for incidence [38,39], hypertension management with a specific focus on risk prediction based on the intelligent algorithm optimised Bayesian network (Du et al., 2021); prediction of poor outcomes in hypertensive patients [40]; and extreme gradient boosting (XGBoost) for timely intervention in hypertension management [38]. A few literature review-based studies have been conducted on using ML to analyse NCD MA [8,41–44]. [41] identified, summarised, and analysed research on using ML for MA assessments. In their evaluation and analysis of published ML predictive models [42], examined how well patients with cardiovascular diseases adhered to their medications [8]. conducted a narrative review of the scientific literature on AI and AI-aided strategies for tracking and optimising MA in NCD patients [43]. demonstrated how habit was conceptualised in the literature on MA interventions or what impact the key technique described in habit formation theory had in these studies [44]. surveyed the literature on applying ML techniques for inflammatory bowel disease, focusing on how the field has changed over time.

The other systematic review studies on NCD MA do not include the ML aspect [45–48]; Odegard, & Letassy, 2016; [49–53]. In light of the most recent information available, none of the few literature review studies has conducted an extensive systematic literature review that collates papers on ML-based analytics in NCD MA. This includes studies using self-reported MA data, clinical data extracted from EMR, and data derived from remote medication measurements extracted in real-time. In light of this, there was a need for a comprehensive systematic review of the literature on ML-based analytics in NCD MA, spanning the various clinical data input types. So, in the context of this study, we developed the following overarching research question: *What research has adopted ML in the analytics of MA in NCD patients?*

The attempt to answer the major research question led to the formulation of several sub-research questions that are presented as follows.

a) What research has adopted ML-based analytics to measure MA in patients with diabetes, hypertension, cardiovascular, cancer and respiratory diseases?
b) What are the basic data sources and data generation techniques enabling the measurement of MA in NCD patients?
c) What are the common thresholds for measuring MA in NCD patients?
d) How have ML algorithms been applied for analysing NCD MA or non-adherence?
e) What evaluation metrics have been used to measure the performance of ML approaches applied in the analytics of NCD MA?

This work provides a comprehensive, well-informed systematic review of the literature analysing NCD MA using ML techniques such as standard ML algorithms, DL, RL, and ensemble learning. Contemporary research that uses ML approaches and exploits data extracted from EMRs, self-reported MA surveys or remote real-time measures of medication administration and therapeutic drug monitoring (TDM) in NCD patients has been compiled and examined retrospectively. Due to the paucity of literature review studies that holistically look at the application of ML-based analytics in NCD MA assessments, this systematic review study's main contributions are listed below.

- This review provides well-collated literature on MA integrated with ML-based analytics. This is consistent with [41]; who noted the paucity of research incorporating ML to measure how well NCD patients adhere to medication.
- Proffer an up-to-date and thorough systematic literature review on how ML algorithms have been used in MA analytics.
- Specifically, it informs NCD patients, health caretakers, and health practitioners on the accuracy of ML algorithms in measuring NCD MA for informed decision-making.
- Presents an open discussion to validate or invalidate the value proposition of ML-based analytics for measuring MA and appropriating interventions.
- Recommend future research trends and directions for using ML to evaluate how well patients take their NCD medication.

2. Research methodology

This section aimed to ensure that the approach to conducting a solid systematic review was followed to prevent bias and ensure that the review addressed the study's focus properly. It presents how to select and review the literature on ML-based analytics in NCD MA. The practical instructions for conducting systematic reviews in computer science provided by Ref. [54] served as the foundation for this systematic review. As a result, keywords, search strings, electronic search engines, and inclusion and exclusion criteria were established.

2.1. Search for keywords

According to Ref. [55]; keywords are required while searching for research articles in electronic literature databases. As a result, many search phrases and synonyms were used interchangeably to conduct a thorough literature search that directly addressed the study's principal research question. According to Ref. [54]; a systematic review should evaluate the specified research questions and extract the initial keywords, then use related publications from the field of study to identify more phrases and synonyms of the previously identified words. As a result, we used phrases derived from the research topic in our search query, such as "Machine Learning in Medication Adherence," which was deemed synonymous with "Machine Learning in Medication Compliance." "Machine learning in NCD Medication Adherence " was the second keyword used. ML, an acronym for "Machine Learning," was used instead of "Machine Learning." Again, "DL," a subset of "Machine Learning," was substituted for "Machine Learning." Medicine was frequently used as a synonym for "medication" keyword. This method allowed electronic literature databases to generate search variants, allowing for easier access to an extensive collection of research articles. The terms "NCD" and "Chronic Disease" were used interchangeably. According to Ref. [56]; researchers utilise keywords in various ways; therefore, it is critical to locate and use many key search terms to make it simpler to find relevant literature.

To discover papers that used ML only in evaluating MA, all of the search phrases, such as "Medication Compliance", "Medication Adherence", "NCD", and "Chronic Disease", were matched with "Machine Learning" or "Deep Learning." To extend the search, an asterisk was used to discover terms that begin with the same letters. This necessitated adding an asterisk (*) at the end of each search query. According to Ref. [57]; adding an asterisk allows the search to include several terms with similar roots. For instance, "Medic" was used as a root for medication, medicine, medical, medicines, and medications. Adhere* also includes adhering, adherings, adhering, adherence, adherences, and adherent. The search technique was improved by using the Boolean operators "OR" and "AND," as well as truncation, to connect search strings and phrases. This method broadens the search results [56]. also suggested enclosing important search terms in quote marks. The strategy helped to turn search results from digital libraries into more precise search extracts.

2.2. Searching the articles

In order to find the required articles, the search was split into two stages. The first round of searches was done in July 2022, followed by the second phase in August 2022. The second search yielded no articles that were distinct from the first.

2.3. Literature databases

During the electronic literature search, the following databases were utilised: Google Scholar, PubMed, IEEE Explore, and Web of Science. According to Coughlan, Ryan, and Cronin (2015), electronic databases are significant literature sources for scholars and researchers. The four electronic literature databases examined in this study are among the world's most well-known and respected academic literature sources. PubMed has millions of scholarly publications in biomedical and life sciences, with some recent works on health sciences incorporating the ML element. Google Scholar offers research in various fields, including computer science divisions such as AI, ML, and DL. The IEEE Explore gives users access to world-class engineering and technology publications of the best possible quality. The Web of Science is a comprehensive multidisciplinary citation database of scientific and scholarly publications, including journals, conference proceedings, books, and data compilations. Extraction of the essential literature from several digital libraries aided in obtaining comprehensive and relevant literature resources on the subject of interest [58].

2.4. Inclusion and exclusion criteria

A set of criteria for inclusion and exclusion were employed to ensure the extraction and usage of appropriate and relevant literature. Only published publications from credible peer-reviewed journals and conference proceedings were included in the review study. This study included articles published in 2010 or later. The evaluation includes empirical research that is naturally based on observed and measurable events. The research covered in this review is original, from credible peer-reviewed publications and conference proceedings, and is available in full-text online. Table 1 summarises the inclusion and exclusion criteria applied in this review study.

2.5. Eligibility

The targeted publications were searched and refined. The search produced 181 hits [55]. argue that removing duplicates is important in refining search results. As a result, 53 duplicates were deleted. An additional 103 publications were eliminated because they did not match the inclusion criteria outlined in Table 1, with the majority focusing on MA but without ML application. Thus, inclusion criteria examined studies that used at least one ML approach for MA analysis. Finally, 25 were deemed eligible for inclusion in this systematic review study.

3. Comparison with previous related reviews

This section summarises and compares the current systematic review paper to previously published review studies on NCD MA, including ML analytics. Though there have been attempts to review research on assessing or evaluating NCD MA using ML, some of these studies were scoping or narrative reviews rather than systematic reviews. For example [41], only conducted a scoping review to categorise, summarise, and evaluate the literature on using ML for MA-related actions. A scoping review is distinct from a systematic review in that its main objective is to provide a high-level overview of previous work in the

Table 1
Inclusion and Exclusion Criteria applied.

Inclusion	Rationale	Exclusion
Articles published in reputable peer review journals and conference proceedings	Access to current and up-to-date literature	Opinions, keynote speeches, editorials and PowerPoint presentation slides were not considered
Full-text articles	Access to comprehensive and complete research information	Abstracts
Articles on MA or non-adherence and incorporating the ML analytics aspect	Specific need to collate and review related literature on MA that only incorporated the ML analytics aspect	Articles on MA or non-adherence but devoid of ML analytics aspect
Empirical studies	Present closely original data and evidence directly linked to the data source	Other systematic review papers

field of study. In contrast, the current systematic review is more extensive and insightful. The survey by Ref. [8] is a narrative review that merely offers an overview of the state of AI in terms of the evaluation and optimisation of MA in NCD patients. A systematic review is more clinical and superior to a narrative review since it is more likely to be directed by a well-defined clinical or basic research topic or question. It is also more methodologically explicit and less susceptible to bias [59].

[44] systematic review was limited to ML use in the context of inflammatory bowel conditions. The study by Ref. [42] assessed published prediction models that use ML to estimate MA among chronic disease patients. Their study was limited to people with CVDs. Our present review study examines ML applications to MA involving many NCDs. As a result, the current systematic review findings are generalizable to a broader range of NCDs than only diabetes and hypertension. In their research [43], focused on habit development in MA strategies for chronic diseases. The studies did not necessarily incorporate ML applications to MA interventions but used ML to review the required literature. Table 2 summarises the reviews done on the analysis of MA among chronic disease patients incorporating the ML analytics component.

Table 2
Summary of previous reviews on ML and MA in NCD patients.

Source	Publication Date	Title of Article	Remark
Bohlmann et al.	2021	ML and MA.	The paper served as a scoping review that classified, summarised, and analysed publications centred on applying ML for actions associated with MA.
Zakeri et al.	2021	Application of ML in predicting MA of patients with CVDs: A systematic literature review.	This research aimed to identify and summarise the literature on ML models for predicting MA in patients with illnesses associated with CVDs or their primary risk factors. RF, SVM, and neural networks were discovered to be the most frequently used ML algorithms in the existing literature in that review study.
Robinson et al.	2022	An ML-assisted review of the use of habit formation in MA interventions for long-term conditions.	Discussed how habits have been conceptualised in the literature relating to interventions for improving MA or what impact the key technique indicated in habit formation theory has in these studies. A review was conducted with the use of ML.
Babel et al.	2021	AI solutions to increase MA in patients with NCDs	They conducted a narrative review to present research on AI and AI-assisted approaches in measuring and increasing MA in NCD patients. Additionally, the advantages of employing AI and ML to improve MA were discussed.
Stafford et al.	2022	A systematic review of AI and ML applications to inflammatory bowel disease, with practical guidelines for interpretation.	Presentation of a systematic survey of the literature on the use of ML techniques for inflammatory bowel disease, with further emphasis on the evolution of the discipline through time.

Despite efforts to gather and compile studies on ML applications related to NCD MA assessments, only a few of these systematic review studies reported on ML-based analytics in NCD MA assessments. These reviews focused on AI and ML applications to inflammatory bowel disease [44] and the application of ML in predicting MA in CVD patients [42]. The other evaluations were not systematic reviews in the typical sense [41]. study was a scoping review of ML and MA, while the research by Ref. [8] collated studies on AI and AI-assisted solutions in monitoring and improving MA in NCD patients. No author (s) has holistically conducted a comprehensive systematic literature survey that pulls together studies on ML-based analytics in MA involving various NCDs for better generalizability of findings. The current study also uniquely reviews studies on ML applications to MA leveraging the various input data types generated in medication processes: self-reported MA data, clinical data extracted from EMR, and data extracted from off-site, real-time measures of the amount of medication taken and TDM. In light of the shortage of comprehensive systematic reviews on ML analytics applications to MA assessments, as depicted in Table 3, the value proposition for conducting this expansive systematic review was amplified.

4. Analytics pipeline on the use of ML

Fig. 1 depicts a comprehensive diagrammatic representation of the data analytics pipeline using ML

4.1. Data sources

Identifying appropriate data sources is a critical first step in the computational analytics of data that requires ML model development. According to Ref. [64]; data are real-world observations. ML analytical processes ingest data gathered from numerous sources. Historical patient health records acquired from existing databases, such as electronic and patient management systems at healthcare facilities, may be used as data sources in medication management and adherence evaluation [65]. This data could include insurance claims, pharmacy prescriptions, or EMRS, according to Ref. [65]. In terms of measuring adherence, just as in prior studies, this data could consist of data from insurance companies regarding reimbursement claims, pharmacies concerning the distribution of medications, and physicians regarding the prescription of medications [66,67]. Primary data collection methods, such as surveys and interviews, could also be used to acquire the required data. Self-reported measures used by patients, including diaries and questionnaires, are used to collect data on topics such as MA evaluation [68]. Real-time electronic medication monitoring systems, such as medication event monitoring systems (MEMS) and CGMs [69], and embedded internet of things (IoT) devices for real-time monitoring and tracking of medication in patients [7] can also create data. Microchips implanted in prescription bottles record the opening of each bottle in electronic medication monitoring systems [68]. Ultimately, electronic healthcare data, electronic medication monitoring data, and self-reported survey data are all valuable data sources that make ML analytics a reality.

4.2. Pre-processing and feature engineering

Pre-processing is among the most crucial first steps in preparing raw data and features for an ML model. As part of the pre-processing data phase, it is crucial to recognise that real-world data typically contains noise, missing values, and an unsuitable format that cannot be directly used for ML models [70]. Therefore, cleaning and preparing the data for creating an ML model requires the operation of data pre-processing, which also improves the model's accuracy and efficiency. Acquiring the dataset, importing libraries, importing datasets, handling missing

Table 3
Performance metrics for classification.

Metric	Description
Accuracy	The ratio of the total number of accurate predictions to the total number of predictions produced given a dataset. The metric is used to determine how frequently the classifier generates accurate predictions [60]. This metric is suitable for use when there is a reasonably even distribution of weights across the target variable classes in the data. A balanced dataset is one that has a distribution of labels that are roughly comparable to one another.
Precision	Measures the percentage of anticipated positives that turn out to be positive [60]; as such, it is a metric that is used to overcome the limitation of the accuracy metric. When one wants to have a high confidence level in a model prediction, precision is an excellent option for an evaluation metric. The precision metric determines the percentage of correct positive predictions [61]. For instance, one needs to be certain about building a model to predict MA. Otherwise, it may result in poor health management interventions. It is possible to compute it as the "true positive", which refers to the predictions that come true in relation to the total positive predictions (true positive and false positive).
Recall or Sensitivity	A recall is the number of correct positive class predictions made from all correct positive cases in the dataset [60]. However, it sounds very much like the precision metric. The objective of the recall metric is to determine the percentage of true positives that were mistakenly classified. It is computed as the ratio of true positives, also known as predictions that turned out to be accurate, to the total number of positives that were either accurately forecasted as positives or wrongly projected as negatives (true positive and false negative).
Specificity	Specificity is a measure that determines a model's ability to estimate true negatives for each of the available categories. Therefore, specificity refers to the number of true negatives the model may accurately predict [60].
F1-Score	The F1-score is a metric that is used to evaluate how accurately a model represents a dataset. The F-score, or the F1-score, is used to evaluate a binary classification model, and this evaluation is based on the predictions that are supplied for the positive class. F-Measure provides a single score that addresses concerns about both recall and precision in a manner that balances the two to produce a single score. Consequently, the F1-score can be calculated by finding the harmonic mean of both precision and recall and assigning the same weight to each variable [60].
AUC	AUC is a performance metric that can be used for classification tasks at positive threshold levels [60]. It indicates how well a model can differentiate between classes on a particular dataset. The AUC will be higher if the model is more accurate at detecting when a 0 is truly a 0 and when a 1 is truly a 1, leading to higher overall model accuracy. In the context of MA evaluation, a model is said to have a higher AUC if it can accurately differentiate between patients who follow their medication regimen and those who do not.
Confusion Matrix	This provides a comprehensive overview of the performance of a prediction model and is a widely used visualisation tool for demonstrating an algorithm's accuracy, sensitivity, and specificity [62]. The metric provides a detailed illustration of the accurate and inaccurate classifications that apply to each class. Using a confusion matrix can be helpful when attempting to differentiate across classes. This is particularly true when the cost of misclassification differs between the two classes or if one class has much larger test results than the other. Making a false positive diagnosis of an illness, for example, as opposed to a false negative diagnosis of the same condition, has distinct ramifications in healthcare management.
Logarithmic Loss (Log-Loss)	There are a few other names for log loss, including cross-entropy and logistic regression loss. In essence, it tests how well a classification model works, where the input is a probability value between 0 and 1 and the parameters are described by probability estimates. It is possible to differentiate it precisely so that it may be understood more readily. Since it is a precision measurement, probability can be utilised as a measurement of confidence due to its nature.

Model evaluation metrics for regression are used to handle the problem of discovering associations between dependent and independent variables in supervised ML. The performance of a regression model can be evaluated based on its prediction errors. Metrics such as the root mean square error (RMSE), mean absolute error (MEA), mean squared error, R squared score and modified R squared are the ones that are most commonly utilised to assess the performance of regression models. The RMSE is a statistical metric for determining how well a regression line corresponds to data points [63]. A fundamental statistic, mean absolute error (MAE), measures the absolute difference between actual and projected values [63]. "Absolute" refers to reading a number as positive. The term "mean squared error" refers to the average of the squared differences between the model's projected and actual values. In a regression model, the amount of variance in the dependent variable that the independent variable can account for is determined by the R squared value, also referred to as the coefficient of determination in some contexts. The problem with R squared can be circumvented by employing adjusted R squared, which invariably produces results that are less significant than those produced by R squared. This is because it adjusts the values of the increasing predictors and only reports progress when a change has occurred.

values, imputation, dataset splitting, exploratory data analysis (EDA), encoding categorical data, feature scaling, feature creation, and selection are all pre-processing activities and feature engineering. As a result, once the data sources have been identified, the necessary datasets must be acquired for use. Some datasets are freely accessible online from sources such as Kaggle, the Global Health Observatory, and the UCI ML Repository. Other datasets can only be accessed and utilised with the data source owners' explicit and official agreement and approval. For example, a specific healthcare facility's EHRs can only be accessed and used if the institution involved grants the researcher the necessary access and usage privileges. Nonetheless, key libraries are imported into a data analytics application to enable various analytical procedures on the data as part of the pre-processing process. In Python, for example, libraries such as Pandas, Numpy, Matplotlib, and Plotly enable operations such as data importing and EDA. The datasets are imported after they have been granted access, and all relevant libraries are then imported. The EDA approach is a critical pre-processing step for both predictive and descriptive algorithms, and aids in the preliminary investigational analysis of data by extracting averages, means, minimums, and maximums (among other descriptive statistics), discovering patterns, spotting outliers and anomalies, finding missing values in the data, testing hypotheses, and checking assumptions by visualising data in graphs such as box plots, scatter plots, and histograms. The ML pipeline is given in Fig. 2.

Missing values are addressed as part of the pre-processing stage by eliminating rows or columns with missing values or through imputation. Imputation is a way to keep most of the data in the dataset by substituting a value instead of missing data [71]. This method might replace missing values with the mode (the most prevalent value in other records) for missing categorical values and the means of combined column values for missing numerical values [72]. Additionally, outliers should be handled during data pre-processing. Outliers are abnormally low and high values in the dataset that could otherwise result in model measurement errors, biassed parameter estimation, and inaccurate findings [73]. If these outliers are not addressed, they have a negative impact on model performance and, therefore, must be dealt with appropriately. Outliers could be addressed by removing outlier-containing records or treating outliers as missing values and replacing them using the appropriate imputation method [73]. Splitting the dataset into training and test sets is another step that must be completed in the data pre-processing stage. The training set is a subset of

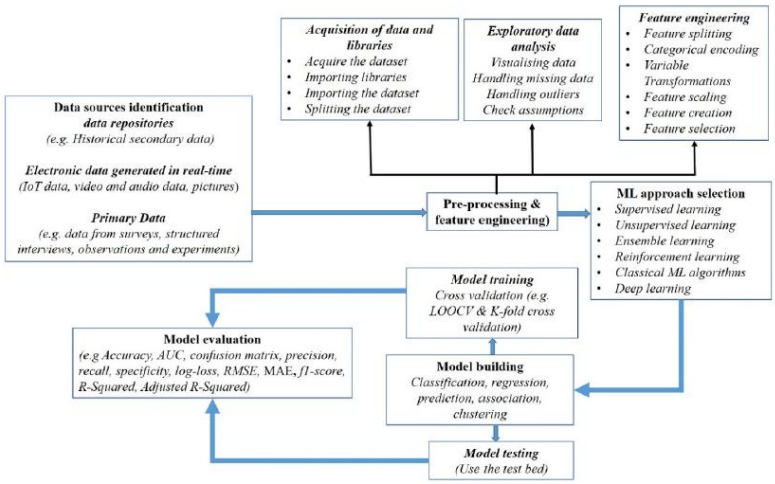


Fig. 1. Analytics pipeline on the use of ML.

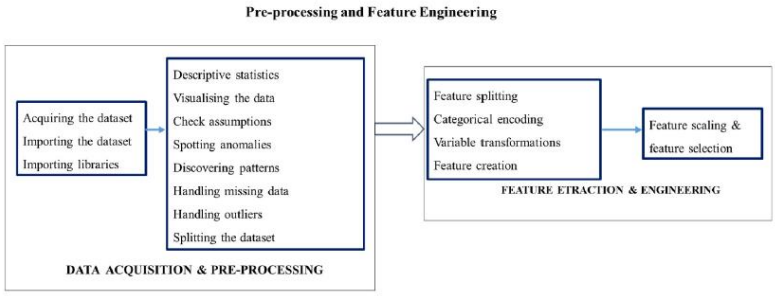


Fig. 2. Procedure of data pre-processing, feature extraction, feature engineering, feature scaling and selection by the ML model.

the full dataset utilised to train the ML model's analytics capabilities. The test bed, a subset of the entire dataset, is used to test the ML model to predict, classify, or cluster the output. Activities for pre-processing data lead right into feature engineering, which includes splitting features, specific encoding, scaling and transforming features, creating features, and feature selection.

The process of transforming unstructured data into features that may be utilised in constructing an analytics model through ML or statistical modelling is called "feature engineering." This is consistent with the definition of feature engineering offered by Ref. [64]; who describes it as extracting features from raw data and converting them into formats suitable for ML models [64]. define feature engineering as extracting and converting features from raw data and converting them. A feature is an attribute or variable that defines some element of particular data items. The data's most appropriate and significant features are determined for model construction by deleting unnecessary data and redundant features. This method can lower the number of variables in a model while maintaining its predictive power. This lessens the computation time and complexity required to generate the model [74,75]. Another advantage of deleting extraneous features and data is enhancing the model's accuracy [76]. Categorical variables should be encoded as numbers since numbers are usually easier for an algorithm to understand. This resonates with [77]; who argue that ML algorithms operate best with numerical inputs; thus, categorical variables must be encoded into numerical values using encoding techniques. One-hot encoding (OHE) is a method for pre-processing categorical data before it is used in ML models. This method changes categorical values into new binary features and gives them 1 and 0 without losing information [78].

One of the last activities of data processing is feature scaling and transformation, which involves minimising variables that dominate other variables by employing standardisation or normalisation procedures to change the dataset's independent variables into a specific range. Feature scaling and transformation are typically conducted during data pre-processing to manage widely changing magnitudes, values, or units, with the benefits of facilitating faster model training and boosting model performance [79]. For example, all variables may be scaled to new values ranging from −1 to 1. The logarithmic transformation is a commonly used transformation technique that is used to compress larger numbers while comparatively expanding smaller ones, transforming the dataset into a normally distributed set [80]. This leads to less skewed data points, particularly in the case of heavy-tailed distributions. Another variable transformation is the square root transformation, which changes a dataset by replacing each value for an independent variable with its square root. Most of the time, the different levels of the independent variable in question have the same variance after a square-root transformation.

The categorical splitting of dataset features into two or more additional features aids algorithms in better understanding and learning the patterns in the dataset. Splitting characteristics into sections might sometimes improve their value toward the goal to be learned. For example, a string variable containing both date and time could be separated to form a sub-feature containing only the "Date" because the

"Date" feature contributes more to the target function than the date and time combined. Apart from feature splitting, feature creation could be accomplished using mathematical operations such as determining differences by subtracting two values, aggregations, multiplication of two values, sums, divisions, means, and medians, allowing the generation of additional features. The existing dataset is thereby enriched in this way. Even though these features come straight from the given dataset, they can affect how well the model performs if they are carefully chosen to relate to the target.

To identify the subset of the most significant features to include in a model, redundant, irrelevant, or noisy features are removed from the original feature set using different feature selection techniques. As shown in Fig. 3, the two basic feature selection strategies are supervised feature selection and unsupervised feature selection [81]. The target variable is not taken into account in unsupervised feature selection methods, which can be used with unlabelled datasets.

The wrapper method is based on a specific ML algorithm that the researcher or data analyst is attempting to fit to a given dataset; it uses a grid search methodology to evaluate all potential feature combinations against the evaluation criterion. The wrapper feature selection technique is a search issue in which multiple combinations are created, assessed, and compared to other combinations. It trains the algorithm by iteratively employing the subset of features until an optimal set of features is found.

Rather than feature selection based on cross-validation performance, filter methods select features based on statistics. This method does not rely on a learning algorithm and instead selects features as a preprocessing phase. The variables are chosen based on their performance in various statistical tests for their relationship with the target variable [81]. The filter approach removes unimportant variables and redundant columns from the model by rating them using various criteria. For example, a chi-square test or correlation value between every input feature and the outcome variable targeted in the research is calculated, and the desired number of variables with the best chi-square or correlation value is chosen. Regarding the Fisher's exact test, the variables' rank on the Fisher's criteria is provided in descending order, and then the variables with a high Fisher's score are chosen. A chosen metric identifies irrelevant qualities and performs recursive feature selection. Filter methods are either univariate, in which an ordered ranking list of features is generated to influence the final selection of a feature subset or multivariate, in which the features are evaluated for relevance, thus detecting redundant and irrelevant features. Compared to wrappers, filter techniques are faster and more generalizable [82].

By considering feature interaction and low computational cost, embedded approaches incorporate the benefits of filter and wrapper methods [82]. The variable selection ML algorithm is part of the learning algorithm in embedded feature or variable selection approaches. This lets classification and feature selection occur at the same time. The features that contribute most to each model training iteration are carefully extracted. Common embedding approaches include RF feature selection, decision tree feature selection, and least absolute shrinkage and selection operator (LASSO) feature selection [83]. used the LASSO regression analysis method to perform both variable selection and regularisation to improve the prediction accuracy and interpretability of the resulting model in a study that evaluated and compared

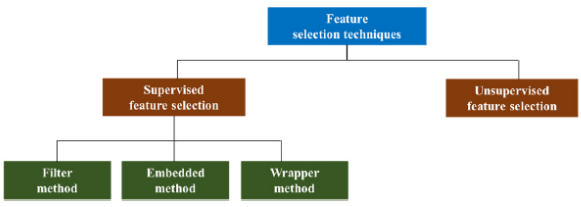


Fig. 3. Feature selection techniques.

three predictive analytic approaches to characterise medication non-adherence.

It is vital to highlight that the most significant variables or features in creating an ML model are chosen when feature selection is applied, which has the advantage of lowering the computational cost of modelling, minimising overfitting, and improving the predictive ML model's performance. As highlighted in Section 4.3 below, several ML techniques could be utilised to handle regression, classification, clustering, association, and forecasting problems.

4.3. ML approaches

Effective ML approaches must be identified and implemented along the ML analytics pipeline to handle the existing real-world challenges and difficulties. The branch of AI known as ML refers to the ability of machines to learn from data, improve their performance based on previous experiences, and make predictions. ML employs a diverse range of algorithmic approaches to processing huge amounts of data to make predictions of output values within a reasonable range. These algorithms learn from the supplied data, which is incorporated into the model and used to carry out a specific analytics task. The most prevalent approaches to ML include supervised, unsupervised, semi-supervised, and RL [84,85]. These approaches could be explained in the form of standard or classical ML algorithms, DL algorithms, and ensemble learning.

Supervised Learning Algorithms: The foundation of supervised learning algorithms is supervision, as the name suggests. This means that in the supervised learning technique, machines are trained using the "labelled" dataset, and the machine predicts the output based on the training [85]. In other words, the values of the input variables and their matching output variables (labels) are known in advance. More significantly, after training using the input and associated output, the machine should be able to predict the output using the test dataset. The supervised learning technique's fundamental purpose is to map the input, or independent variable, with the outcome or dependent variable. Classification and regression algorithms are two types of supervised ML algorithms, depending on the situation at hand. The ML algorithm must form a conclusion from observed values and decide which category new observations belong to in classification tasks. For example, when categorising MA behaviour as "adherence" or "non-adherence" based on current medical records, the algorithm must categorise medication-taking habits appropriately. Classical supervised ML techniques that are widely used include RF, SVM, LR, and decision tree algorithms. Regression algorithms, on the other hand, are used to handle regression situations in which the input and output variables have a linear relationship. Regression analysis is particularly beneficial for prediction and forecasting since it focuses on one dependent variable and a sequence of other changing variables. These algorithms are used to predict continuous outcome variables, such as patient medication-taking patterns or forecasting prescription refill adherence using EHRs and dispensation data, as [86] have done in past research. The most popular regression techniques are the simple LR, multiple regression, and LASSO regression methods.

Unsupervised Learning Algorithms: Unsupervised learning does not require supervision, in contrast to learning that is supervised. Consequently, in unsupervised ML, the computer learns from an unlabelled dataset and makes independent predictions about the results on its own [87]. The values of the input variables are known in this type of ML technique, but there are no associated values for the output variables. Thus, the unsupervised learning algorithm's main goal is to group or categorise the unsorted dataset based on similarities, patterns, and differences. The ability of unsupervised learning algorithms to generate hidden patterns from underlying structures in input data makes them so versatile [88]. Clustering and association are two tasks that fall under unsupervised learning [85]. Clustering is the process of grouping comparable data sets based on defined criteria. It is useful for segmenting data into groups and analysing each data set to uncover trends.

Traditional unsupervised ML methods include the self-organizing map (SOM), K-NN, and the K-means clustering techniques. In contrast, association rule learning discovers significant relationships between variables in a huge dataset. The primary goal of this learning technique is to determine the dependence of one data item on another and map those variables accordingly. Apriori and FP-growth algorithms are two of the most well-known ways to find association rules.

Semi-Supervised Learning Algorithms: Both labelled and unlabelled data are used in semi-supervised learning. Labelled data has meaningful tags, so the algorithm can interpret it, whereas unlabelled data does not. ML algorithms can learn to label unlabelled data using this combination. The basic goal of semi-supervised learning is to use all data efficiently, rather than only labelled data, as in supervised learning.

Reinforcement Learning (RL): RL is a feedback-based method in which an AI agent or software application explores its surroundings automatically by hitting and trailing, taking action, learning from experiences, and improving its performance. RL is the most similar to human learning. There is no labelled data in reinforcement learning, as in supervised learning; agents learn solely from their experiences. The algorithm or agent learns by interacting with its surroundings and receiving a positive or negative reward [85]. As a direct consequence, the agent will receive rewards for positive actions and punishments for negative ones. The RL agent aims to earn as many rewards as possible [89]. Following the definition of the rules, the ML algorithm attempts to explore several options and possibilities, monitoring and assessing each output to determine which is ideal. RL instructs machines through trial and error. It learns from previous experiences and changes its approach to problems to get the best outcome. Based on this concept, a prior study was done to examine whether using RL to optimise response and adapt the involvement of engaged patients may optimise adherence to diabetes medications [90]. Participants in that study who were having trouble keeping their diabetes under control and were already taking oral diabetic medication were split into two groups at random: one received the reinforcement-learning intervention. At the same time, the other served as the control. Participants were given electronic pill bottles to use as part of the intervention, and those assigned to the intervention arm received up to daily SMS. After that, an RL prediction algorithm was employed to personalise each message using daily pill bottle adherence statistics individually. So, RL could be used to help solve problems in the real world by giving important insights on a large scale. Some examples of RL algorithms are deep adversarial networks, Monte Carlo RL, and Q-Learning (Kim, 2020).

Deep Learning (DL): The subfield of ML and AI, known as DL, is designed to mimic humans and their actions to make effective decisions. Deep neural networks, also known as DNNs, are one of the many breakthroughs made in AI, and they stand out as a particularly promising expansion of the shallow ANN structure [91,92]. DL algorithms are layered in a hierarchy of increasing complexity and abstraction, as opposed to traditional ML algorithms, which are linear. DL is essentially a three- or perhaps more-layer neural network [93]. The building blocks of a deep neural network are layers upon layers of interconnected nodes. Deep neural networks are built with multiple layers, each improving and optimising the prediction or classification computations. According to Ref. [89]; the DL method splits the data into several layers, each of which can gradually extract features and pass them on to the following layer in the hierarchy. Calculations go forward via a network according to a process known as forward propagation. The layers that are visible in a deep neural network are the ones that are located at the input and output vertices. After completing the data processing in the input layer, the DL model's output layer is responsible for making the ultimate prediction or classification. The subsequent levels each take their input from the one before them, which means that the layer before them must have produced some output. Back propagation is another strategy that can be used. This method generates prediction errors by employing methods such as gradient descent and then adjusts the weights and biases of the function to train the model by recursively iterating through the layers [94]. A neural network can generate predictions and correct errors by integrating forward and backpropagation. The DL algorithm gradually becomes more accurate over time. Multilayer perceptrons (MLPs), SOMs, RNNs, CNNs, and LSTMs are all examples of DL methods. With the help of a diagram, Fig. 4 shows how a DL architecture in the form of a deep neural network has input, hidden, and output layers.

One of the most significant advantages of DL is its capacity to deal with complex challenges, the resolution of which requires locating obscure patterns in the data and possessing a deep understanding of the complexities underlying the relationships between a large number of interrelated variables [95]. Deep neural networks are comprised of multiple layers, each of which enables models to learn more complex features and carry out computationally demanding tasks more effectively. This allows the models to execute multiple complex operations concurrently. It outperforms ML for machine perception tasks, which need it to interpret unstructured datasets like images, sounds, and videos in the same way that a human would. The fact that DL algorithms can eventually learn from their errors explains why this capability exists. It can check the accuracy of its forecasts or outputs and make any necessary adjustments due to its capabilities. On the other hand, conventional approaches to ML require varying degrees of human participation to determine whether or not the output is accurate. DL generally functions most effectively with a huge dataset. Because of the enormous amounts of data analysed, DL is by far the most accurate subset of ML [96]. DL is exceptional in that its performance is unaffected by adding additional data to the model. Rather, DL models may be trained on enormous data and improve as more data is added to their training set. DL is extremely scalable due to its capacity to analyse big data and conduct various computations quickly and cost-effectively. This directly impacts productivity since it allows for more rapid deployments or rollouts, increases modularity, and enables using trained models for a wider range of problems. In comparison, legacy algorithms fail to increase the performance once a certain stage has been reached. The creation of more helpful decision rules by DL algorithms is made possible by merging the patterns that are discovered in the data.

DL excels at solving complex problems such as image labelling, natural language processing, and speech recognition, naming just a few examples of the challenges it can effectively address [95]. Most ML algorithms struggle to analyse unstructured data, resulting in less utilisation of this type of data. It is common to practise for traditional ML algorithms to have limits when it comes to analysing unstructured data, which is unfortunate because unstructured data is a significant source of knowledge. Apparently, DL has the most impact in this particular domain. DL models can be trained to optimise practically every function of an organisation, including medical institutions if they are provided with unstructured data and proper labelling. Another important advantage of a DL approach is its capacity to carry out feature engineering independently. It searches through the data to locate features that are correlated with one another. Then it combines those features to

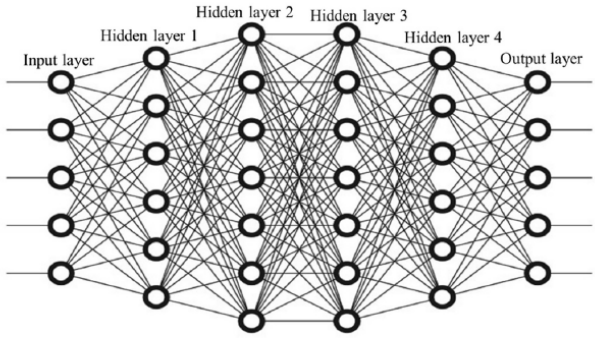


Fig. 4. A six-layered DL architecture.

enhance speedier learning without being specifically instructed. Without any further guidance from humans, DL algorithms can generate new features by focusing on a small subset of the characteristics included in the training data. This demonstrates that DL can complete complex tasks normally requiring a significant amount of feature engineering. This means organisations can roll out applications or technologies more quickly and precisely. Even though their feature extraction capabilities necessitate additional training data, this boosts their ability to comprehend sophisticated and distinctive patterns across a wide range of data classes [97]. Compared to ML, DL is computationally more demanding and hence requires more powerful processors and longer processing times [98]. Because of the vast amounts of data required for their training, DL techniques are inappropriate for using phenomena with relatively small datasets [99]. As a direct consequence, researchers and data analysts often use traditional ML algorithms, despite the drawbacks of these approaches. Classical ML may be better for feature engineering tasks that are relatively simple and do not involve the analysis of unstructured data [95].

Ensemble Learning: One of the most powerful ML approaches, ensemble learning, employs the combined output of two or more models (weak learners) to handle a specific computational intelligence task [100]. Ensemble models are better at making predictions than single models because they combine the results of many individually trained supervised learning models and use those results in many different ways. Among the most prominent ensemble learning approaches are bagging, boosting, and voting. Bagging, which is also known as bootstrap aggregation, is an approach that combines the predictions of multiple models, each of which was trained on its own individual set of randomly generated training data, to improve the accuracy of predictions and lower the amount of model variation [100]. The final output of the ensemble model will be determined by taking the average of all of the individual estimators' predictions. The RF is a good example of ensemble learning partly because it is made up of many different decision trees. The other ensemble method, boosting, allows each member to learn from the mistakes of the previous member and generate better predictions for the future. In contrast to the bagging approach, all weak base learners are grouped in a sequential sequence in boosting to learn from the mistakes of their preceding learners. As a result, all poor learners are transformed into strong learners, resulting in a superior predictive model with dramatically increased performance. AdaBoost is one such boosting algorithm. Voting generates multiple models of various types, and the predictions are combined using some basic statistics, such as computing measures of central tendency (the mean or median). The final projection will incorporate this prediction along with additional data. Voting in ensemble learning aggregates the results of each classifier fed into the voting classifier and predicts the output class depending on the majority of votes. Instead of making separate models and ascertaining their accuracy, a single model learns from different models and predicts output based on which output classes have the most votes.

4.4. Model evaluation

One of the most critical aspects of developing an effective ML model is evaluating its performance. Model evaluation is the process of analysing the ML performance of the model, as well as its weaknesses and limitations, using various assessment criteria. Evaluating a model is necessary for determining the effectiveness of a model in any ML-based research work, and it also has a part to play in the process of monitoring models. Measurements of this kind are referred to as performance metrics or evaluation metrics, and they are used to evaluate the quality of the model or its performance. Given the available data, these evaluation metrics provide insight into how the ML model performed. It is possible that the performance of the model can be improved by adjusting the hyper-parameters. Each ML model has the goal of performing well when applied to data that has not been seen or used before, and performance metrics are used to determine how well the model generalises when applied to the new dataset. Most of the time, problems in ML are classified into classification or regression. As a result, not all metrics can be applied to every type of problem. Both regression and classification tasks require different assessment criteria. The most typical measures to evaluate a model's classification performance are accuracy, precision, recall, and specificity scores, the F1-score, confusion matrix, logarithmic loss, and area under the curve (AUC). It is critical to note that the AUC, ROC-AUC, or AUROC are interchangeably used in the existing literature. Table 3 contains more information regarding these measures.

5. Recent work on ML application to NCD MA

This section presents recent work on the application of ML in the analytics of MA in patients with various NCDs (hypertension, diabetes, CVDs, respiratory diseases, and cancer, among others). The articles reviewed in this study are summarised in Table 4.

5.1. Recent work on ML analytics application to hypertension MA

Three articles on analysing hypertension MA using ML-based analytics have been extracted from the existing literature [86,103,104]. [86] focused on predictive and pattern analysis of prescription refill adherence using EMRs and dispensing data. The research produced medication-taking prediction models using four well-known ML algorithms: RF, LR, GB, and K-NN. K-means clustering was used in the research to identify consistent PDC patterns over two years. The performance of the model was tested using three different metrics: precision, recall, and the AUC or ROC-AUC statistic. When baseline predictors were employed and history information was provided by incorporating features of earlier prescriptions, the RF and GB algorithms had the best outcomes on both the validation and test sets. This was the case regardless of which set was being used. In the temporal split scenario, when patients with only one prescription were eliminated from the test set, the best model, RF, had AUCs between 0.90 and 0.91. This was the case when using baseline and, correspondingly, baseline predictors plus history.

[103] examined whether it would be feasible to apply ML techniques such as RF, ANN, SVR, and SOM to identify and determine characteristics associated with the adherence levels of hypertension patients from a tertiary hospital in Kuala Lumpur, Malaysia. Using the backwards elimination approach with RF, the study chose features from the ranked variables strongly correlated with the patients' adherence levels. Their analysis compared the ability of RF, ANN, and SVR to accurately predict patients' adherence to their hypertension medication using the identified characteristics. In order to evaluate how well the ML models were performing, the RMSE was used to calculate the values that were specified for them: 1.53 for RF and 1.55 for SVR. According to the Wilcoxon signed ranked test, there was no significant difference between the actual scores and the predictions the ML models generated. The RF variable importance technique found that education level, married status, general overuse, monthly income, and specific issues were the most important variables.

[104] used decision trees to construct a clinical prediction model of MA in hypertension patients using data from a Chinese hospital. This model estimated patients' likelihood of taking their prescribed medications. The PDC of the prescribed antihypertensive medications served as the study's evaluation criterion. The study retrospectively analysed patients' adherence to hypertension drugs based on Intelligent Chronic Disease Management System data. Based on a ROC-AUC of 0.810, the adherence prediction decision tree model predicted compliance to antihypertensive medications with 0.78 sensitivity and 0.69 specificity. According to the study, an adherence-predicting model could be employed in community-based hypertension care.

Table 4
Summary of recent work on ML analytics application to MA.

Details of Source	Title	Method/Structure	ML Approach	Results	Strengths (+)/Limitations (−)
Author(s): Lo-Ciganic et al. Source: Medical Care Year: 2015	Using ML to examine MA thresholds and risk of hospitalisations.	• Between 2007 and 2011, examinations were performed on 33 130 patients with TD2 who were identified using administrative claims data from the Pennsylvania Medicaid program.	Random Survival Forest (RSV)	• The most discriminating adherence levels for hospitalisation for any reason ranged from 46% to 94%.	(+) The model's prediction error was above 25% since it didn't contain clinical and social-behavioural information like HbA1C, which is connected to adherence and hospitalizations.
		• The types of data that were gathered include Medicaid claims and encounter data for outpatient, inpatient, long-term care, and professional services, as well as prescription drug claims with details such as fill date, quantity dispensed, days of supply, and prescriber information.	Survival Tree (ST)		(−) The Medicaid population in Pennsylvania could not be generalised to other Medicaid populations or commercially insured populations due to the differences in demographics and programmatic characteristics between the groups.
		• The study employed ST and the probabilistic hazard model to find hospitalisation predictors and adherence levels that successfully differentiate hospitalisation risk and PDC.			(+) It was anticipated that the Pennsylvania Medicaid programme would be fairly representative due to the gender distribution of the programme (42% men), which is comparable to Medicaid across the country, as well as access to and utilisation of healthcare comparable to Medicaid across the country.
Author(s): Mohebbi et al. **Source:** *IEEE* **Year:** 2017	A DL approach to adherence detection for T2D.	• Using a modified version of the MVP model, originally made for T1D patients, it was possible to simulate a wide range of CGM data for T2D patients. Different classification algorithms were evaluated with the help of these signals through a comprehensive grid search.	LR	• The accuracy of the classical LR model was 65,2 ± 0.8% better than that of random, whereas the highest performing models were created using DL, which had an accuracy of 77.5 ± 1.4% with CNN and 72.5 ± 3.5% with MLP.	(−) Based on the study's findings, it was recommended that, in the future, when access to a considerable amount of actual CGM data becomes available, the feasibility of patient-specific detection systems based on DL models be examined. Concerns are raised due to the study's use of limited data.
			MLPs CNN	• CNN achieved the highest results, achieving an average classification accuracy of 77.5%.	(+) When comparing the classification algorithms, a thorough grid search was utilised. This is a method for locating the optimal model hyper-parameters that make predictions as accurate as possible.
Author(s): Chen et al. **Source:** *Diabetes* **Year:** 2019	ML application to predict patient risk of non-adherence in T2D management using U.S. claims databases.	• To predict adherence, ML models were trained on 111 180 T2D patients who were beginning metformin monotherapy. The amount of time covered by the baseline (pre-index) data for eligible patients was six months, while the amount of time covered by the follow-up (post-index) data was two years.	Random Forest (RF)	• RF classifiers that employed a training set with a random split of 80% and a test set with a random split of 20% had an accuracy of 0.73 and a sensitivity of 0.73 in the early investigation of the data.	(+) A highly specific cohort of non-adherent T2D patients was employed instead of a generic one. As a result, the findings directly and abundantly address the specific issue of non-adherence.
		• PDC $\geq$ 0.8 indicates good adherence. Age, gender, race, T2DM-related medications and procedures, other health problems, and metformin use were included in the model.	XGboosting		(+) The dataset for prediction comprised clinical data that had not been investigated extensively, namely the duration of unadjusted hypoglycaemic medication.
			BART and super learner LR		(−) The investigation was conducted in a single centre with small sample size, limiting the generalizability of the findings. External validation will require a large-scale, forward-thinking, multicentre investigation

(*continued on next page*)

Table 4 (*continued*)

Details of Source	Title	Method/Structure	ML Approach	Results	Strengths (+)/Limitations (−)
Author(s): Wu et al. **Source:** *BMJ Open Diabetes Research and Care* **Year:** 2020	Predictive models of medication non-adherence risks of patients with T2D based on multiple ML algorithms.	• The data for this research were extracted from the EMRs. During the period between April 2018 and March 2019, a face-to-face questionnaire survey was carried out in the outpatient clinic of the Sichuan Provincial People's Hospital. In total, 401 people took part in the study. • Face-to-face questionnaires were used to collect information about patients with T2D, including their demographics, disease and treatment, diet and exercise, mental health, and degree of treatment adherence.	Bayesian network Neural net SVM LR K-NN LSVM RF C 5.0 model Tree-AS CHAID Quest C&R Tree Ensemble model	• The AUC values for the created modelling algorithms ranged from a low of 0.557 (SD 0.051) to a high of 0.866 (SD 0.082), with the best approach being the ensemble model, which comprised five models and utilised oversampling to balance the data after data imputing but without data binding.	(+) A dataset taken from the real world served as the basis for the variables that were collected and investigated concerning MA. The predictive accuracy of the model improved as a result of this, despite the fact that the sample size was only small. (−) When the authors examined the sample size, they discovered no inflexion point in the AUC curve as the sample size increased. This was something that the authors acknowledged. This suggested that a larger sample size was still necessary for the investigation.
Author(s): Gu et al. **Source:** *IEEE* **Year:** 2021	Predicting injectable MA via a smart sharps bin and ML.	• Over the course of three years, 165 223 records about injection disposal were collected from 5915 different HealthBeacon units.	Extremely Randomized Trees	• With a ROC-AUC of 0.86, the suggested ML approach demonstrated very strong prediction performance.	(+) Multiple weak learners and the fusion of multiple different types of ML classifiers were used to improve the accuracy of predictions and lower the risk of overfitting.
		• HealthBeacon Ltd.'s "SSB," a connected IoT gadget, was utilised to track and monitor injection disposal at patients' homes.	RF	• Predicting a patient's likelihood of missing a prescription medication on time with an accuracy of 81.3% was obtained using HealthBeacon SSB data.	(+) In the 10-fold cross-validation method, grid search and random search were utilised to get the optimal values for the model's hyper-parameters.
		• The study utilised an architecture called "majority voting," in which the majority of the predictions made by the five models would be used as the final answer.	XGBoost Gradient Boosting and MLP through Ensemble learning	• Furthermore, the recall/sensitivity from the confusion matrix is 91%, indicating that 91% of the prediction was correct for individuals who took medication on time.	
Author(s): Thyde et al. **Source:** *Journal of Diabetes Science and Technology* **Year:** 2021	ML-based adherence detection of T2D patients on once-daily basal insulin injection.	• A group with type 2 diabetes receiving once-daily insulin infusions were modelled using simulated CGM data.	CNN	• Each of the three expert-engineered, feature-based classification models achieved an average accuracy of 78.6%, 78.2%, and 78.3%, respectively.	(+) Use of real-time clinical research data aided the generation of accurate results based on the originally collected data.
		• The study simulated CGM data from people with type 2 diabetes labelled adherent with their once-daily insulin injections or non-adherent. The well-known and T2D-modified MVP model was used to simulate the in-silico CGM data, which showed how plasma glucose levels changed.		• Both classification models that incorporated expert-engineered learned attributes achieved an average accuracy percentage of 79.7%. The average accuracy of the two classification models, each of which uses both expert-designed and learned features, was 79.7% and 79.8%, respectively.	(−) Due to the fact that the adherence rate was only 95%, the findings may only be applicable to a certain group of T2D patients who are extremely adherent. In clinical settings, levels of patient adherence can vary substantially, highlighting the importance of conducting additional studies into different types of patient adherence.
Author(s): Lauffenburger et al. **Source:** *British Medical Journal* **Year:** 2021	RL to improve non-adherence for diabetes treatments by optimising response and customizing engagement (REINFORCE): a pragmatic randomised trial study protocol.	• Brigham and Women's Hospital in Boston served as the research site. Adults with T2D aged 18 to 84 who took 1–3 oral medications daily and had an HbA1c level of 7.5% were eligible. EHR data were used to analyse these criteria. Patients who met the criteria for using electronic pill bottles had smartphones with internet data plans or wi-fi at home, and their desire to use them was observed. • RL, or a control group, was randomly allocated to 60 people with suboptimal oral	Reinforcement Learning	• Findings showed a 10% disparity in average adherence between the groups. Six months of monitoring were performed. This result was obtained using the following parameters: significance level = 0.05, power = 0.8, and standard deviation = 12.5%. Also, it would pick up on a difference in diet adherence of 50% and an HbA1c difference of 1% (assuming SD = 1.3).	(+) In this study, long-term drug use and clinical effects such as glucose management were investigated by employing a 6-month follow-up to evaluate patient adherence and the study's findings. (+) A pragmatic randomised trial with two arms called the RL to improve non-adherence

(*continued on next page*)

Table 4 (*continued*)

Details of Source	Title	Method/Structure	ML Approach	Results	Strengths (+)/Limitations (−)
		diabetes management. The computerised pill bottles and daily text messages were distributed to the intervention arm. An RL prediction system based on pill bottle adherence was used to personalise messages. A validated three-item self-report measure and the follow-up survey determined self-reported adherence.			for diabetes treatments by optimising response and customizing engagement trial tests a highly scalable approach to personalise communication using RL to improve MA and diabetes control. The clinical trial's goals were to increase the percentage of diabetes patients who successfully control their condition and reduce the number of diabetic patients who fail to take their medicine as prescribed. (−) Although there is a possibility that monitoring could affect adherence, electronic pill bottles are quite dependable in terms of capturing the amount of medication that is really consumed. Nevertheless, observer effects generally disappear over time and would be comparable in the control and intervention arms.
Author(s): Galozy & Nowaczyk **Source:** *Journal of Biomedical Informatics* **Year: 2020**	Prediction and pattern analysis of medication refill adherence through HER and dispensing data.	• The data repository known as Region Halland (RH) was searched to obtain data regarding patients treated at public primary, secondary, and specialised care facilities in Halland, Sweden. • Patients diagnosed with essential hypertension (ICD10 code i10) at any point between November 1, 2014, and July 31, 2018, were considered to be a part of the group. The patients included in the study range from 18 to 90 years old. As of October 2018, records from 545 652 patients who made 7 548 301 visits were included in the data collected from 29 care facilities (24 primary care, 3 hospitals, and 2 emergency care). Medical services were the primary area of concentration, including the frequency and purpose of visits, information regarding earlier prescriptions and PDCs, and a few demographic details pertaining to both the patient and the prescriber. • The major metrics utilised to evaluate model performance were the ROCAUC or AUC, precision, and recall.	K-Means RFs (LR) Gradient Boosting (GB) K-NN.	• In the temporal split scenario, the best model, RF, produced AUCs of approximately 0.90 and 0.91 using the baseline predictors and baseline plus history, respectively. • In contrast, the temporal split with the most recent prescription in the test set had the worst performance, with AUCs for the best model (GB) of approximately 0.77 and 0.80 for baseline predictors and baseline plus history, respectively. • Various data divisions result in varied model performances (AUC test set: 0.77–0.89). Adding historical data tends to produce a marginal improvement in performance across the board (an increase in absolute AUC of somewhere between 1% and 2%). • The models did not predict index prescriptions or patients whose PDC levels suddenly dropped (AUC test set: 0.56–0.66) (recall: 0.58–0.63).	(−) The patient data used in this study came from a single region in Sweden, which has specific regulations regarding healthcare and pharmacies that might not apply to other countries healthcare systems. (−) The simulation results were not validated by comparing them to the real consumption patterns of the cohort. The study focused on simple patterns because there was a lack of such data; nevertheless, actual consumption patterns may be more complex. (+) Robust model training employs ML algorithms using four unique data splits, including stratified random, patient, and temporal forward prediction with and without index patients, to retain model performance when using new data instead of training data. (+) Robustness of results due to application and comparison of multiple ML algorithms in the prediction and pattern analysis of medication refill adherence.
Author(s): Aziz et al.	Determining hypertensive patients' beliefs about medication and MA	• In total, there were 160 participants in this study. In 2011, patients were collected	RF	• The RMSE values produced by the ML models built with the selected variables were	(−) The authors of the study said that it was not yet possible to claim that the

(*continued on next page*)

Table 4 (*continued*)

Details of Source	Title	Method/Structure	ML Approach	Results	Strengths (+)/Limitations (−)
Source: *PeerJ* **Year:** 2020	associations using ML methods.	from the outpatient clinics at the University Kebangsaan Hospital. All adult patients diagnosed with essential hypertension and those using at least one antihypertensive medication for longer than a year were considered candidates for the study.		1.50 for ANN, 1.53 for RF, and 1.55 for SVR. • The accuracy of the dichotomized scores provided accuracies of 65% (ANN), 78% (RF), and 79% (SVR), respectively. This accuracy was calculated based on a percentage of	findings had a broad generalizability because the study was based on a limited number of clinical data. This was because the study was conducted at a single institution.
		• The Malaysian MA Scale and the validated Beliefs about Medicines Questionnaire were utilised in the construction of the questionnaire. Both of these instruments permitted an evaluation of the general attitudes of MA and were utilised in developing the questionnaire.	ANN	correctly recognised adherence values and was employed as an extra model performance parameter.	(+) The application of SOM demonstrates how clinical data can be seen in a two-dimensional representation by coupling it with dimensionality reduction techniques that map higher-dimensional data onto lower-dimensional space. This makes it possible to simplify complex problems to have a better understanding of them.
		• High- drug adherent patients had an overall score between 6 and 8. To assess how well the ML model performed, the RMSE was utilised.	Support Vector Regression (SVR)		
Author(s): Gao et al. **Source:** *American Journal of Hypertension* **Year:** 2020	A clinical prediction model of MA in hypertensive patients in a Chinese Community Hospital in Beijing	• Data on all patients who were treated at the Fangzhuang Community Health Service Center between June 1, 2014, and December 31, 2018, were retrieved using the Intelligent Chronic Disease Management System.	Decision trees	• The predictive model for antihypertensive MA had a ROC-AUC of 0.81, a sensitivity of 0.78, and a specificity of 0.69.	(−) In order to evaluate the accuracy of the prediction model, the study solely applied basic cross-validation and external validation. Again, the methodology should be evaluated using hypertensive patient databases from different community hospitals.
		• Data regarding patients' adherence to antihypertensive medication was analysed using a retrospective approach beginning one year prior to the initial prescription and continuing for another six months following the patient's final prescription. A total of 7638 people with hypertension participated in the research study.			(−) Since the study only used one ML algorithm, decision trees, it did not use and compare multiple ML algorithms to measure how well hypertensive patients adhered to their medicine.
		• The PDC was determined by multiplying the total amount of antihypertensive medications prescribed in a community hospital over one year. A PDC of less than 0.8 means adherence is low, while a PDC of more than 0.8 means that adherence is strong.			(+) The Chi-squared test for significance was used to screen characteristic variables in hierarchical or binary data to ensure that only statistically significant variables were included in the analysis. The Wilcoxon signed-rank test was used to screen characteristic variables in continuous data to ensure that only statistically significant variables were included.
Author(s): Son et al. **Source:** *Health Informatics Research* **Year:** 2010	Application of SVM for prediction of MA HF patients.	• A self-reported questionnaire was distributed to 76 heart failure patients at a university hospital to see how effectively they took their meds. Running mathematical simulations to determine the variables that best predict how well patients take their medicine resulted in an SVM model.	SVM	• 77.63% was the most accurate detection accuracy that was achieved. According to the research findings, SVM modelling is an effective classification approach that may be used to predict MA in patients with heart failure.	The primary problem with this study is that it used such a small sample size that it is difficult for any results to be statistically significant.
		• The dataset was subjected to LOOCV to see how well the SVM models' estimations held up.			(+) This was one of the first studies to use SVM to find out what made Korean patients with HF adhere to or not adhere to their medications. This made the study useful because it generated new

(*continued on next page*)

Table 4 (*continued*)

Details of Source	Title	Method/Structure	ML Approach	Results	Strengths (+)/Limitations (−)
					information that researchers could use in the same field in the future.
Author(s): Karanasiou et al. **Source:** *Healthcare Technology Letters* **Year:** 2016	Predicting adherence of patients with HF through ML techniques.	• The classifiers were assessed using data from 90 cases that were collected retrospectively. The information was extracted from the Department of Cardiology at the University Hospital of Ioannina.	RF	• When it came to the first classification problem (global adherent), the best detection accuracy was 82%, and when it came to the second classification problem (medication adherent), it was 91%.	(+) Comprehensive analysis of adherence in HF patients using more than ten ML and DL algorithms, allowing comparison of the performance of the various algorithms for the recommendation of the best classification ML or DL algorithm for adherence prediction.
		• Patients with heart failure who were at least 18 years old, receiving optimal care, and who had recently undergone hospitalisation, an emergency admission, or specialised consultation for decompensated HF were included in the study. Additionally, patients needed at least one ECG, typically associated with heart failure within the previous year. Patients were classified as either global or medication adherents, depending on the results of medical assessments.	Random Trees (RT) Logistic Model Trees (LMTs) Rotation Forest, SVM Radial Basis Function Network (RBF Network) Bayesian Network Naive Bayes MLP Simple CART		(+) The filter approach determines the value of the retrieved features and validates that the chosen variables are independent of the learning process. In addition, it eliminates the risk of overfitting the data and ensures that the conclusions apply to a wide range of contexts.
Author(s): Franklin et al. **Source:** *Health Services Research* **Year:** 2016	Observing versus predicting: initial patterns of filling predict long-term adherence more accurately than high-dimensional modelling techniques.	• A survey was conducted involving people receiving Medicare and having CVS Caremark as their pharmacy benefit provider. Respondents had to be at least 65 years old. Patients who started taking statins or statin combinations between January 1, 2006, and December 31, 2008, were considered for inclusion in this study. The index date was determined by the date of the first statin prescription filled after 180 days without medication.	LR	• When only baseline data were used, the prediction was poor, with a maximum cross-validated C-statistic of 0.606 and 0.577 for patients with an index supply of 30 days and >30 days, respectively. This difference was because patients with an index supply of 30 days had more information available to them.	(−) Since the study used a single analysis method (logistic regression), it lacks an evaluation of adherence prediction based on different ML algorithms.
		• In the model, there were 77 703 statin initiators used. Patients must have been enrolled in Medicare and Caremark for 180 days before the index date to be eligible for the study. Patients were considered for inclusion in the study if they had one or more health care claims filed and one or more prescriptions filled six months before the index fill.		• Predictions were substantially more accurate for patients whose initial statin prescriptions were shorter when using only markers of first statin adherence (C = 0.827/0.518). Adding factors chosen by the researchers made predictions much more accurate (C = 0.842/0.596).	(−) Patients who were actively involved in the healthcare system and remained insured by Medicare and their Part D prescription plan for 180 days before and 365 days after the initial statin dispensing were excluded from the study. Because of the requirement that participants in the study had continuous health insurance for at least 18 months, it is possible that this cohort did not accurately represent all statin initiators in Medicare. As a consequence of this, it is possible that this group had a higher proportion of adherence, and it is also possible that the findings cannot be generalised to other Medicare patients.
		• Patients were disqualified from the study if they had lost their eligibility for Medicare or Caremark, moved into an assisted living facility or hospice, or had passed away			(−) Since younger people were not included in the study, the findings may not be applicable to a younger working population taking statins or predicting how well people will take other

(*continued on next page*)

Table 4 (*continued*)

Details of Source	Title	Method/Structure	ML Approach	Results	Strengths (+)/Limitations (−)
		within the first year of the follow-up period.			medications for long-term conditions. Additionally, the results may not be useful for predicting how well people will adhere to other medications.
Author(s): Lucas et al. **Source:** *PLoS One* **Year:** 2017	An EHR-based model predicts statin adherence, LDL cholesterol, and CVD in the United States military health system.	• EHR data came from the MilitaryHealth System, which keeps administrative records for military personnel who have retired from the military and dependents eligible for medical benefits. • Patients who had been prescribed a statin for the first time in 2005 and 2006 were the survey subjects. The data from baseline billing, laboratory, and pharmacy claims were gathered and summarised using non-negative matrix factorization in the two years leading up to the first statin prescription. The follow-up statin prescription refill data were used to determine the adherence outcome, greater than 80% of the days covered. • The dataset contained 138 731 individuals.	RF	• Using the patient statin prescription, the anticipated disease risk, and the EHR features, a cross-validated C-statistic with a value of 0.736 was ascertained to predict statin non-adherence. Following the addition of the initial refill, the C-statistic increased to 0.81.	(−) Since the study only used one ML method, RF, it lacked the application and comparison of various ML algorithms in analysing statin adherence. (−) The study relied on prescription refill data, which are just a proxy for MA because the degree of actual medication use could not be determined. (+) The research employed a dimension reduction technique to reduce the number of variables by combining data from the pharmacy, the laboratory, and the billing into linked characteristics. As missing data is common in EHRs, this technique simplifies using the prediction model. Therefore, when attempting to predict whether a new patient will take a statin, several data points can be substituted for one another without impacting the model's performance. (+) The model lays the groundwork for future individualised studies into the causes of non-adherence by identifying different patient clusters at a higher or lower adherence risk. These patient clusters can either have a higher or lower likelihood of not following the medication regimen.
Author(s): Kumamaru et al. **Source:** Journal of Managed Care & Specialty Pharmacy **Year:** 2018	Using previous MA to predict future adherence.	• The research looked at the prior adherence of 89 490 persons who were starting statins for the first time and evaluated their prior adherence to other chronic preventative drugs over a baseline of 365 days. The study was conducted in the United States. A PDC greater than or equal to 80% was used to determine strong statin adherence. • The study identified 18-year-olds who started a statin between July 2010 and December 2011. Initiation of a statin was defined as a new statin dispensation after 365 days with no statin dispensation. The cohort	LR	• The C-statistic for a prediction model based solely on demographics was 0.578 (95% confidence interval: 0.573–0.584). • After considering the patient's comorbidities, utilisation of medical services, and medication use, the C-statistic came out to be 0.665 with a 95% confidence interval ranging from 0.659	(−) The study used a single analysis approach (logistic regression); thus, the study lacks evaluation of adherence prediction based on multiple ML algorithms. (−) The study measured statin adherence in claims data using PDC to define high adherence. Although this has been shown to correlate well with other adherence measures, including drug presence

(*continued on next page*)

Table 4 (*continued*)

Details of Source	Title	Method/Structure	ML Approach	Results	Strengths (+)/Limitations (−)
		entry date was the first statin dispensation. Only the first patient cohort entry was included.		to 0.670. The C-statistics for models with the individual previous MA measurements as the only explanatory variable ranged from 0.533 for not getting a second fill to 0.666 for the highest PDC.	measured by serum levels, and is a widely used measure of adherence, dispensing patterns may not exactly correspond to patient medication-taking behaviour.
		• The research project developed LR models by employing a range of categorical variables and historical adherence indicators to predict high adherence in a random sample of 50% of the total. C-statistics were used to verify that its discrimination was accurate. By fitting a modified model, the researchers also investigated whether or not there was a correlation between past and subsequent statin adherence.		• These results were found for models with the individual previous MA measurements as the sole explanatory variable. The C-statistic for the combined model also includes a mean PDC range from 0.695 to 0.700. If a patient's prior mean PDC was less than 25%, they were about half as likely to take their prescription statins (risk ratio = 0.49, 95% CI = 0.46–0.50), whereas if it was larger than 80%, they had a relatively higher chance to take their statins.	(−) selected a lookback period of 365 days for assessing baseline covariates and previous adherence. Requiring longer lookback periods may lead to a more complete capture of chronic conditions. Still, it would reduce the number of eligible patients, potentially limiting the generalizability of the prediction models. (+) it is one of the few studies to create models to predict future adherence using prior MA. This generates new information by laying a foundation based on the previously established MA factor.
Author(s): Zullig et al. **Source:** *Health Services Research* **Year:** 2019	Novel application of approaches to predicting MA using medical claims data	• The data came from 11 969 Medicare recipients who submitted claims to Medicare Parts A, B, and D for acute myocardial infarction (MI)-related hospitalizations between the years 2007 and 2012 and filled a statin prescription either at the time of discharge or as soon as they were able to do so after the event.	LR with the backward selection of covariates	• The three analytic approaches had moderate discrimination (C-index ranging from 0.664 to 0.673).	(−) The dataset used was limited to patients who had already filled their prescriptions. As a result, the study could only look at issues during the implementation phase of taking medications, not at the start or primary non-adherence.
		• The C-index was used to evaluate the level of discrimination exhibited by the model, and decile plots were utilised to compare the projected values and the observed event rates.	LASSO	• Although the LASSO regression model selected over 90% of all possible predictors, there was only a slight difference between the three distinct analytical approaches (C-index ranged from 0.664 to 0.673).	(−) The analytic models used administrative claims data that lacked clinical, socioeconomic, and behavioural variables that may influence MA and information about treatment non-adherence reasons, such as provider determination and drug prices. (−) Since characteristics associated with lipid-lowering MA vary across groups, the use of Medicare fee-for-service claims limits generalizability to other patient populations (for example, the younger and uninsured) or payer systems (for example, commercial or Medicaid). (+) Using three different models allows for comparing how well they predict MA in patients.
			RF		
Author(s): Janssoone et al. **Source:**https://doi.org/10.48550/arXiv.1811.12234 **Year:** 2018	ML on EHRs: Models and features usages to predict medication non-adherence	• Consumption and reimbursement data for breast cancer patients were obtained from the French Health Insurance System (SNIIRAM, the French National Health System), which covered 99.8% of the French population (66 million people). The dataset included hospitalizations, medicine purchases, and patient-specific information	LR Decision Tree Gradient boosting MLP	• The AUCs were approximately 0.70, and the best result was achieved with Gradient Boosting.	(+) SNIIRAM is one of the world's largest organised databases of health data, and it was used in this study. This then helps in the process of generalising the findings from the investigation. (−) Due to a constraint related to the labelling of the data, it is necessary to devise a plan that will investigate automatic labelling and anomaly discovery to locate

(*continued on next page*)

Table 4 (*continued*)

Details of Source	Title	Method/Structure	ML Approach	Results	Strengths (+)/Limitations (−)
		like age, government assistance, and geographic data.			information in the data that is more accurate.
Author(s): Meneveau et al. **Source:** *Breast Cancer Research and Treatment* **Year:** 2020	Predicting adjuvant endocrine therapy initiation and adherence among older women with early stage breast cancer.	• Using the Surveillance, Epidemiology, and End Results—Medicare database, a search was conducted to identify women aged 70 and older who were in the early stages of ER and breast cancer. Comorbidities, socioeconomic factors, socio-economic status, prescription medications, and demographic data were collected and analysed as potential predictors.	LR	• Within the cohort, 8703 (78.9%) patients received adjuvant endocrine therapy, and 6685 (60.6%) remained on AET for at least one year.	(−) The Medicare database that was used is not exhaustive, despite the fact that it contains a sizeable proportion of American women over the age of 70 who are having treatment for breast cancer in its early stages.
		• Eleven thousand thirty-seven patients were eligible to participate in the trial after meeting the criteria. Using logistic regression, a stepwise selection of significant factors was utilised to produce classifiers for initiation and adherence. The C-statistic was one of the metrics that was used in the process of evaluating how effective the model was.		• AET initiation and adherence bivariate factors were comparable. The AET initiation and adherence classifiers with the highest C-statistics, 0.65 and 0.60, respectively, were found to be the most accurate.	(−) Medicare Part D claims data were used in the calculation of the MPR. Because this information depends on whether a prescription was filled, it may not be a reliable indicator of whether or not AET was used. (−) Since this study only used information from SEER, the projected models may not have been as accurate as they could have been. This is because the models didn't consider patient-specific characteristics like attitudes toward healthcare and personality traits. (+) It is one of the few recent studies that add to the corpus of knowledge about adherence evaluation in breast cancer patients using ML.
Author(s): Yerrapragada et al. **Source:** *JCO Clinical Cancer Informatics* **Year:** 2021	ML to predict Tamoxifen Non-adherence among US commercially insured patients with metastatic breast cancer	• The IBM MarketScan Commercial Claims and Encounters and Medicare Claims datasets contain information on 3022 women who had breast cancer from 2012 to 2017 that had progressed to other body areas. This dataset was used in the study. Non-adherent patients had less than 80% PDC the year after starting treatment.	LR	• Logistic regression and feedforward neural networks performed similarly (AUROC 0.64) and outperformed both enhanced LR (AUROC 0.61) and RF (AUROC 0.61). (AUROC 0.62).	(+) The research was methodologically sound since it underwent extensive training, testing, and comparison of four ML approaches to predict tamoxifen non-adherence.
		• Before and after the index date, patients who participated in the trial should have had a continuous period of at least 12 months of medical and pharmaceutical coverage to meet the requirements for 24 months of monitoring.	RF	• After using the synthetic minority oversampling technique to balance the data, RF (AUROC = 0.93) and feedforward neural networks (AUROC = 0.79) produced more reliable predictions. On the other hand, the LR algorithm (with an AUROC of 0.57) and the boosted LR algorithm (with an AUROC of 0.56) performed somewhat worse than when the data was not balanced.	(+) This study utilised a large and varied sample of patients that was indicative of care settings that are found in the real world.
		• The databases contained medical and pharmacy claims for employees, their spouses, and dependents who were covered by employer-sponsored private health insurance. Additionally, the databases contained medical and pharmacy claims for retirees eligible for Medicare and had Medicare supplemental plans funded by large employers'	Boosted LR	• When given more balanced data, the RF model was also better at classifying patients as either adhesive (97% negative predictive value) or non-adherent (65% positive predictive value), with age 65 or older continues to be the most important variable.	(−) The study was also restricted to patients who regularly enrolled, representing healthy users. Given this, the results are likely to be a solid representation of individuals who received insurance via their employment or who had ongoing benefits with fairly steady coverage.

(*continued on next page*)

Table 4 (*continued*)

Details of Source	Title	Method/Structure	ML Approach	Results	Strengths (+)/Limitations (−)
		health benefit programmes. They also had records about the treatments and admissions of inpatients, outpatient services, and purchases of prescription drugs.			
Author(s): Fan et al. **Source:** *Frontiers in Pharmacology* **Year:** 2019	ML approaches to predict risks of diabetic complications and poor glycaemic control in non-adherent T2D.	• PDC was used to assess non-adherence. The AUROC and F-score were used to test four models. • Face-to-face interviews and the Sichuan Provincial People's Hospital's Electronic Health Medical Record System (EHRS) were used to collect data for this study. Every subject was a patient with T2D. The final batch included 165 patients. (+) The raw data earmarked for the study were randomly grouped ten times by modifying the seed value of the "partition." This made it possible to do the same experiments more than once without introducing bias, which can happen when datasets are randomly put into categories.	Feedforward Neural Networks ANN	• Ensemble diabetes models outperformed all other predictive models for nephropathy and diabetic angiopathy complications, with AUCs of 0.902 ± 0.040 and 0.889 ± 0.059, respectively.	
Discriminate (D) and Ensemble models		The results of the ensemble models were based on a summary of the best three models (as measured by AUC) and the voting principle. • Every single participant had T2D. The final group consisted of 165 different patients. The results of the ensemble models were based on the voting method and a summary of the results from the top three models (as shown by AUC). • One example of a complication of diabetes is diabetic peripheral neuropathy (DPN). Other complications include diabetic angiopathy (DA), diabetic nephropathy (DN), and diabetic eye disease (DED).	Bayesian network (BN) CHAID CRT QUEST	• Discriminate (D) performed the best of all diabetic peripheral neuropathy and diabetic eye disease models. Its AUCs were 0.859 ± 0.050 for diabetic peripheral neuropathy and 0.832 ± 0.086 for diabetic eye disease. BN was the most accurate glycosylated haemoglobin A1c (HbA1c) model, with a maximum AUC of 0.825 ± 0.092.	(+) The length of uncorrected hypoglycemia treatment was a clinical variable that had not been thoroughly explored in the past but was present in the dataset that was used for prediction in the study. This was because previous research had focused mostly on other aspects of the condition. (−) Due to the fact that this was a study conducted at a single centre using a very small sample size, the performance of the final models was not compared to that of the known clinical reference tools. This is because doing so would limit the validity of the verification findings.
Author(s): Pettas et al. **Source:** *IEEE* **Year:** 2019	Recognition of breathing activity and MA using LSTM neural networks.	• At the University of Patras, three healthy people were given an inhaler for 12 s, the surrounding area was kept quiet, and the room's acoustics were carefully controlled. • The inhalation device was equipped with a microphone that collected audio signals during breathing and drug activation. Additionally, the inhalation device communicated with a mobile smartphone through Bluetooth.	LSTM RNN LSTM Neural Networks	• The LSTM-based approach, which achieved a prediction accuracy ranging from 92% to 94% while considering samples containing a mixture of patterns corresponding to more than one of the specified classes, gave more accurate findings than the conventional ML methods.	The research only involved three participants who made recordings inside a building with its acoustics carefully controlled. It would have been better to validate the results with a larger sample of subjects. (+) Even though there were only a few participants in the study, it utilised empirically collected data, which is more credible than other types of data because it represents real-life experience instead of just theories.

(*continued on next page*)

Table 4 (*continued*)

Details of Source	Title	Method/Structure	ML Approach	Results	Strengths (+)/Limitations (−)
		Each individual audio recording was sampled at a frequency of 8 kilohertz, and the depth was also 8 bits.			
Author(s): Ntalianis et al. **Source:** *IEEE* **Year:** 2019	Assessment of MA in respiratory diseases through deep sparse convolutional coding.	• The audio samples that were obtained were then represented in the frequency domain so that the monitor audio samples could be differentiated from one another more easily. The study focused on each sound source's spectrogram, which shows the signal power at different frequencies over time.	RF • Audio files were recorded from twelve healthy people in indoor and outdoor settings. All of the subjects used the same inhaler device, but each of the subjects used a different canister. The recordings were divided into four categories: inhaler activation, inhalation, exhalation, and environmental noise or other sounds. Each of these categories was then annotated with relevant information.	Deep sparse CNN	• The classification accuracy of these models ranged from 94 to 95%, and their cross-entropy loss was between 0.20 and 0.25, which suggests that this method could also be used in an embedded system whose only task would be to monitor medication levels.
		(+) In contrast to the research carried out by Ref. [101]; which only managed to record audio files from three subjects in a regulated indoor environment using an inhaler device, the research carried out by Ref. [102] involved twelve subjects and was successful in recording audio files in both indoor and outdoor environments, providing greater generalizability of results in terms of the environment.			
• A total of 1980 recordings were gathered, with each class having 495 recordings. The used windows of the n = 4000 samples were resized to 250 x 16 to serve as input for the classifier. Experimenting with the matrix in several situations led to determining its dimensions; 125 by 32 produced comparable results.	(−) This study did not include evaluating the performance of other algorithms using the same dataset because it only used one algorithm to evaluate MA.				
Author(s): Koesmahargyo et al. **Source:** *Psychiatry Research* **Year:** 2020	Accuracy of ML-based prediction of MA in clinical research.	• A smartphone app video recorded 4182 clinical trial participants taking their prescribed medication.	XGBoost	• The classification models made accurate predictions of adherence for the entirety of the trial (AUC = 0.83), the following week (AUC = 0.87), and the following day (AUC = 0.87).	(+) The study used primary data obtained through remote, real-time medication dose assessment. This provided the benefits of authentic, accurate, and up-to-date data,

(*continued on next page*)

Table 4 (*continued*)

Details of Source	Title	Method/Structure	ML Approach	Results	Strengths (+)/Limitations (−)
					which were all advantages of the study. (−) Patients with mental health disorders were overrepresented in the data collected, and the ratio of adherent patients to non-adherent patients was skewed, even though non-adherent patients are usually a smaller part of the patient population as a whole.
		• The people who took part in the study suffered from various illnesses, including congestive HF and chronic obstructive pulmonary disease. The study considered the patients' primary diagnosis, demographic information, and past signs of adherence or non-adherence to make the following predictions: (1) 80% adherence rates across the clinical trial; (2) 80% adherence for the next week; and (3) 80% adherence for the next day. • The AUC was used to assess the classifier's performance. The average AUC was used as the performance measure for classification in the 5-fold cross-validation.			
Author(s): Lee et al. **Source:** Health Informatics Research **Year:** 2013	Predictors of MA in elderly patients with chronic diseases using support vector machine models.	• A cross-sectional design descriptive survey was carried out in the outpatient clinics of a teaching hospital in the city of Cheonan in the Republic of Korea. The data from 293 people over 65 who suffered from chronic diseases were analysed between January and May of 2011.	SVM LR	• The classification accuracy was 71.1% when LR was employed and 97.3% when SVM was used.	(−) The study only did a single assessment of self-reported adherence, not a longitudinal one. When compared to other techniques of assessment, self-reports tend to overstate adherence behaviour and, in general, have high specificity but low sensitivity. Additionally, self-reports are often employed.
		• Morisky's self-report was used to evaluate drug adherence in older adults with normal cognitive function who had been taking medication for more than six months for asthma, hypertension, diabetes, chronic obstructive pulmonary disease, liver cirrhosis, stroke, or heart disease. Respondents who had consented but could not complete the survey on their own were invited to participate in the face-to-face interviews.		• The accuracy was 72.4% when using self-efficacy as the only variable in the model. The findings of both the LR and SVM analyses indicate that self-efficacy is a significant factor in determining the degree to which older people in Korea adhere to their medication regimens.	(−) The findings of the study cannot be generalised to patients in younger groups because the data were collected from patients with chronic diseases who were older than 65 years old. (+) The application of LR and SVM for modelling the same dataset enabled comparisons to be made between the two methods.
Author(s): Wang et al. **Source:** *Patient Prefer Adherence* **Year:** 2020	Applying ML models to predict medication non-adherence in Crohn's disease maintenance therapy.	• These people had Crohn's disease and were treated at the GI Department of Shanghai Ruijin Hospital, a facility connected with Shanghai Jiaotong University School of Medicine. They were either admitted to the hospital or travelled there themselves.	LR	• The three models used in this research provided reliable predictions, each achieving an accuracy of at least 81.6% and an AUC of 0.896. • The SVM had an accuracy of 87.7%, a recall of 86.2%, a precision of 85.6%, an F1 score of 0.855, and an AUC of 0.930.	(+) The model's good performance could be because of the feature selection method, which used RF and univariate analysis to build an eight-dimensional vector feature set with low dimensionality. This approach would shorten the model computation time and help avoid overfitting, improving model generalisation and classification.
		• Patients were given questionnaires regarding drug adherence, anxiety and depression, medication necessity and concerns, and pharmacological knowledge, while other information was collected from their EMRs.	Back-Propagation Neural Network		(−) Since this was a single-centre study in Shanghai, the patient profile may have been skewed and not reflective of the country as a whole. Data interpretation requires caution when extrapolating to Crohn's Disease patients in general.
		• The process of manually collecting patient health data from hospital computerised medical records resulted in creating a single-centre	SVM		

(*continued on next page*)

Table 4 (*continued*)

Details of Source	Title	Method/Structure	ML Approach	Results	Strengths (+)/Limitations (−)
		database for treating Crohn's disease that contained 128 elements. EMR system that kept track of the patient's demographic information, socioeconomic information (such as education level, employment status, and income), clinical symptoms, laboratory test results and diagnoses, treatment plans, and follow-up data was used.			
		• The models were evaluated using the AUC, the F1 score, recall, accuracy, and precision.	RF		

5.2. Recent work on ML analytics application to CVD medication and statin adherence

From the existing literature, five research papers on the application of ML analytics to CVD medicine and statin adherence emerged [20,30,83,105–107]. SVM was used by Ref. [30] to predict MA in patients suffering from heart failure. To assess the dependability of the SVM model estimations, leave-one-out cross-validation (LOOCV) was used on the MA data obtained from 76 patients diagnosed with heart failure (HF) who were treated at a university hospital and had completed a self-reported questionnaire. One model had seven predictors—age, education, monthly income, ejection fraction, Mini-Mental Status Examination—Korean, medication knowledge, and functional class—while the other had only five. The models with the best classification of MA in HF patients were the ones with seven predictors and five predictors, respectively. It was determined that a detection rate of 77.63% was the most accurate. According to the study's findings, SVM modelling is a useful classification approach that may accurately predict MA in patients suffering from heart failure.

[106] used ML techniques to build a model for predicting adherence in HF patients. Two questions regarding classification were discussed, the first of which asked whether or not the patient was globally adherent and the second of which asked whether or not the patient was medication adherent. RF, random trees (RT), logistic model trees (LMTs), rotation forest, SVM, radial basis function network (RBF network), BN, naive Bayes, MLP, and a simple classification regression tree were among the eleven classification techniques that were utilised. The best detection accuracy for the first classification problem was 82% and 91% for the second classification problem. The suggested methods' ability to predict how well HF patients will adhere to their medication regimens with satisfactory model prediction accuracy suggests that it can improve how HF patients are managed.

[107] used RF to construct an EHR-based model for statin adherence. This model was associated with clinical outcomes in patients who were taking statin medicine. Statins are a class of drugs regularly recommended to those with a high risk of developing CVDs, as they can lower blood levels of low-density lipoprotein cholesterol [108]. The required EHR data was collected with the help of the Armed Health System, responsible for maintaining administrative data on active duty personnel, retirees, and dependents of United States military service members who received health benefits. The adherence outcome was figured out by using the data gathered from multiple refills of the statin medicine. The detection accuracy was 82% for the first classification problem, while it was 91% for the second classification problem. The classification of statin non-adherence using an RF predictive model based on patient statin medication, predicted disease risk, and EHR parameters as potential inputs gave a cross-validated C-statistic of 0.736. When the initial prescription refill was added to the model, the C-statistic went up to 0.81. Statins were linked to major improvements in lowering cholesterol and fewer hospitalizations for heart attacks, strokes, and coronary artery disease.

Better adherence prediction algorithms were proposed and put to the test by Ref. [105]. When LR was used in patients beginning treatment with statins, the prediction made using baseline data alone was unsatisfactory. The maximum cross-validated C-statistics for patients with an index supply of at least 30 days were 0.606 and 0.577, respectively. Using only markers of initial statin adherence significantly improved prediction accuracy (C = 0.827/0.518) among patients whose initial dispensing was shorter. Prediction accuracy was further improved when paired with investigator-specified variables (C = 0.842/0.596) [20]. used LR to analyse several metrics of past MA to predict future statin adherence. Their database was substantial and contained administrative claims from the United States. The confidence interval for the 95% credible interval for the C-statistic of a model that incorporated information on patient comorbidities, utilisation of medical services, and medication use was 0.665. In models that included only one of the prior drug adherence characteristics as an explanatory variable, the absence of a second fill resulted in a C-statistic of 0.533 (95% confidence interval = 0.529–0.537), whereas the highest PDC resulted in a C-statistic of 0.666 (95% confidence interval = 0.661–0.671) When the mean PDC from the combined model was taken into account, the c-statistic came out to be 0.695 (95% CI: 0.690–0.700).

[83] assessed and compared three predictive analytic approaches for evaluating medication non-adherence and determining under what conditions each method would be most effective based on a prescription for a statin that was given at the time of discharge or shortly after. Standard LR with backward covariate selection, LASSO, and RF were the three analytics methodologies used. The C-index measure (range 0.5, non-informative, to 1.0, perfect prediction) was used to test models for discrimination. The models were calibrated using decile plots, which compare the predicted event rates of the model to the actual event rates. In every model, previous statin use was the most important factor in determining future adherence. C-index values ranging from 0.664 to 0.673 indicate minimal variation among the three analytical approaches, even though the LASSO regression model chose approximately 90% of all available variables.

5.3. Recent work on ML analytics application to cancer treatment MA

Three studies that evaluated ML analytics' application in measuring patient adhesion to cancer medication were also gathered and reviewed [109–111]. [109] used LR, decision tree, gradient boosting, and multilayer perceptron to predict pharmaceutical non-adherence while providing doctors with insights into the underlying causes of the medication drop-outs. Consumption data were collected from breast cancer patients, while data on payments were obtained from SNIIRAM,

the French National Health System. The collection consisted of data on hospital stays, purchases of pharmaceuticals, and contextual patient data such as age, access to public services, and geographical particulars. The models' AUCs were around 0.70, with gradient boosting having the greatest predictive performance of patient non-adherence.

[110] developed an LR model for the beginning and maintenance of adjuvant endocrine therapy (AET) to assist in the decision-making process regarding the omission of radiation therapy. Using the SEER-Medicare database, the researchers identified women over 70 in the early stages of oestrogen receptor and breast cancer. They gathered information on comorbidities, socioeconomic factors, socioeconomic status, prescription medicines, and demographic data as potential predictors. To generate LR classifiers for initiation and adherence, an iterative procedure that involved choosing significant variables was utilised. The AET initiation and adherence classifiers had C-statistics of 0.65, whereas the adherence classifier's C-statistic was 0.60. The strongest models in their analysis were only moderately accurate at predicting adherence. This shows that predicting adherence is challenging because the factors that affect starting and consistency of taking AET are complex and vary from patient to patient.

[111] created an ML model to screen women with metastatic breast cancer for tamoxifen non-adherence in the first year of treatment. They used freely available baseline real-world data as their data source. Non-adherence was measured and assessed by PDC using LR, boosted LR, RF, and feedforward neural networks. The models were created and internally validated using the area under the receiver operating characteristic curve (AUROC). According to their findings, using baseline administrative data and leveraging ML effectively predicted tamoxifen nonadherence, baseline claims were insufficient to distinguish between levels of adherence. Moreover, further validation with extended longitudinal data enhanced model performance, particularly with the RF model.

5.4. Recent work on ML analytics application to MA in respiratory diseases

Research has shown that ML can also be used to analyse MA in individuals with respiratory diseases [101,102]. In order to provide a data-driven solution for tracking pressurised metered dose inhaler MA [101], used RNN-equipped long short-term memory (LSTM) units and spectrogram characteristics. This enabled the researchers to track MA in the simulation. Three healthy individuals were observed and recorded in an indoor environment that was both acoustically controlled and devoid of extraneous noise. The audio signals that were used to capture the inhalers lasted 12 s each. The audio samples obtained were then represented in the frequency domain so that the monitor audio samples could be differentiated from one another more easily. During that inquiry, the spectrogram of each audio sample was obtained. The power of the time-localized signal was shown at several different frequencies on the reconstructed spectrogram.

A deep sparse convolutional neural network (CNN) was employed as a classifier to track how well people comply with their medications in real-time, according to Ref. [102]. The identical inhaler device with various canisters collected audio data from twelve healthy subjects in indoor and outdoor settings. Inhaler activation, inhalation, exhalation, and background noise or other sounds were the four categories into which the recordings were split. Using the recordings, the suggested model was able to categorise with a level of accuracy of 95%. This shows that this method could be used on an embedded device to monitor MA. When samples contain patterns from more than one specified class, the LSTM-based approach was more accurate than standard ML algorithms, with prediction accuracy ranging from 92% to 94%.

5.5. Recent work on ML analytics application to MA in diabetes

Previous research has revealed that ML is used to analyse MA in diabetic patients. Eight studies on this issue were identified [7,22,69,90,112–115]. To determine the best levels of adherence for hospitalisation risk discrimination [112], employed ML to study the relationship between oral hypoglycemic MA and hospitalisation avoidance. Using administrative claims data from Pennsylvania Medicaid, the researchers conducted a retrospective longitudinal cohort analysis on non-dual eligible medical aid participants between 18 and 64 diagnosed with T2D. They identified hospitalisation predictors using RSFs and fitted survival trees (ST) to experimentally determine the adherence levels that best distinguish hospitalisation risk when combined with the PDC. The strictest adherence requirements for risk of all-cause hospitalisation ranged from 46% to 94%, depending on the patient's health and the pharmaceutical regimen's complexity. However, it was noted that ML techniques show promise as a simple and effective approach for optimising healthcare delivery and generating personalised approaches to MA.

A DL approach for identifying adherence in T2D was created by Ref. [69]. This method was constructed using simulated CGM. To simulate a wide range of CGM signals for patients with T2D, an adapted version of the Medtronic virtual patient (MVP) model originally developed for type 1 diabetes was utilised. In order to evaluate and compare LR, MLPs, and CNN, these signals were put through an extensive grid search. According to the study's findings, DL proved beneficial for tracking the adherence of T2D patients [113]. predicted patient risk of non-adherence in T2D care using data from U.S. claims by employing LR, RF, XGBoosting, BART, and super learners. In order to train the ML models and provide accurate predictions on the degree to which patients would adhere to metformin monotherapy, the Truven database was searched for type 2 diabetes (T2D) patients who had initially started taking metformin by itself (the index date). The PDC was utilised to ascertain the degree of metformin adherence. They compared the LR model with other non-linear ML models such as XGboosting, BART, and super learner to optimise the accuracy and sensitivity. With an 80% random split training set and a 20% random split test set, RF classifiers showed accuracy and sensitivity of 0.73 in early analysis.

[114] analysed various ML algorithms and established a model that could be used to predict the likelihood that T2D patients would not adhere to their medications as prescribed. The ML algorithms that were utilised were BN, neural net, SVM, LR, K-NN, LSVM, RF, C5.0 model, Tree-AS, CHAID, Quest, C&R tree, and the ensemble model. The medication possession ratio was utilised, and ML modelling methodologies were used to perform the MA evaluation on the patients. The AUC values for the built-in modelling approaches ranged from a minimum of 0.557 (SD 0.051) to a maximum of 0.866 (SD 0.082), with 0.866 being the most accurate. The best ML approach was an ensemble of five models that used oversampling to ensure that the data were balanced following the data imputation.

[7] used a smart sharps bin and ML to predict injectable drug adherence using random trees, RF, XGBoost, gradient boosting, and MLP via ensemble learning. This study monitored and kept track of patients' injection disposal practises in their homes by using real-time data that was generated by a connected (IoT) system known as a "Smart Sharps Bin (SSB) ". Both random search and grid search, along with 10-fold cross-validation, were utilised to optimise the model's hyperparameters. With a ROC-AUC score of 0.86, the developed ML method did quite well in terms of prediction. According to the study's findings, data from HealthBeacon SSB can be utilised to determine with an accuracy of 81.3% whether or not a patient is most likely to skip a prescription dose. A strategy that accurately identifies patients at risk of future non-adherence can be obtained by using data from the HealthBeacon SSB in conjunction with an ML model, as the results of this study have demonstrated. An early warning system for adherence detection was constructed by Ref. [22] using CNN. The system was based on massive in-silico CGM and injection data. CGM data from individuals with T2D was simulated using labelled adherence or nonadherence to the once-daily insulin injection that was recommended to them. The

well-established and T2D-modified MVP model served as the foundation for the simulations of simulated plasma glucose concentration excursions. When there was more CGM data available on the day when the classification was performed, there was an increase in the accuracy of detecting adherence.

[90] used RL to enhance diabetes MA by optimising response and personalising engagement. Patients participating in the trial who had poor control of their diabetes and who were using oral diabetic medications were randomly allocated to either the RL intervention or the control condition. Electronic pill bottles were employed in the study's control and intervention groups; however, the intervention group also received daily text messages. The instructions were personalised for each patient using an RL prediction algorithm that was derived from daily pill bottle adherence measurements. The REINFORCE experiment aimed to investigate whether or not it was feasible to boost MA in T2D patients by utilising ML techniques to personalise the content of text messages. Since the intervention was discovered to be successful, this approach will be replicated and implemented in other clinical settings, as well as for a wider range of health behaviours. The research demonstrated the potential of improving other self-management systems by increasing the scope of RL.

[115] assessed the predictive power of ANN, BN, CHAID, CRT, QUEST, discriminate (D), and ensemble models in non-adherent T2D. Diabetic peripheral neuropathy (DPN), diabetic angiopathy (DA), diabetic nephropathy (DN), and diabetic eye disease (DED) were among the consequences studied. The majority of the assessed models performed satisfactorily. Ensemble diabetes models outperformed all other predictive models of nephropathy and diabetic angiopathy problems, with AUCs of 0.889 ± 0.059 and 0.902 ± 0.040, respectively. Discriminate (D) performed the best out of the diabetic peripheral neuropathy and diabetic eye disease models, with AUCs of 0.859 ± 0.050 and 0.832 ± 0.086, respectively. The model BN, with an AUC of 0.825 ± 0.092, predicted glycosylated haemoglobin A1c (HbA1c) the best. After the developed prediction models in this work have been tested and screened, the final models could be useful for T2D patients' general practitioners, endocrinologists, and various other specialists in the medical field.

5.6. Recent work on ML analytics application to MA in other NCD situations

There is also published research on the use of ML MA in patients who suffer from a variety of clinical ailments such as asthma, stroke, hypertension, diabetes, liver cirrhosis, HF, diabetes, CVDs, and Crohn's disease, among others [21,26,71]. [21] examined the precision of dynamic non-adherence prediction by utilising data from remote, real-time medicine dosing measurements by employing XGBoost. This was done to determine whether or not XGBoost could predict non-adherence accurately. Patients suffering from congestive HF, chronic obstructive pulmonary disease, and various other health conditions participated in the trial. Empirically reported adherence was the best indicator of future adherence or non-adherence, which was a significant finding. The ability to appropriately predict adherence was demonstrated in the trial (AUC = 0.83), the subsequent week (AUC = 0.87), and the following day (AUC = 0.87). According to the study's results, accurate prediction of dynamic drug adherence can be achieved by utilising real-time dosage measurement.

SVM models and LR were utilised by Ref. [32] to determine the features that influence the degree to which older patients with chronic illnesses adhere to taking their prescribed medication. A cross-sectional design descriptive survey was carried out in the outpatient clinics of a teaching hospital in the city of Cheonan in the Republic of Korea. The data for the study were collected from 293 individuals over the age of 65 who were diagnosed with chronic conditions (such as asthma, hypertension, diabetes, chronic obstructive pulmonary disease, liver cirrhosis, stroke, and CVDs) but had normal cognitive function. This data collection took place between January and May of 2011. The classification accuracy was 71.1% when LR was used, but SVM improved it to 97.3%. The accuracy was 72.4% when self-efficacy was the only variable that was used in the analysis. The results of the LR and SVM show that self-efficacy is an important factor in determining how well Korean senior citizens adhere to the medication regimens that they have been prescribed.

To expedite the intervention process [26], developed ML models to predict non-adherence to azathioprine among patients with Chron's disease. They created the models using SVM, LR, and backpropagation neural networks. An AUC of 0.896 and a minimum accuracy of 81.6% were shown to be present in each of the three models utilised in the analysis. According to the study, the SVM is significantly more effective than the linear regression and the backpropagation neural network. It had a higher F1 score of 0.855 and a higher AUC of 0.930. Its accuracy was greater, coming in at 87.7%, its recall was higher, showing a value of 86.2%, and its precision was higher, reaching 85.6%.

6. Results

Considering the period spanning from January 1, 2010, to August 31, 2022, a total of 25 studies on the application of ML to the analytics of NCD MA met the eligibility selection criteria depicted in Table 1. Seven (7) of the 25 full-text articles used ML to analyse MA in diabetes patients, six used ML to evaluate MA in CVD medication and statin adherence, four used ML to analyse MA in cancer patients, three used ML to evaluate MA in hypertension patients, two used ML-based analytics to measure adherence in respiratory disease patients, and three used ML analytics application of MA in other NCD situations. The results of this review are expressed and discussed in four sub-sections. These include MA measurement thresholds in NCD patients, techniques and data sources that enable the application of ML to measure MA, ML algorithms used for MA analytics, and evaluation of the ML models for NCD MA analytics. Fig. 5 illustrates these findings.

6.1. Thresholds of measuring MA in NCD patients

The proportion of days covered (PDC) was a typical way to measure drug adherence or non-adherence in the research under consideration. It is generally accepted that the PDC refers to the portion (or percentage) of days following the beginning of treatment during which a patient can obtain their prescribed medication, with the exact number of days obtainable being contingent on the fill dates and the total number of days' supply for each dispensing method (Malo et al., 2017). The PDC is widely used to evaluate MA using administrative data throughout the implementation phase, from the beginning to the end of medication use (De Geest et al., 2012). The PDC was dichotomized at a threshold of more than 80% (>80%), which demonstrates MA, and a threshold of less

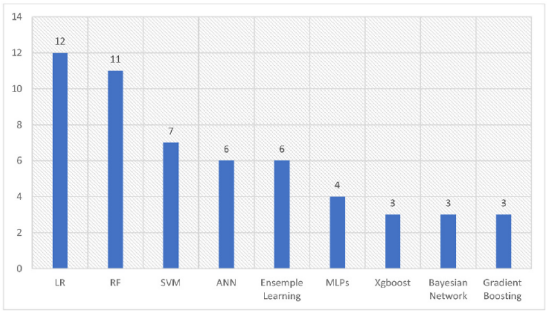


Fig. 5. ML/DL approaches used in at least three past studies in the analytics of NCD MA.

than 80% (<80%), which represents non-adherence to MA [107,111]. Other examined studies characterised high adherence as PDC ≥80% or poor adherence as PDC ≤80% [20,21,104,113], which differs slightly from the outright threshold above or below the 80% threshold. The 80% threshold was used when measuring MA based on prescription refill data and when measuring adherence to statins after treatment initiation [20] and examining non-adherence using data from remote real-time medication dosing measurements [21]. A standardised MA assessment scale, such as the Malaysian MA scale (MALMAS), was utilised, with total values ranging from 0 to 8 [103]. This MA measure allows for an analysis of overall MA perceptions, with a score of 6–8 (75%) indicating high adherence. Strong adherence is generally defined as at least 75% of the PDC, with PDC ≥ 80% being the most common high adherence threshold used. In other cases, NCD patients were classified as adherent to prescribed medication based on clinician assessment [106].

6.2. Data sources and data generation techniques enabling the application of ML to analyse MA

The primary goal of ML modelling in medicine or public health is to incorporate data from a wide variety of sources. These sources may include clinical measures and observations, biological data, experimental findings, environmental data, data generated through wearable technology, and data generated by other electronic systems. MA can be measured via electronic mechanisms [90,101,113], self-reported surveys [21,30,32,102,103], electronic medical records (EMRs) and prescription refill or dispensation data [105,112]; Karanasiou et al., 2016; [20,83,86,104,107,109–111], therapeutic drug monitoring [22,69], and using both EMR and a survey [26,114,115].

Most of the studies (11 research articles) on the application of ML in analysing MA included in this systematic review used EMRs, prescription refills, or insurance claim data. These could be administrative claims data extracted from claims databases; detailed patient records that were kept at public facilities that provided primary, secondary, and speciality care; historic injection disposal records collected from a cloud-connected MA technology database that supports self-administering of injectable medication in the home environment; or data extracted from an eHealth database. The following variables were present in the EMRs and prescription refill datasets: patient visit frequency, prior prescription information and prior PDC, medication possession ratio, type of disease, treatment adherence, information on long-term care, insurance claims and encounter data for outpatients and inpatients, therapeutic regimen, prescription drug claims, date of prescription filling, the amount is given, and days of supply. Pharmacy claims data and EMR records with information about drug prescriptions, drug consumption rates, medical care, and patient information allow ML and DL algorithms to be used to analyse how well NCD patients adhere to their medications.

Other studies used real-time monitoring and tracking of medicine using IoT data to measure MA. This finding was substantiated by extensive data extraction that was carried out using a connected IoT system known as “Smart Sharps to monitor and track the injection disposal of patients in their home environment” [7]. The data collected by the IoT device would then be utilised to create an ensemble learning model for predicting MA with very high predictive performance. The other electronic method was using electronic pill bottles in the daily adherence measurement for patients with a computing device and an internet connection [90]. Daily text messages were sent to patients who received this intervention, reminding them to take their medications. To monitor patients’ adherence, bidirectional text messaging or electronic pill bottles were utilised (Mheta, 2019). To support MA, the framing of SMS messages sent to patients would then be uniquely personalised using RL prediction algorithms based on daily adherence measures from the pill bottles. In another study, the capacity to recognise respiratory activity and MA was evaluated by listening to audio signals of breathing and drug actuation obtained from a microphone linked to a patient’s inhalation equipment and connecting with a mobile phone over Bluetooth [101]. The gathered audio samples were then represented in the frequency domain so that the ability of the audio samples to vary from one another could be optimised and monitored. The spectrogram of each audio sample was extracted explicitly for the study. A spectrogram displays the time-localized signal power at various frequencies. RNNs employed LSTM units and spectrogram features to track MA in patients using pressurised metered-dose inhalers.

A study similar to the one carried out by Ref. [101] was conducted by Ref. [102]. This study collected audio data from twelve healthy individuals who used an identical inhaler device with different canisters indoors and outdoors. The activation of the inhaler, inhalation, exhalation and any other sounds that may have been present in the environment were analysed and divided into four categories. The sparse CNN algorithm was used as a classifier based on the audio recordings to provide a real-time evaluation of MA in patients suffering from respiratory illnesses who were utilising the inhaler device. Patients with a wide range of diseases, like congestive heart failure (HF) and chronic obstructive pulmonary disease (COPD), were observed using a smartphone app that recorded videos of patients taking their medicine [21]. Furthermore, while the use of CGM incorporates an electronic method for evaluating MA, the role of therapeutic drug monitoring (TDM) in such a system cannot be overlooked. Based on simulated CGM signals, the results showed that T2D patients could also use DL-based adherence detection [22,69]. A continuous glucose monitoring (CGM) monitor is a medical device that gives real-time glucose data, allowing continuous tracking and monitoring of diabetes patients’ blood glucose levels throughout the day [116]. The acquisition of real-time glucose data has made it possible for medical professionals to make clinical decisions for the management of diabetes that are both more rapid and more proactive. The authors found that it is possible to replicate in-silico CGM data from T2D patients who have labelled their adherence or non-adherence to the required once-daily insulin injection. The well-known and T2D-modified MVP model obtained the in silico CGM data. This model can simulate changes in the amount of plasma glucose (PG). As more CGM data became available on the day of classification, it became simpler to ascertain whether or not an individual was following the prescribed medication schedule provided to them.

MA was also measured using self-reported surveys. For example, in a study by Ref. [103]; self-reported MA was based on the beliefs about medicines questionnaire (BMQ) and a standard MA scale with total values ranging from 0 to 8. This MA measure assessed overall perceptions of MA, with a score of 6–8 (≥75%) considered high adherence. In addition, a cross-sectional descriptive survey of older adults with chronic conditions has been conducted to ascertain determinants of drug adherence. Morisky’s self-report was employed in this MA assessment technique. Morisky’s questionnaire was administered verbally to consenting responders who could not complete it themselves. According to Ref. [117]; the Morisky MA scale is a regularly used adherence screening instrument that consists of dichotomized questions on historical medication usage patterns that need yes or no replies. The questionnaire method is often used during medication history interviews because it is quick and easy to use.

6.3. Machine learning and DL algorithms used for the analytics of NCD MA

In the analysis of NCD MA, various ML and DL approaches have been applied: LR, RF, SVM, ANN, MLP, ensemble learning, XGBoost, CNN, BN, gradient boosting, and decision trees such as the C5.0 model, CHAID, CART, and QUEST. Additional algorithms and approaches used in no more than two identified and included studies include Bayesian additive regression, K-NN, reinforcement learning, K-means, SVR, LMTs, ST, radial basis function networks (RBFN), naive Bayes, simple CART, LASSO, boosted LR, and LSTM RNN. Fig. 5 summarises and depicts the ML/DL approaches used in at least three research articles in MA analytics.

SVM (including linear support vector machine [LSVM]), ANN is interchangeably expressed as neural net (including feedforward neural networks, back-propagation neural networks, & LSTM neural networks), ensemble learning (including super learner, extremely randomized trees, random trees).

Fig. 5 shows that LR (n = 12), RF (n = 11), SVM (n = 7), ANN (n = 6), ensemble learning (n = 6), MLPs (n = 4), XGBoost (n = 3), BN (n = 3), and gradient boosting (n = 3) were the most common ML and DL methods used to identify, predict, classify, or cluster MA in NCD patients. The study shows that ML, DL, and ensemble learning can predict MA or non-adherence in patients. Using EMRs, real-time medication dosing data, or direct pharmacy transactional data, ML, DL algorithms, reinforcement learning, and ensemble learning can be used to inform caregivers about how well a patient is adhering to or not adhering to their medications, as well as the risk of dropping a drug at certain points in the treatment. Fusing various heterogeneous classifiers such that they work as a set to complement one another is the basis for the observed high performance of ensemble learning. In ensemble learning, model outputs are generated through a voting mechanism or architecture that combines the findings from a number of different models in order to make more accurate predictions [7,115]. The use of ensemble learning was a reliable and practical technique in the study that examined the use of ML in predicting risks of complications and inadequate glycemic control in non-adherent T2D patients with AUCs over 0.88. This was discovered in the research that looked at the application of ML in this area [115]. In other cases, DL in the form of long short-term memory RNNs performed exceptionally well, achieving at least 92% prediction accuracy [101]. The deep sparse CNN demonstrated that DL algorithms could be implemented on an embedded device built to monitor MA [102].

Random forest (RF) models also performed well. Random forests are considered ensemble classifiers because they are built out of several decision trees, each of which votes for one of the classes [106]. The final classification of a sample is then decided using a majority-vote method. The class that receives the majority of votes from the trees is the one to which the classified sample is most likely to belong. This is because votes come from the trees. It is important to highlight that the ML algorithms were used to build adherence prediction models and feature engineering as part of the overall data analytics pipeline. For example [26], built credible LR, neural net, and SVM models with a minimum accuracy of 81.6% and an AUC of 0.89 to predict non-adherence to Crohn's disease maintenance medication. These models were able to predict non-adherence accurately. This level of model performance was credited to the feature selection strategy, which combined RF and univariate analysis to generate an eight-dimensional vector feature set of low dimensionality. This approach hastened the development process, making it easier to avoid overfitting, resulting in improved model generalisation and classification. Similarly [113], compared the LR model. They optimised accuracy and sensitivity through feature engineering and non-linear ML methods such as XGboosting, BART, and super learner. In the next section (6.4), model evaluation and performance measures are discussed in more depth.

6.4. Evaluation of the developed ML/DL models

This study shows how different metrics may be used to assess how well ML or DL models are performing. These include the AUC or ROC-AUC and the confusion matrix - linked to the evaluation metrics listed below: accuracy, precision, positive predictive value (PPV), negative predictive value (NPV), F1 score, recall/sensitivity, and specificity. RMSE, cumulative accuracy profit curve (CAP) or Lorenz curve, decile plots, LOOCV, mean ensemble test accuracy (META), and the k-fold cross-validation procedure were also used to evaluate the models. The AUC or ROC-AUC were used widely in the process of evaluating the predictive performance of the ML, DL, and ensemble learning model solutions [21,26,86,104,109,111,114,115].

For instance [104], used decision trees to build a clinical prediction model of drug adherence in hypertension patients. With a ROC-AUC of 0.810, the prediction model exhibited 0.78 sensitivity and 0.69 specificity for predicting antihypertensive MA. This aligns with the findings of [7]; who used metrics like accuracy, specificity, sensitivity, precision, the F1 score, and AUC to evaluate the performance of ensemble learning and DL models. The recall and/or sensitivity of the confusion matrix was 91%, which indicates that 91% of the prediction was made for those individuals who would take their medication at the appropriate time. Their innovative ML approach exhibited very good predictive performance, with an AUROC of 0.86 and an accuracy of 81.3%, respectively. In other cases, five measures were used in a single study to evaluate and compare model performance [26]. These measures included accuracy, recall, precision, F1-score, and AUC. According to previous research, the SVM outperformed neural networks and LR in nearly every category, with an accuracy of 87.7%, recall of 86.2%, precision of 85.6%, F1 score of 0.855, and an AUC as high as 0.930. This research also revealed very accurate predictions with a minimum accuracy of 81.6% and an AUC as high as 0.896 [26].

Another study reveals the combined use of AUCs or the CAP curve to evaluate the capability of LR, decision trees, gradient boosting, and MLPs to predict medication discontinuations [109]. The CAP provides second and third measurements, demonstrating a model's ability to reliably identify a patient at risk. The CAP was employed as a tool in ML to visualise the discriminative capability of classification models, with CAP values ranging from 0.47 to 0.7. This highlights the importance of model evaluation criteria in comparing different ML or DL models built on a shared dataset. A model evaluation metric known as the "C-statistic, " also referred to as the "C-index" or the "concordance statistic," was utilised in some of the reviewed studies [20,105,107,110]. The C-statistic was utilised to evaluate the appropriateness of the developed models for the binary outcomes. The C-statistic is one example of a statistic that can be used in clinical research. In one clinical research example, the C-statistic indicates the frequency with which the model accurately differentiates between patients who take their medication as directed and those who do not. The C-statistic can take on values anywhere from 0.5 to 1, with values close to 1 signifying a robust model.

It is important to note, however, that the models used by the various ML algorithms vary depending on the state and type of predictors used, the number of refills used in a year for a dataset, feature selection, and the length of consumption history, as well as the specific medication consumption behaviours in question. This is consistent with the findings of [86]; who demonstrated very strong predictive RF model performance (AUC 0.90) when patients with a single prescription from the test set were included and used baseline predictors and history. The RF model first achieved a cross-validated C-statistic of 0.736 for identifying statin non-adherence based on EHR variables in the study by Ref. [107]. When the initial refill was included in the model, the C-statistic rose to 0.81. This means that the number of refills in the dataset for predicting MA can affect the model's performance. Prediction models evaluated using C-statistic initially yielded 0.578 when only demographic models were included [20]; the C-statistic value improved to 0.665 after including patient comorbidities, health care services utilisation and medication use as additional input variables, and it improved, even more, when previous MA and mean PDC were included as additional input variables.

In other cases, balancing data with a dataset balancing technique known as synthetic minority oversampling technique (SMOTE) improved predictive abilities for RF (AUC 0.93) and neural net (AUC 0.79); otherwise, RF had similar predictive abilities (AUC ranging from 0.61 to 0.64) when compared to LR, boosted LR, and neural net (AUC ranging from 0.61 to 0.64) [111]. In the research carried out by Ref. [7]; the balanced sampling strategy was utilised in conjunction with data imputation, which is a method that estimates and replaces missing data with some replacement value to keep the majority of the data in the dataset, and binning, which is a process that converts numerical

variables into their categorical counterparts. This was done to improve the prediction models' accuracy by reducing the amount of noise or non-linearity in the dataset. In that study, feature selection and model evaluation were made using the AUROC metric. A slight variation involved using the C-index to test model discrimination and decile plots to evaluate the performance of an MA model by comparing projected values to observed event rates [83]. Both of these methods are described in more detail below. The C-index ranged from 0.664 to 0.673, indicating that all three analytic techniques (LR with backward covariate selection, LASSO, and RF) showed moderate discrimination abilities. Decile charts are a data visualisation tool that separates a data series into ten equal segments. A series, for example, can be divided into 10-decile groups regardless of size. Using data from medical claims, these groups are used to test the outputs of the predictive model to see how well it predicts how well people adhere to their medications.

The accuracy metric was one of the model evaluation metrics mentioned in the research studies we reviewed [101–103,113]. Accuracy is a metric used to measure an ML model's performance. This metric indicates the proportion of test data predictions or the number of times the ML model is accurate. [103]; for example, reported the performance accuracy of their models based on a percentage of correctly classified adherence values; the ANN model had 65% accuracy, the RF model had 78% accuracy, and the SVR model had 79% accuracy in assessing the beliefs of hypertension patients and their relationships regarding MA. According to the Wilcoxon signed ranked test, there was no significant difference between the actual scores and the predictions the ML models generated. In the same study by Ref. [103]; the ANN had an RSME value of 1.42, the RF had an RSME value of 1.53, and the SVR had an RSME value of 1.55. These values also show how well each model performed. The RSME assists in measuring how well a model makes quality predictions or quantifies a model's error in producing predictions. Thus, accuracy and RSME measurements agree in terms of evaluating model performance in terms of making correct or incorrect predictions. The LSTM RNN produced higher accuracy and superior model performance using DL techniques than standard ML methods, such as the RF, with prediction accuracy ranging from 92% to 94% [101]. DL-based models again had the highest accuracy, achieving 77.5 ± 1.4% accuracy with CNN and 72.5 ± 3.5% accuracy with MLP. This is in comparison to the accuracy of the traditional LR model, which was 65.2 ± 0.8% [102]. successfully achieved good model performance by utilising another DL (deep sparse CNN) technique, achieving a classification accuracy of 95%. At one point, the accuracy of ensemble learning was calculated by adding the standard deviation to the mean ensemble test accuracy, also known as META. This resulted in a precision of 0.776. Sensitivity (0.776), specificity (0.776), PPV (0.778), and NPV were the other reported measures (0.816).

Another cross-entropy loss metric was used in research that evaluated MA among patients with respiratory-related diseases. This metric measures how effectively a classification ML mode performs [101,102]. The test cross-entropy loss in the study by Ref. [102] was between 0.20 and 0.25, indicating that the created CNN model performed well in identifying MA. This holds because the loss (or error) is expressed as a number ranging from 0 to 1, with a test loss relatively close to 0 indicating a very close to being a perfect model.

According to the findings, LOOCV and K-fold cross-validation were also used to measure how well ML algorithms performed while making predictions using data that had not been used in the model's training [21,26,30]. LOOCV was used on the dataset to evaluate the accuracy of the estimations provided by the SVM models for the prediction of MA. The models achieved a maximum accuracy of 77.63% in their predictions. During ML classifier training, the LOOCV plays a vital role in avoiding the overfitting of a classifier on the training set. AUC estimates for each subsequent iteration were used with 5-fold cross-validation [21]. The cross-validation technique was useful in determining how well the ML models performed after training on the unknown data. Finally, utilising data from remote real-time measurements of drug administration and the 5-fold cross-validation technique, researchers acquired an AUC of at least 0.83 and higher generalisation ability on the testing set and unforeseen data. In related research, two-way cross--validation improved model evaluation accuracy in studies that created prediction models of medication non-adherence risks in T2D patients using various ML algorithms [114]. [26] created an accurate prediction in their modelling of medication non-adherence by utilising stratified 10-fold cross-validation. This allowed them to achieve a minimum accuracy of 81.6% and an AUC of 0.896. This procedure assists in the detection of selection bias or overfitting. It provides insight into how the model would generalise to a dataset not part of the training set or original study. Fig. 6 summarises the various ML/DL approaches and model evaluation metrics.

7. Discussion

Though clinician estimates could be used to characterise MA or non-adherence among NCD patients, increasing research has employed the PDC as the basis for reporting MA or non-adherence. MA is generally considered high when it exceeds 75% of the PDC, with PDC $\geq$ 80% being the most common high adherence criterion. EMRs are the most commonly used data source in developing ML models for the prediction or analytics of MA. Among other variables, the historical ERMs used medication refill data, health insurance claims, dispensing data, patient visit frequency, days of medication supply, past PDC, disease types, clinical presentations, medication possession ratio, drug usage rates, therapeutic regimen, and demographic information of the patients. Incorporating patients' MA histories, such as past prescriptions and initial and historical patterns of medicine refills, enhances the ML models' performance and greatly improves prediction accuracy. By leveraging ML approaches, self-reported adherence through self-completion surveys about previous medication patterns has been used to generate the data required for measuring MA. Measuring drug adherence in real time via electronic mechanisms is also common. In that vein, some of the electronic mechanisms used include real-time monitoring and tracking of medication taking, taking advantage of data generated by IoT devices such as electronic pill bottles, bidirectional text messaging in the daily measurement of adherence for patients, TDM such as CGM to generate real-time glucose data for continuous monitoring of diabetes patients' blood glucose. In other cases, audio files were recorded, and DL algorithms such as CNN and RNN were used as classifiers on the data to give a real-time assessment of MA in respiratory disease patients utilising an inhaler device.

In general, the accuracy of adherence detection increases when additional real-time electronic medication data, such as CGM data, become available on the day of categorization. This means that the availability of big data on MA among NCD patients can increase ML models' categorization and predictive capabilities. As a result, empirical measurement of medication dose, real-time surveillance of medication use, and self-reported MA surveys can give enormous amounts of data for modelling MA prediction. Techniques for monitoring drug adherence might be either direct or indirect. Direct measurement is described by Ref. [118] as the direct observation of drug administration or the detection of a drug or its metabolite in a biological fluid. Continuous real-time tracking and monitoring of diabetes patients' blood glucose levels throughout the day using the CGM is a direct measurement approach in the context of our current investigation. However, while this method is accurate, it also has the disadvantages of being expensive and labour-consuming for healthcare practitioners. It comes with challenging logistics for completing these assessments when working with large patient groups and underprivileged communities. This resonates with [118]; who notes that despite these drawbacks, many authorities consider that electronic medication monitoring gives the most accurate and pertinent data on MA, especially in complex clinical circumstances [9]. In our study, however, the indirect approach to measuring MA comprised self-reported MA and examining previous EMRs based on the

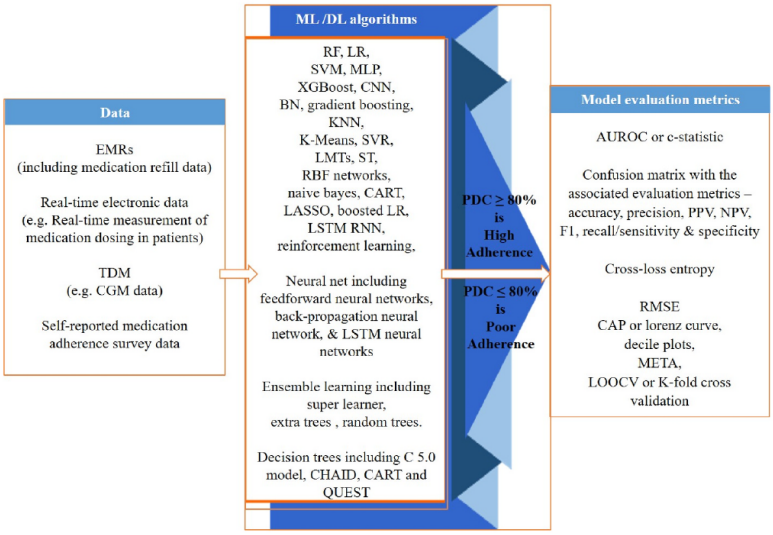


Fig. 6. ML/DL approaches and model evaluation metrics.

PDC. Indirect evaluation methods are becoming more popular because they are easy to use and cost less to set up [68,118]. Thus, with costly and labour-consuming electronic medication monitoring, leveraging historical EMRs for assessing MA is a good fit for underdeveloped communities.

This paper shows how to apply multiple supervised and unsupervised ML, DL, and ensemble learning models to predict, classify and analyse MA in NCD patients. Popular ML techniques include LR, RF, SVM, ANN, and ensemble, learning models. In some cases, ML algorithms were used to develop models for predicting or classifying MA and to select variables to employ predictors with significant statistical significance when building the models. In addition to feature engineering, the predictive capability of the ML models was improved by using techniques such as balanced sampling, data imputation, binning, and cross-validation. With an AUC of 0.88, ensemble learning, which integrates the performances of different ML algorithms based on voting principles or design, is demonstrated to be robust in creating superior predictions. Ensemble learning models, including RF, are among the ML models that performed well because they are classified as ensemble classifiers, which are naturally composed of several decision trees, each of which votes for one of the classes.

RF is effective with both categorical and continuous data due to its foundation in the bagging method and its application of the ensemble learning technique. It constructs as many trees as it can on the subset of data and then integrates the results of several decision trees to reach a single output. As a result, both the overfitting problem in decision trees and the variance can be decreased, leading to an increase in the accuracy of regression and classification tasks. Because RF can execute both regression and classification tasks accurately, it is a technique commonly used in data science. This is demonstrated by using RF in at least eleven studies on developing classification or prediction models to analyse MA in NCD patients. The application of RF makes it easier to evaluate the contribution or relevance of a variable to the model. In most cases, the Gini importance and mean drop in impurity (MDI) metrics are used to measure the extent to which the model's accuracy worsens due to removing a specific variable. In one of the studies that used RF [106], used the Gini index throughout their research to determine which characteristic of each tree node indicated the optimal split. This was done to discover which tree node attribute best represented the optimum split. The validity of each RF sub-branches was confirmed using data acquired from samples that were not chosen for development. Before creating test models [103], assessed the relevance of the features using the RF feature importance approach to determine which ones were most relevant. [119]; who also investigated the predictor importance (PI) by mixing the OOB predictor observations within the RF classifier model, found this consistent with their findings. The RF classifier is a useful method for estimating missing values due to feature bagging since it retains its accuracy even when portions of the data are missing. This is due, once again, to feature bagging. RF methods can handle massive data sets, making more accurate predictions. Still, because they compute data for each decision tree, they can process data more slowly than other algorithms.

Adopting DL models such as ANN, CNN, RNN, and LSTM neural networks enhanced MA analytics performance when applied to NCD patients. DL approaches used to recognise MA and respiratory activity, such as LSTM neural networks, performed remarkably well in identifying MA, with prediction accuracy between 92% and 94%. These techniques were utilised to detect both MA and respiratory activity. A study conducted by Ref. [26] revealed the high-value proposition of DL in the form of algorithms. In this study, the researchers used LSTM and recurrent neural networks to predict the progression of Alzheimer's disease and reported an accuracy of 99% $\pm$ 0.0043. This indicates that the algorithms are very accurate. Even though DL models require significant computational resources, such as powerful Graphical Processing Units (GPUs) and large amounts of memory, which can be expensive and time-consuming, their performance has dramatically increased in a wide range of applications, including the medical sciences and they are particularly effective at revealing complex architecture in high-dimensional data. Furthermore, their performance in a wide range of applications has significantly improved. Many firms, including Google and Microsoft, use DL approaches to achieve significant gains in various classification and regression challenges and datasets [120]. According to the findings of this study, one of the benefits of DL algorithms is their ability to analyse both structured and unstructured data, such as images, text, audio, and any other types of historical data. DL approaches, which are renowned for their exceptional ability to learn from past data, have the potential to play a critical role in the development of intelligent data-driven systems that meet today's standards. Over and above analysing MA based on historical textual data, it emerged that a DL algorithm was useful in recognising breathing activity

and MA using LSTM neural networks. Again in another application of DL, the assessment of MA in respiratory diseases through deep sparse convolutional coding was based on audio files that were recorded from people in indoor and outdoor settings using an inhalation device.

Interestingly, DL techniques such as RNNs and LSTM neural networks are most suited for processing sequential data types such as time series, speech or audio, and text. These algorithms may be able to preserve context and memory across time, allowing them to make predictions based on previous inputs. Even when given unstructured data, DL models may be trained to optimise nearly any function in any domain. This is a huge benefit. According to Ref. [121]; DL outperforms other standard ML methods and shallow networks in fields involving unstructured data analysis and potentially larger datasets. When analysing unstructured data, it is common to practise for traditional ML techniques to have limits. This is problematic because unstructured data is a substantial data source today. From this vantage point, it would appear that DL wields the most influence. The drawback of DL-based classification algorithms is that they typically need very large data sets. Since there are difficulties in training models with many features on small data sets and finding solutions that generalise effectively to the population, researchers continue to rely on traditional ML techniques despite their limited sample sizes. As noted by Ref. [122]; while DL takes a long time to train a model with many dataset features, it runs quickly during testing compared to classical ML techniques.

It is critical to note that the performance of models across the various Ml algorithms can be observed to be dependent on the state and type of predictors used, the quantity of data used, the number of refills used in a year for a dataset, feature selection, and length of consumption history considered, as well as the specific medication consumption behaviours in question. This can be seen in four ways: data adequacy, predictor number, the statistical significance of predictors, and length of consumption history. As a result, if a model is properly fed, it will only produce sound findings. The most frequent model, performance evaluation metrics, used in creating MA analytics models include the AUC or ROC-AUC and the confusion matrix–linked with accuracy, precision, recall/sensitivity, the F1 score, and specificity. When procedures such as the LOOCV or K-fold cross-validation are employed during the training of ML classifiers to prevent overfitting the classifiers on the training set, the validation result is presented as an AUC. These K-fold cross-validations, which can also be 5-fold and 10-fold cross-validations, enable trustworthy model prediction with high accuracy. Overall, the study shows that ML, DL, and ensemble learning can predict patients' MA or non-adherence.

8. Strengths and limitations

Given the paucity of literature on studies that have used ML approaches to analyse MA among NCD patients, this review provides well-collated literature on MA literature with integrated ML-based analytics. This article provides a cutting-edge, comprehensive, systematic review of how ML algorithms have been integrated into MA analytics. The expansive nature of the systematic literature review is evident in the extensive inclusion of studies on ML application in the analytics of MA in patients with various NCDs, such as diabetes, hypertension, cancer, CVDs, and respiratory diseases, among others. Studies that applied the various standard ML algorithms, DL approaches, ensemble learning models, boasting ML algorithms, and tree-based ML algorithms were included. This current review study informs NCD patients, caregivers, and practitioners about the accuracy of ML algorithms in measuring NCD MA for informed decision-making. The research had certain limitations. The evaluation was limited to research that examined MA using ML methods; thus, the number of studies analysed was relatively small.

9. Conclusion

It should be highlighted that using standard ML, DL, and ensemble learning in various analytics, such as prediction, classification, and clustering of adherence or non-adherence, offers huge potential for analysing MA among NCD patients. However, the results show that the models perform differently based on the dataset size, the state, type, and significance of variables employed, and the length of past consumption behaviours. As a result, data scientists or analysts must exercise caution when selecting the dataset and variables to utilise in developing ML models to predict MA or non-adherence. Similarly, the accuracy of each ML or DL model must be understood in light of the research parameters. In general, ML approaches in medicine or public health enable the incorporation of data from multiple sources, such as direct and indirect clinical measurements and observations, biological data, experimental results, environmental information, wearable devices, and other electronic systems, for modelling MA or non-adherence among NCD patients. Future researchers should concentrate on an empirical examination of the application of ML to evaluating MA among NCD patients based on real-world datasets in Africa, considering environmental factors. Another systematic review can focus on how ML is used to analyse MA and other long-term and contagious diseases, such as tuberculosis.

Ethical approval

N/A.

Authors' contributions

All authors contributed equally in the writing of this manuscript.

Data availability

All data generated/analysed and used to support the findings of this study are included in the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The manuscript was not funded by any organisation nor was any support revied from anyone.

References

[1] World Health Organization. Adherence to long-term therapies: evidence for action. World Health Organization; 2003. https://apps.who.int/iris/bitstream/handle/10665/42682/9241545992.pdf;jsessionid=13A1E77459F57E9933B22387CDA75439?sequence=1.
[2] Cutler RL, Fernandez-Llimos F, Frommer M, Benrimoj C, Garcia-Cardenas V. Economic impact of medication non-adherence by disease groups: a systematic review. BMJ Open 2018;8(1). https://doi.org/10.1136/bmjopen-2017-016982.
[3] Omotosho A, Ayegba P. Medication adherence: a review and lessons for developing countries. International Association of Online Engineering; 2019. https://www.learntechlib.org/p/218048/.
[4] Baveja L, Jain D. You can't manage what you can't measure: medication adherence. https://us.milliman.com/en/insight/you-cant-manage-what-you-cant-measure-medication-adherence-in-chronic-disease-management in chronic disease management; 2021.
[5] Cutler DM, Everett W. Thinking outside the pillbox? Medication adherence as a priority for health care reform. N Engl J Med 2010;362:1553–5. https://doi.org/10.1056/NEJMp1002305.
[6] Mongkhon P, Ashcroft DM, Scholfield CN, Kongkaew C. Hospital admissions associated with medication non-adherence: a systematic review of prospective observational studies. BMJ Qual Saf 2018;27:902–14.
[7] Gu Y, Zalkikar A, Liu M, Kelly L, Hall A, Daly K, Ward T. Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data. Sci Rep 2021;11. https://doi.org/10.1038/s41598-021-98387-w.

[8] Babel A, Taneja R, Mondello MF, Monaco A, Donde S. Artificial intelligence solutions to increase medication adherence in patients with non-communicable diseases. Frontiers in Digital Health 2021;3:1–9.

[9] Lam WY, Fresco P. Medication adherence measures: an overview. BioMed Res Int 2015. https://doi.org/10.1155/2015/217047.

[10] Singla R, Singla A, Gupta Y, Kalra S. Artificial intelligence/machine learning in diabetes care. Indian Journal of Endocrinology and Metabolism 2019;23:495–7.

[11] Ho PM, Rumsfeld JS, Masoudi FA, McClure DL, Plomondon ME, Steiner JF, et al. Effect of medication nonadherence on hospitalization and mortality among patients with diabetes mellitus. Arch Intern Med 2006;166:1836–41. https://doi.org/10.1001/archinte.166.17.1836.

[12] Cadarette SM, Wong L. An introduction to health care administrative data. Can J Hosp Pharm 2015;68:232–7.

[13] Abegaz TM, Tefera YG. Target organ damage and the long term effect of nonadherence to clinical practice guidelines in patients with hypertension: a retrospective cohort study. Int J Hypertens 2017;749. https://doi.org/10.1155/2017/2637051.

[14] Lehmann AP, Ahmed R, Celio J, Gauchet A, Bedouch P, et al. Assessing medication adherence: options to consider. Int J Clin Pharm 2014;36:55–69.

[15] Sackett DL, Haynes RB, Gibson ES, et al. Randomised clinical trial of strategies for improving medication compliance in primary hypertension. Lancet 1975;1:1205–7.

[16] Andrade SE, Kahler KH, Frech F, Chan KA. Methods for evaluation of medication adherence and persistence using automated databases. Pharmacoepidemiol Drug Saf 2006;15:565–74. https://doi.org/10.1002/pds.1230.

[17] Baumgartner PC, Haynes RB, Hersberger KE, Arnet I. A systematic review of medication adherence thresholds dependent of clinical outcomes. Front Pharmacol 2018;9. https://doi.org/10.3389/fphar.2018.01540.

[18] Franklin JM, Krumme AA, Shrank WH, Matlin OS, Brennan TA, Choudhry NK. Predicting adherence trajectory using initial patterns of medication filling. Am J Manag Care 2015;21:537–44.

[19] Lauffenburger JC, Franklin JM, Krumme AA, Shrank WH, Matlin OS, Spettell CM, Brill G, Choudhry NK. Predicting adherence to chronic disease medications in patients with long-term initial medication fills using indicators of clinical events and health behaviors. Journal of Managed Care & Specialty Pharmacy 2018;24:469–77.

[20] Kumamaru H, Lee MP, Choudhry NK, Dong YH, Krumme AA, Khan N, Brill G, Kohsaka S, Miyata H, Schneeweiss S, Gagne JJ. Using previous medication adherence to predict future adherence. Journal of Managed Care & Specialty Pharmacy 2018;24(11):1146–55.

[21] Koesmahargyo V, Abbas A, Zhang L, Guan L, Feng S, Yadav V, Galatzer-Levy IR. Accuracy of machine learning-based prediction of medication adherence in clinical research. Psychiatr Res 2020;294:1–7. https://doi.org/10.1016/j.psychres.2020.113558.

[22] Thyde DN, Mohebbi A, Bengtsson H, Jensen ML, Mørup M. Machine learning-based adherence detection of type 2 diabetes patients on once-daily basal insulin injections. J Diabetes Sci Technol 2021;15(1):98–108.

[23] Ellahham S. Artificial intelligence: the future for diabetes care. Am J Med 2020;133:895–900.

[24] Venkatachalam J, Abrahm SB, Singh Z, Stalin P, Sathya GR. Determinants of patient's adherence to hypertension medications in a rural population of Kancheepuram District in Tamil Nadu, South India. Indian J Community Med: official publication of Indian Association of Preventive & Social Medicine 2015;40(1):33.

[25] Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN. Computer Science 2021;2(3):160.

[26] Wang L, Fan R, Zhang C, Hong L, Zhang T, Chen Y, Liu K, Wang Z, Zhong J. Applying machine learning models to predict medication nonadherence in crohn's disease maintenance therapy. Patient Prefer Adherence 2020;14:917–26.

[27] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019:1347–58.

[28] Tsoi K, Yiu K, Lee H, et al. The HOPE Asia Network. Applications of artificial intelligence for hypertension management. J Clin Hypertens 2021;23:568–74.

[29] Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inf Decis Making 2010;10(1):16.

[30] Son Y-J, Kim Y-G, Kim EH, Choi S, Lee S-K. Application of support vector machine for prediction of medication adherence in heart failure patients. Healthcare Informatics Research 2010;16(4):253–9.

[31] Almansour NA, Syed HF, Khayat NR, Altheeb RK, Juri RE, Alhiyafi J, Alrashed S, Olatunji SO. Neural network and support vector machine for the prediction of chronic kidney disease: a comparative study. Comput Biol Med 2019;109:101–11.

[32] Lee SK, Kang B-Y, Kim H-G, Son Y-J. Predictors of medication adherence in elderly patients with chronic diseases using support vector machine models. Health Informatics Research 2013;19(1):33–41.

[33] Farran B, Channanath AM, Behbehani K, Thanaraj TA. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait a cohort study. BMJ Open 2013;3(5). https://doi.org/10.1136/bmjopen-2012-002457.

[34] Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inf Decis Making 2011;11(1):51.

[35] Golino HF, Amaral L, Duarte SFP, et al. Predicting increased blood pressure using machine learning. Journal of Obesity 2014. https://doi.org/10.1155/2014/637635.

[36] Lopez-Martinez F, Nunez-Valdez ER, Crespo RG, Garcia-Diaz V. An artificial neural network approach for predicting hypertension using NHANES data. Sci Rep 2020;10(1). https://doi.org/10.1038/s41598-020-67640-z.

[37] Soh DCK, Jahmunah V, San TR, Acharya UR. A computational intelligence tool for the detection of hypertension using empirical mode decomposition. Comput Biol Med 2020;118.

[38] Ye C, Fu T, Hao S, et al. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. J Med Internet Res 2018;20(1). https://doi.org/10.2196/jmir.9268.

[39] Kanegae H, Suzuki K, Fukatani K, Ito T, Harada N, Kario K. Highly precise risk prediction model for new-onset hypertension using artificial intelligence techniques. J Clin Hypertens 2019;22(3):445–50.

[40] Lacson RC, Baker B, Suresh H, Andriole K, Szolovits P, Lacson E. Use of machine-learning algorithms to determine features of systolic blood pressure variability that predict poor outcomes in hypertensive patients. Clinical Kidney Journal 2018;12(2):206–12.

[41] Bohlmann A, Mostafa J, Kumar M. Machine learning and medication adherence: scoping review. J Med Internet Res 2021;2(4):e26993. https://doi.org/10.2196/26993.

[42] Zakeri M, Sansgiry SS, Abughosh SM. Application of machine learning in predicting medication adherence of patients with cardiovascular diseases: a systematic review of the literature. Journal of Medical Artificial Intelligence 2022;5(5):1–16.

[43] Robinson L, Arden MA, Dawson S, Walters SJ, Wildman MJ, Stevenson M. A machine-learning assisted review of the use of habit formation in medication adherence interventions for long-term conditions. Health Psychol Rev 2022:1–23.

[44] Stafford IS, Gosink MM, Mossotto E, Ennis S, Hauben M. A systematic review of artificial intelligence and machine learning applications to inflammatory bowel disease, with practical guidelines for interpretation. Inflamm Bowel Dis 2022;20:1–11.

[45] Cramer JA. A systematic review of adherence with medications for diabetes. Diabetes Care 2004;27:1218–24.

[46] Demonceau J, Ruppar T, Kristanto P, Hughes DA, Fargher E, Kardas P, De Geest S, Dobbels F, Lewek P, Urquhart J, Vrijens B. Identification and assessment of adherence-enhancing interventions in studies assessing medication adherence through electronically compiled drug dosing histories: a systematic literature review and meta-analysis. Drugs 2013;73(6):545–62.

[47] McGovern DP, Kugathasan S, Cho JH. Genetics of inflammatory bowel diseases. Gastroenterology 2015;149(5):1163–76. https://doi.org/10.1053/j.gastro.2015.08.001.

[48] Capoccia K, Odegard PS, Letassy N. Medication adherence with diabetes medication: a systematic review of the literature. Diabetes Educat 2016;42:34–71.

[49] McGovern A, Tippu Z, Hinton W, et al. Comparison of medication adherence and persistence in type 2 diabetes: a systematic review and meta-analysis. Diabetes Obes Metabol 2018;20:1040–3.

[50] Walsh CA, Cahir C, Tecklenborg S, Byrne C, Culbertson MA, Bennett KE. The association between medication non-adherence and adverse health outcomes in ageing populations: a systematic review and meta-analysis. Br J Clin Pharmacol 2019;85:2464–78. https://doi.org/10.1111/bcp.14075.

[51] Tola GA, Regassa LD, Weldesenbet AB, Merga BT, Legesse N, Tusa BS. Adherence to antihypertensive medications and associated factors among hypertensive patients in Ethiopia: systematic review and meta-analysis. SAGE Open Medicine 2020;8. https://doi.org/10.1177/2050312120982459.

[52] Evans M, Engberg S, Faurby M, Fernandes JDDR, Hudson P, Polonsky W. Adherence to and persistence with antidiabetic medications and associations with clinical and economic outcomes in people with type 2 diabetes mellitus: a systematic literature review. Diabetes Obes Metabol 2021;24(3):377–90.

[53] Paneerselvam GS, Aftab RA, Baig MAI, Hariadha E. The pharmacist role in improving medication adherence in dialysis patients: a systematic review. Biblio 2021;12(3):761–8.

[54] Weidt F, Silva R. Systematic literature review in computer science-a practical guide. Relatórios Técnicos Do DCC/UFJF; 2016. p. 1.

[55] Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA. Cochrane handbook for systematic reviews of interventions. second ed. Chichester (UK): John Wiley & Sons; 2019.

[56] Polit DF, Beck CT. Essentials of nursing research: appraising evidence for nursing practice. ninth ed. Philadelphia: Wolters Kluwer; 2018.

[57] Bettany-Saltikov J, McSherry R. How to do a systematic literature review in nursing: a step-by-step guide. https://research.tees.ac.uk/en/publications/how-to-do-a-systematic-literature-review-in-nursing-a-step-by-ste-3; 2016. second ed.

[58] Subirana M, Sola I, Garcia J, Gich I, Urrutia G. A nursing qualitative systematic review required MEDLINE and CINAHL for study identification. J Clin Epidemiol 2005;58(1):20–5.

[59] Pae CU. Why systematic review rather than narrative review? Psychiatry Investigation 2015;12(3):417–9.

[60] Jiao Y, Du P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quantitative Biology 2016;4(4):320–33.

[61] Vujović ZD. Classification model evaluation metrics. Int J Adv Comput Sci Appl 2021;12(6):599–606.

[62] Orozco-Arias S, Piña JS, Tabares-Soto R, Castillo-Ossa LF, Guyot R, Isaza G. Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements. Processes 2020;8:2–19.

[63] Steurer M, Hill RJ, Pfeifer N. Metrics for evaluating the performance of machine learning based automated valuation models. In 36th International Association for Research in Income and Wealth Virtual General Conference; 2021. p. 1–37.

[64] Zheng A, Casari A. Feature engineering for machine learning principles and techniques for data scientists. https://www.repath.in/gallery/feature_engineering_for_machine_learning.pdf; 2018.

[65] Haas K, Ben Miled Z, Mahoui M. Medication adherence prediction through online social forums: a case study of fibromyalgia. J Med Internet Res 2019;21(4). https://doi.org/10.2196/12561.

[66] Hess LM, Raebel MA, Conner DA, Malone DC. Measurement of adherence in pharmacy administrative databases: a proposal for standard definitions and preferred measures. Ann Pharmacother 2006;40(7–8). 1280–88.

[67] Dixon BE, Jabour AM, Phillips EO, Marrero DG. An informatics approach to medication adherence assessment and improvement using clinical, billing, and patient-entered data. J Am Med Inf Assoc 2014;21(3):517–21.

[68] Kreys E. Measurements of medication adherence: in search of a gold standard. Journal of Clinical Pathways 2016;2(8):43–7. https://www.hmpgloballearningnetwork.com/site/jcp/article/measurements-medication-adherence-search-gold-standard.

[69] Mohebbi A, Aradottir TB, Johansen AR, Bengtsson H, Fraccaro M, Morup M. A deep learning approach to adherence detection for type 2 diabetics. Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2017:2896–9. https://doi.org/10.1109/EMBC.2017.8037462.

[70] Kotsiantis SB, Kanellopoulos D, Pintelasata PE. Pre-processing for supervised leaning. Int J Comput Sci 2006;1(2):111–7.

[71] Kang H. The prevention and handling of the missing data. Korean Journal of Anesthesiology 2013;64(5):402–6.

[72] Shumeiko D, Rozora I. Handling missing values in machine learning regression problems in II international scientific symposium «intelligent solutions» IntSol-2021. Kyiv-Uzhhorod, Ukraine, http://ceur-ws.org/Vol-3106/Short_5.pdf; 2021.

[73] Seliem M. Loading handling outlier data as missing values by imputation methods: application of machine learning algorithms. Turkish Journal of Computer and Mathematics Education (TURCOMAT) 2022;13(1):273–86.

[74] Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013. https://link.springer.com/book/10.1007/978-1-4614-6849-3.

[75] Nargesian F, Samulowitz H, Khurana U, Khalil EB, Turaga D. *Learning feature engineering for classification* in proceedings of the twenty-sixth international joint conference on artificial intelligence main track. 2529-2535, https://doi.org/10.24963/ijcai.2017/352; 2017.

[76] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Bioinformatics Advances 2015. https://doi.org/10.1155/2015/198363.

[77] Potdar K, Pardawala T, Pai C. A comparative study of categorical variable encoding techniques for neural network classifiers. Int J Comput Appl 2017;175(4):7–9.

[78] Lantz B. Machine learning with R. Birmingham: Packt Publishing Ltd.; 2013.

[79] Ahsan MM, Mahmud MAP, Saha PK, Gupta KD, Siddique Z. Effect of data scaling methods on machine learning algorithms and model performance. Technologies 2021;9:52.

[80] Feng C, Wang H, Lu N, Chen T, He H, Lu Y, Tu XM. Log-transformation and its implications for data analysis. Shanghai Archives of Psychiatry 2014;26(2):105–9.

[81] Akhiat Y, Manzali Y, Chahhou M, Zinedine A. A new noisy random forest based method for feature selection. Cybern Inf Technol 2021;21(2):10–28.

[82] Bouchlaghem Y, Akhiat Y, Amjad S. Feature selection: a review and comparative study. 2022.

[83] Zullig LL, Jazowski SA, Wang TY, et al. Novel application of approaches to predicting medication adherence using medical claims data. Health Serv Res 2019;54:1255–62.

[84] Alpaydn E. Introduction to machine learning. second ed. London: The MIT Press; 2010.

[85] Russell. Machine learning: step-by-step guide to implement machine learning algorithms with Python. California, US: CreateSpace Independent Publishing Platform; 2018.

[86] Galozy A, Nowaczyk S. Prediction and pattern analysis of medication refill adherence through electronic health records and dispensation data. J Biomed Inf 2020;112:1–13.

[87] Alpaydn E. Introduction to machine learning. third ed. London, England: The MIT Press; 2014.

[88] Singh D, Samagh JS. A comprehensive review of heart disease prediction using machine learning. Journal of Critical Reviews 2020;7(12):281–5.

[89] Mathew A, Arul A, Sivakumari. Deep learning techniques: an overview. In: Advanced machine learning technologies and application; 2021. https://doi.org/10.1007/978-981-15-3383-9_54.

[90] Lauffenburger JC, Yom-Tov E, Keller PA, et al. REinforcement learning to improve non-adherence for diabetes treatments by Optimising Response and Customising Engagement (REINFORCE): study protocol of a pragmatic randomised trial. BMJ Open 2021;11:1–9.

[91] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Comput 2006;18(7):1527–54.

[92] Schmidhuber J. Deep learning in neural networks: an overview. Neural Network 2015;61:85–117.

[93] IBM. AI vs. machine learning vs. deep learning vs. neural networks: what's the difference?. https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks; 2020.

[94] Panigrahi A, Chen Y, Kuo CCJ. Analysis on gradient propagation in batch normalized residual networks. https://dblp.org/rec/journals/corr/abs-1812-00342.html; 2018.

[95] Moshayedi AJ, Roy AS, Kolahdooz A, Shuxin Y. Deep learning application pros and cons over algorithm deep learning application pros and cons over algorithm. EAI Endorsed Transactions on AI and Robotics 2022;1(1).

[96] Rani KMS. A compendium of deep learning frameworks. Int J Appl Eng Res 2019;14(10):2462–5.

[97] Hinton G, LeCun Y, Bengio Y. Deep learning. Nature 2015;521(7553):436–44.

[98] Zohuri B, Moghaddam M. Deep learning limitations and flaws. Mod. Approaches Mater. Sci 2020;2:241–50.

[99] Camilleri D, Prescott T. Analysing the limitations of deep learning for developmental robotics. July 26–28, 2017, Proceedings 6. Stanford, CA, USA: In*Biomimetic and Biohybrid Systems: 6th International Conference, Living Machines 2017*; 2017. p. 86–94 [Springer International Publishing].

[100] Zhang C, Ma Y. Ensemble machine learning: methods and applications. New York, NY: Springer; 2012.

[101] Pettas D, Nousias S, Zacharaki EI, Moustakas K. Recognition of breathing activity and medication adherence using LSTM neural networks. Institute of Electrical and Electronics Engineers; 2019. p. 941–6. https://doi.org/10.1109/BIBE.2019.00176.

[102] Ntalianis V, Nousias S, Lalos AS, Birbas M, Tsafas N, Moustakas K. Assessment of medication adherence in respiratory diseases through deep sparse convolutional coding. 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA) *IEEE*; 2019. p. 1657–60.

[103] Aziz F, Malek S, Ali A, Wong MS, Mosleh M, Milow P. Determining hypertensive patients' beliefs towards medication and associations with medication adherence using machine learning methods. PeerJ 2020;8:1–23.

[104] Gao W, Liu H, Ge C, Liu X, Jia H, Wu H, Peng XA(. Clinical prediction model of medication adherence in hypertensive patients in a Chinese community hospital in Beijing. Am J Hypertens 2020;33(11):1038–46.

[105] Franklin JM, Shrank WH, Lii J, Krumme AK, Matlin OS, Brennan TA, Choudhry NK. Observing versus predicting: initial patterns of filling predict long-term adherence more accurately than high-dimensional modeling techniques. HSR: Health Serv Res 2016;51(1):220–39.

[106] Karanasiou GS, Tripoliti EE, Papadopoulos TG, et al. Predicting adherence of patients with HF through machine learning techniques. Healthcare Technology Letters 2016;3(3):165–70.

[107] Lucas JE, Bazemore TC, Alo C, Monahan PB, Voora D. An electronic health record based model predicts statin adherence, LDL cholesterol, and cardiovascular disease in the United States Military Health System. PLoS One 2017;12(11):1–17.

[108] National Health Services. Overview-statins. https://www.nhs.uk/conditions/statins/; 2022.

[109] Janssoone T, Bic C, Kanoun D, Rinder P, Hornus P. Machine learning on electronic health records: models and features usages to predict medication non-adherence. 1-5, https://arxiv.org/pdf/1811.12234.pdf; 2018.

[110] Meneveau MO, Keim-Malpass JT, Camacho F, Anderson RT, Showalter SL. Predicting adjuvant endocrine therapy initiation and adherence among older women with early stage breast cancer. Cancer Research and Treatment 2020;184(3):805–16.

[111] Yerrapragada G, Siadimas A, Babaeian A, Sharma V, O'Neill TJ. Machine learning to predict tamoxifen nonadherence among US commercially insured patients with metastatic breast cancer. JCO Clinical Cancer Informatics 2021:814–25.

[112] Lo-Ciganic WH, Donohue JM, Thorpe JM, Perera S, Thorpe CT, Marcum ZA, Gellad WF. Using machine learning to examine medication adherence thresholds and risk of hospitalization. Med Care 2015;53:720–8.

[113] Chen X, Fernandes G, Chen J, Liu Z, Baumgartner R. 1311-P: machine learning (ML) application to predict patient risk of nonadherence in Type 2 diabetes management using U.S. claims databases. American Diabetes Association 2019;68(1).

[114] Wu X-W, Yang H-B, Yuan R, Long EW, Tong RS. Predictive models of medication non-adherence risks of patients with T2D based on multiple machine learning algorithms. BMJ Open Diabetes Research & Care 2020;8(1):1–11.

[115] Fan Y, Long E, Cai L, Cao Q, Wu X, Tong R. Machine learning approaches to predict risks of diabetic complications and poor glycemic control in nonadherent Type 2 Diabetes. Front Pharmacol 2021;12. https://doi.org/10.3389/fphar.2021.665951.

[116] Gothong C, Singh LG, Satyarengga M, Spanakis EK. Continuous glucose monitoring in the hospital: an update in the era of COVID-19. Curr Opin Endocrinol Diabetes Obes 2022;29(1):1–9.

[117] Shalansky SJ, Levy AR, Ignaszewski AP. Self-reported Morisky score for identifying nonadherence with cardiovascular medications. Ann Pharmacother 2004;38(9):1363–8.

[118] Osterberg L, Blaschke T. Adherence to medication. N Engl J Med 2005;353(5):487–97.

[119] Gottlieb A, Yatsco A, Bakos-Block C, Langabeer JR, Champagne-Langabeer T. Machine learning for predicting risk of early dropout in a recovery program for opioid use disorder. InHealthcare 2022, January;10(2):223 [MDPI].

[120] Karhunen J, Raiko T, Cho K. Unsupervised deep learning: a short review. Advances in independent component analysis and learning machines 2015:125–42.

[121] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–8.

[122] Xin Y, Kong L, Liu Z, Chen Y, Li Y, Zhu H, Wang C. Machine learning and deep learning methods for cybersecurity. IEEE Access 2018;6:35365–81.

2.2 Chapter Summary

This expansive literature review chapter on the application of ML approaches to analysing MA highlights the use of various ML techniques for MA analytics. In NCD MA analytics, LR, ensemble learners, SVM, and neural nets were the most commonly used ML approaches. The review chapter reported the various data sources, including EMRs, real-time electronic data, and self-reported MA survey data. The study revealed that a high rating for MA requires it to meet a commonly acknowledged adherence level of at least 75% of the PDC. The review highlights research gaps, including the need for more robust, generalisable ML models and diverse big data integration. Subsequent chapters cover the generation of datasets and ML models for MA analytics based on the insights about diverse data sources, evaluation metrics, and ML approaches reviewed and reported in this chapter. The next chapter (Chapter 3) conducts another systematic literature review on feature selection and predictors of MA from an ML research perspective, all in the context of developing robust ML models for MA analytics.

CHAPTER 3

Systematic literature review on feature selection and importance of predictors of medication adherence from ML research perspectives (Paper 2)

3.1 Introduction

The focus of Chapter Three (3) was to identify, select, and evaluate the most significant features that impact MA in patients with NCDs as reported in existing empirical studies. The chapter commences with a synopsis of feature selection approaches frequently employed in ML, including filter methods, wrapper methods, and embedding methods, emphasising their strengths and weaknesses. The chapter subsequently examines the importance of several predictors, of MA ascertained in existing ML research to conduct an in-depth study of the features that past studies have identified as the most significant. The chapter shows how ML techniques are useful for analysing big datasets with many features and potential interconnections between them, which can be difficult to do using conventional statistical techniques.

The chapter also presents an adapted feature selection pipeline, which includes different data sources (such as EHRs, real-time medical dosage data, and questionnaire data), data pre-processing techniques, and various feature selection approaches. These data sources are crucial for conducting a comprehensive MA analysis. The chapter examines various research studies that used ML analytics to identify the potential features that contribute to the prediction of medical adherence in patients with NCDs, such as hypertension, CVDs, diabetes, and opioid use disorders (OUDs), among other NCDs. This chapter assessed feature selection algorithms from previous studies, such as recursive feature elimination (RFE), univariate and bivariate feature analysis, multivariate feature analysis, LASSO regression, Waikato environment for knowledge analysis (WEKA), SHAP, and random forest feature importance mechanisms, in order to determine the most influential features.

Paper status: Published

Paper link: https://doi.org/10.1016/j.imu.2023.101232

Journal Name: Informatics in Medicine Unlocked

Journal impact factor: NA (Q2)

Student's contribution: Conceptualisation, data collation, methodology, results and original manuscript writing

Other author's contributions: Supervision, editing, and revision

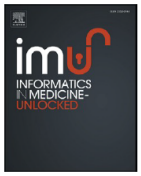
Percentage of student's contribution: 90%

Informatics in Medicine Unlocked 38 (2023) 101210

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Machine learning approaches to medication adherence amongst NCD patients: A systematic literature review

Wellington Kanyongo [a,b,**], Absalom E. Ezugwu [c,*]

[a] Department of Computer Science, Faculty of Science Engineering, Bindura University of Science Education, Bindura, Zimbabwe
[b] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg Campus, Pietermaritzburg, 3201, South Africa
[c] Unit for Data Science and Computing, North-West University, 11 Hoffman Street, Potchefstroom, 2520, South Africa

ARTICLE INFO

Keywords:
Medication adherence
Machine learning
Deep learning
Non-communicable disease
Chronic disease
NCD patients
Diabetes
Hypertension
Cardiovascular disease (CVD)
Cancer
Respiratory disease

ABSTRACT

Non-adherence to prescribed medication is a major public health concern that escalates the risk of morbidity and death as well as incurring extra expenses associated with hospitalisation. According to the World Health Organization (WHO), only 50% of people suffering from chronic diseases follow the treatment recommendations despite the counsel provided to patients on the importance of medication adherence (MA). Early detection of non-communicable disease (NCD) patients poorly adhering to recommended medications using analytics based on machine learning (ML) may improve the outcomes of NCD patients positively. This paper presents a systematic review of literature involving the application of ML in evaluating MA amongst NCD patients. The articles considered in this study were extracted from Web of Science, Google Scholar, PubMed, and IEEE Explore. Twenty-five articles in total met the criteria for inclusion. These were articles that utilised ML techniques to analyse MA in NCDs, with patients suffering from diabetes (n = 8), hypertension (n = 3), cardiovascular disease (CVD) and statin adherence (n = 6), cancer (n = 3), respiratory diseases (n = 2), and other NCD conditions (n = 3). The proportion of days covered (PDC) was typically used to evaluate MA. It emerged from the study that for MA to be considered high, the adherence threshold should be at least 75% of the PDC, a universally accepted threshold. In MA analytics research and practice, a PDC ≥80% threshold is typically regarded as a high level of adherence to prescription medication. Logistic regression (LR) (n = 12), random forest (RF) (n = 11), support vector machine (SVM) (n = 7), neural net (n = 6), ensemble learning (n = 6), MLPs (n = 4), XGBoost (n = 3), Bayesian network (BN) (n = 3), and gradient boosting (n = 3) were the most frequently applied ML techniques in the analytics of MA amongst NCD patients. It should be underscored that leveraging standard ML, deep learning (DL), and ensemble learning has enormous potential for measuring MA amongst NCD patients based on various analytics such as prediction, regression, classification, and clustering. Moreover, a further study could be conducted to comprehend how the application of alternative ML-based techniques can be used to measure MA among patients with chronic infectious diseases.

1. Introduction

Non-adherence to medication regimens that medical specialists have prescribed has been pinpointed as a major problem globally (Sabat et al., 2003; [1–4]. Medication non-adherence is associated with several adverse consequences, notably increased patient health risks, an additional financial burden for clinicians and the patients themselves, healthcare providers, and the healthcare sector, among other stakeholders [5], as well as increased morbidity, mortality, and hospital admissions [6]. Medication adherence (MA) is defined by Ref. [7] as the degree to which a patient's behaviour regarding medications is consistent with the recommended medical instruction or guidance from a healthcare expert. MA is essential for non-communicable disease (NCD) patients to achieve their expected clinical results [8]. The current systematic review focused on NCDs such as diabetes, hypertension, CVDs, cancer, and respiratory diseases which claimed approximately 33.2 million people worldwide in 2019, a 28% increase compared to year-2000 [1]. Consequently, identifying patients predisposed to poor adherence can lower the risk of hospitalisation and adverse health consequences.

The level of MA must always be assessed to detect non-adherent

* Corresponding author.
** Corresponding author. Department of Computer Science, Faculty of Science Engineering, Bindura University of Science Education, Bindura, Zimbabwe.
E-mail addresses: wkanyongo@gmail.com (W. Kanyongo), Absalom.Ezugwu@nwu.ac.za (A.E. Ezugwu).

https://doi.org/10.1016/j.imu.2023.101210
Received 3 January 2023; Received in revised form 3 March 2023; Accepted 4 March 2023
Available online 8 March 2023
2352-9148/© 2023 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Nomenclature description

AIC	Akaike Information Criterion
AI	Artificial Intelligence
ANF	Adaptive Neuro-Fuzzy
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BAASIS	Basel Assessment of Adherence to Immunosuppressive Medication Scale
B–H	Benjamini-Hochberg
BMQ	Beliefs about Medicines Questionnaire
BRT	Boosted Regression Tree
CABG	Coronary Artery Bypass Graft
CART	Classification and Regression Trees
CHAID	Chi-squared Automatic Interaction Detection
CFS	Correlation-Feature Selection
CVD	Cardiovascular Disease
DL	Deep Learning
GSES	General Self-Efficacy Scale
T2D	Type 2 Diabetes
EMR	Electronic Medical Record
FS	Feature Selection
HBM	Health Belief Model
HF	Heart Failure
HR	Hazard Ratio
IoT	Internet of Things
K-NN	K-Nearest Neighbour
LASSO	Least Absolute Shrinkage and Selection Operator
LMTs	Logistic Model Trees
LR	Logistic Regression
MA	Medication Adherence
MAE	Mean Absolute Error
MEPS	Medical Expenditure Panel Survey
MI	Mutual Information
ML	Machine Learning
MLP	Multi-Layer Perceptron
NCD	Non-Communicable Diseases
NYHA	New York Heart Association
OR	Odds Ratio
OOB	Out-of-Bag
OUD	Opioid Use Disorder
PDC	Proportion of Days Covered
PI	Predictor importance
PSSS	Perceived Social Support Scale
QUEST	Quick Unbiased Efficient Statistical Tree Algorithms
RTP	Renal Transplant Patient
RF	Random Forest
RFE	Recursive Feature Elimination
RMSE	Root Mean Square Error
RSV	Random Survival Forest
SOM	Self-Organizing Map
SHAP	SHapley Additive exPlanations
SSB	Smart Sharps Bin
ST	Survival Trees
SVM	Support Vector Machines
SVR	Support Vector Regression
TPB	Theory of Planned Behaviour
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organisation
XGBoost	Extreme Gradient Boosting

diabetes (1.5 million cases) are the next leading causes of NCD deaths after cardiovascular diseases (CVDs), which claim 17.9 million lives each year. Despite the worrisome prevalence and death rates linked with NCDs, WHO asserts that up to half of all prescribed medications for chronic illness are not used as directed [2]. Globally, there is widespread concern about improper medication use and nonadherence. For instance, 40.2% and 56.3% of patients with hypertension do not take their medication as directed [3,4]. Poor medication adherence (MA) is a significant barrier to providing NCD patients with effective clinical care. Lack of adherence to prescribed treatment plans can diminish the efficacy of therapy, negatively impacting patients' health, well-being, and ability to pay medical bills, which, in turn, places a substantial strain on the nation's healthcare budget [5].

Medication non-adherence is described as failing to take prescribed medications as prescribed or abruptly discontinuing therapy [6]. Metrics of "adherence," such as the percentage of days covered (PDC), consider the number of days during the therapy period the patient has supplied. A patient is considered "non-adherent" if their adherence measures are less than 80%, which are continuous values that are often dichotomized [7]. Adherence is also synonymous with "persistence" (sometimes referred to as "continuation"), which denotes the whole course of medication before a major halt in supply [8]. Therefore, persistence is typically understood as the time between two administrations of the same treatment before a discernible change occurs [9]. Given the global prevalence of non-adherence, it is essential to investigate the factors influencing MA in NCD patients to devise effective treatments to improve MA. To develop corrective and preventive actions, it might be helpful to identify at-risk individuals and why people don't take their medications as prescribed [10]. For years, experts have claimed that many factors, such as poor doctor relationships and a lack of thorough health literacy, can affect adherence [11]. Sometimes patients forget to refill their prescriptions, encounter undesirable side effects, and decide they no longer require the treatment [12]. Non-adherence is a phenomenon that can be attributed to a wide range of contributors, including patient-specific variables, disease-specific factors, societal and economic issues, and factors related to the healthcare system [13].

Historically, static and group-based features such as drug tolerability, diagnosis, duration of therapy, and demographics have been used to predict MA or non-adherence problems (Yeam et al., 2018; [14]. Classical modelling approaches predict health outcomes using a limited set of easily observable features and traditional statistical approaches. Using traditional statistical approaches is challenging due to the vast number of features and their possible interrelationships [11]. Clinicians prefer easy-to-use methods for identifying not just non-adherence patients but also the features that are crucial in predicting MA. Predictive analytics is a proactive and more effective method for identifying patients at risk of medication nonadherence [15]. Waiting until patients stop taking their medications before providing an intervention is a reactive approach; thus, predicting MA using machine learning (ML) is preferable for ensuring patients take their medications as prescribed. By leveraging predictive analytics, it is possible to classify patient groups based on their known predisposition to not adhere to treatment regimens. Predictive analytics makes it possible to identify patients who are likely not to follow their medication regimens early on and then focus interventions on these people to make the most use of the limited available resources. It is possible that the outcomes of patients with NCDs, such as CVDs can be improved through the use of ML predictive models for early detection of patients with poor MA [16]. The health sector's attempts to go digital over the past ten years have made a variety of data sources accessible, which can help develop prediction models for MA. Examples of such sources include electronic medical records [EMRs] [17], insurance reimbursement claims data, medicine

dispensing data from pharmacies, survey data gathered from NCD patients, and pharmaceutical prescription data from healthcare professionals [18]. It is possible to deploy multivariate ML models across specific chronic diseases by using patient clinical data, where mining big data repositories such as EMRs could help clinicians make critical decisions (Julius et al., 2021).

Machine learning (ML), a form of computational learning in artificial intelligence (AI), is used in many healthcare systems. ML offers more intrinsic benefits than traditional statistical techniques in its knowledge discovery process. This includes identifying features associated with clinical outcomes and managing complex interactions among features [19,20]. ML models have advantages over conventional statistical approaches, such as nonlinear pattern identification, less biased auto-learning, and greater flexibility to prevent overfitting [21]. Data-driven ML models can forecast MA by using specific features from patients' past medical information. These models can help with treatment planning, predict negative outcomes, and speed up the management of perioperative healthcare. According to Ref. [22]; ML approaches are appropriate for identifying intricate relationships between patient characteristics and MA. Using ML algorithms to identify relevant features for predictive analytics can help save time on time-consuming clinical activities.

MA has been extensively reviewed; however, most studies provide a limited scope of the issue, focusing on a particular disease or treatment [23]. systematic review focused on the use of ML in the context of inflammatory bowel disease. The systematic review study by Ref. [16]; which identified and summarised published predictive models that used ML to evaluate MA, was only focused on patients with CVDs and the major risk factors associated with the CVD condition—hypertension and hypercholesterolemia. Unlike the systematic review conducted by Ref. [16]; which was limited to one condition (CVD) and a relatively small number of studies (12), the current presented systematic review was more comprehensive because its scope covered multiple NCDs (diabetes, CVDs, respiratory diseases, and OUD, among others) across more than 25 published articles. This meant that our research was positioned to report a broad range of risk factors for MA in patients with various NCDs. Despite efforts to review the literature on the evaluation of MA using ML among NCD patients, some of these studies were scoping or narrative reviews as opposed to systematic reviews [24,25]. The current systematic review is more thorough and employs repeatable systematic and explicit methods to identify, select, and critically evaluate relevant research and collect and analyse data from the studies included in the review based on well-formulated questions. In contrast, by its very nature, the scoping review only allowed [25] to review, classify, and summarise literature related to the use of ML for MA evaluation [24]. research is a narrative review that provides an overview of the state of AI in evaluating and optimising MA in NCD patients. Narrative reviews are less formal than systematic reviews because they do not necessitate the presentation of the more rigorous aspects of a systematic review, such as reporting methodology, search terms, databases used, and inclusion and exclusion criteria [26].

Again, the review study by Ref. [27] was a bibliometric analysis rather than a systematic review targeted at ML and blockchain technology applications in the smart healthcare industry. Though their study reviewed literature with an ML component, the articles reported on the countries with the highest output, the major research subjects, funding funds, and research hotspots in this field. However, this study was focused explicitly on articles incorporating the analytics of FS and the importance of predictors of NCD MA from ML research perspectives. The bibliometric analysis visually examines the research area by reviewing data from the database, such as titles, abstracts, keywords, and references [27]. Other studies examined MA monitoring technologies [28, 29]. [28] conducted a review that concentrated on recent advances in technology-based medication adherence approaches, emphasising their technical monitoring aspects. Their study's taxonomy space includes various techniques, such as sensor systems, proximity sensing, vision systems, and combinations of these without the use of machine learning (ML) [29]. conducted a scoping review and bibliometric analysis of big data applications for MA to map the current state-of-the-art MA measurement and assessment techniques used in chronic conditions. Their research was particularly interested in the methods, and assessment procedures presently used to manage MA.

Similarly [30], focused on providing a summary of techniques to evaluate MA based on a review of the state of measuring MA via a literature review study [31]. reviewed the techniques commonly used for adherence evaluation and presented a case study assessing the variables driving long-term adherence. In contrast to Ref. [28]; who focused on the technical aspects of MA monitoring technologies, and [29]; who focused on MA measurement and assessment methods used in chronic conditions, the novelty of this systematic review is evident in systematically reviewing articles that specifically incorporate ML while also demonstrating FS and the importance of MA predictors in NCD patients. The current review's taxonomy space covers multiple NCDs, including hypertension, CVDs, diabetes, OUD, and other NCDs. This is consistent with the findings of [25]; who observed a dearth of research incorporating the use of ML to assess how well NCD patients adhere to medication. The authors' resolve to conduct the systematic review study on FS and the importance of predictors of NCDS from ML research perspectives was reaffirmed when we discovered that some of the existing review studies only examined the literature on risk factors for MA or non-adherence without taking ML into account [32–34].

Even though many different features may affect MA [35]; Seng et al., 2020), only a few studies have examined how these factors affect MA. It is critical to note that the features that affect MA differ from one study setting to the next. This is due to differences in the type of NCDs being investigated, feature selection techniques applied, culture, region, level of economic development, education level, social desirability bias, living habits, living in urban or rural areas, social and healthcare systems, and other factors [36,37]. [21] claim that based on the research design, patient profile, data collection, and analysis, the results may be contradictory because characteristics such as gender, smoking status, and degree of education may indicate a significant association in one scenario but are irrelevant in another. These issues make developing accurate adherence prediction models and patient interventional programmes more challenging. Therefore, it is crucial to examine the key factors that affect MA and their importance in MA models across diverse NCDs and regions of the world. Past research has employed regression-based methodologies to explore the effects of several patients' and provider-level factors on adherence and health outcomes [38, 39]. However, there is limited data to support the use of ML algorithms to predict medication nonadherence early among NCD patients, especially regarding comprehensive literature review studies. Moreover, a few papers have discussed the application of ML approaches in evaluating features associated with MA among NCD patients [22]. Accordingly, we embarked on undertaking a systematic review around feature selection (FS) and the importance of predictors of NCD MA from an ML research perspective. Understanding a population's current level of adherence and the factors that affect adherence is critical for controlling healthcare costs [31]. Thus, the fundamental objective of this study is to answer the question: *What ML research has demonstrated FS and feature importance in predicting MA among NCD patients?* This main research question is further broken down into three specific research questions:

1. What potential features contribute to predicting MA or non-adherence among NCD patients in ML research?
2. What ML algorithms are used in research incorporating feature selection and feature importance screening?
3. In ML-based research, are there any techniques for FS and evaluating the contribution of predictors to NCD MA or non-adherence?

The specific technical contributions listed below illustrate this paper's value proposition:

- Since non-adherence rates are so high, it is essential to predict MA and its features so that specialists can improve patient monitoring and therapy as necessary.
- To demonstrate how well ML works in the FS process so that many predictors can be made that can help with proactive MA prediction and clinical decision-making. This aids medical practitioners in developing intervention plans, identifying the features of non-adherence, and delivering individualised care based on the underlying conditions of each patient group.
- By correctly identifying patients likely to have low MA, linked features may help to develop effective ways to boost MA. The results have implications for clinical decision-making and the management of healthcare organisations.
- By examining features common to many NCD conditions and using FS and importance ranking techniques, analysts can take the first step toward making automated ML models and applications to deal with medication non-adherence in different populations and patients with different NCD conditions. These can then be used as important decision-support tools in various clinical settings.
- Due to the diversity and inherent complexity of how patients take their medications, it is essential to understand the relationships between MA and the associated features. This would make it easier for doctors, caregivers, patients, and administrators of healthcare systems to get beyond the limitations and challenges brought on by the present non-intelligent drug adherence and intervention systems. For ML applications in clinical practice and research, understanding the optimum features set for predictive modelling in MA evaluation is crucial.
- Due to the growing number of scientific papers, it is now almost impossible for a scientist to keep up with all the new study findings in the scientific literature about ML-based analytics of MA among NCD patients. As a result, the body of scientific literature has become fragmented, with various scientists focusing on sections related to their particular fields of interest. Interestingly, this paper collects, unearths, and summarises a substantial body of literature on ML-based analytics of MA in NCD patients, focusing on MA predictors. The results of the study help researchers create and test new research hypotheses. Scientists will be inspired to find new research fields due to the time and effort savings provided by our study.

2. Research methodology

To ensure that the subject of the study is thoroughly addressed while also minimising the amount of bias incorporated into the systematic review process, a standard approach for performing a systematic review in computer science was selected and followed. To that purpose, this systematic review follows the guidelines [40] laid out for conducting such studies in computer science. The research questions, keywords for literature searches, search strings, and search engines were defined following the guidelines by Ref. [40]. The search strings were improved and executed as part of the article retrieval process, and the search results were downloaded and saved. The article selection method generally comprises defining the inclusion and exclusion criteria, selecting papers to evaluate, and finally, extracting the results relating to the research questions.

2.1. Article retrieval process

The keywords used in this study were directly related to its main research question: "What ML research has shown feature selection and feature importance in predicting MA among NCD patients?" Based on this research question, five keywords were retrieved and used to create the search strings. The key search terms include "machine learning", "feature selection", "feature importance", "medication adherence", and "non-communicable disease". By combining the main search phrases with the Boolean operator "AND," our search conditions were improved. Search terms like "machine learning feature importance and non-communicable disease medication adherence", "machine learning feature selection and non-communicable disease medication adherence", and "machine learning and non-communicable disease medication adherence" were generated as a result. Considering that different writers use different terminologies for the same phrase, we extensively and exhaustively searched the necessary literature using previous research synonyms and alternative terms that matched our initially determined keywords. In our thorough literature search, the term "feature" was therefore akin to "predictor," "variable," "factor," "dimension," "attribute," and "driver". Chronic disease was used as a synonym for non-communicable disease during the article retrieval process. Once more, "medication adherence" was interchangeable with "medication compliance," "drug adherence," "medication persistence," and "treatment adherence." Guided by this, the electronic literature databases created different search options that gave the researchers access to a sizable number of research papers.

Articles were retrieved from Google Scholar, Research4Life, IEEE Xplore, and PubMed between November 1, 2022, and December 31, 2022. Research4Life is an effective and highly valued literature database that allows access to scientific research articles in subscriber organisations. The essential source of reliable scientific data for the research community is subscription-based journals. Interestingly, PubMed is a huge, reliable, authoritative repository for information in the medical and health sciences, including current research incorporating ML techniques. The Institute of Electrical and Electronics Engineers (IEEE) and its affiliates publish a lot of scientific and technical literature that is easy to find and read through the IEEE Xplore platform. Google Scholar offers studies from various areas, including scientific, technical, and medical disciplines and computer science, where ML emerged.

2.2. Article selection process

If an article's full text was available in English and published in a reputable peer-reviewed journal, it was considered eligible. The articles were expected to have utilised at least one ML technique for MA prediction amongst NCD patients. The primary focus of inclusion was on articles that clearly exhibited the determination of MA predictors or demonstrated FS or feature importance in ML research for analysing MA among NCD patients. As a result, articles that may have used ML for MA predictive modelling but did not report any MA predictor were excluded from the review. Articles relying solely on LR to assess MA in NCD patients were excluded. They were only included if they would have been used in conjunction with typical ML methods such as tree-based algorithms, boosting algorithms, and neural nets. Though LR can be considered an ML algorithm, existing literature acknowledges that typical ML methods are superior and are actually used as an alternative to traditional LR because of their capacity to model the intricate interactions among the predictors [41]. ML models have the strengths of including nonlinear associations, less biased auto-learning, and higher flexibility to avoid over-fitting. Opinions, keynote addresses, editorials, and PowerPoint presentation slides were ineligible to provide access to comprehensive and complete study information. The expanded search generated substantial results from the earmarked electronic literature databases using the inclusion and exclusion criteria for eligibility as a guide. The initial searches produced two thousand eight hundred sixty-four (2,864) hits. During the screening process, 2602 duplicate papers and irrelevant publications were removed. A further 235 were excluded for not meeting the requirements, especially for lacking the application of ML and reporting predictors of MA. Finally, the review's overall scope encompassed a total of 27 papers.

3. Machine learning

A subfield of AI known as ML explains a computer's capacity to identify patterns and extrapolate meaning from various organised,

unstructured, and semi-structured data (Samer & Nour, 2019). ML aims to build computer systems that can adapt and learn from their experiences [20]. The healthcare system generates enormous volumes of data that can inform clinical decisions, just like other disciplines such as engineering, education, marketing, retail, social media, and financial services. These big datasets are required when building ML models for analysing MA. Recently, several ML algorithms were implemented for enormous volumes of clinical data to create healthcare systems and models that can accurately identify patterns, classify cases, and make predictions. ML has been used in healthcare systems to identify specific clinical features and develop data-driven decision rules [42]. ML algorithms can help identify individuals who should undergo additional evaluation and help prevent time-consuming routine clinical activities [43]. In feature engineering, an ML model is typically an agent trained with a set of predictors to produce a desired result [10].

Recent advances in ML algorithms used in the healthcare sector have improved model classification and prediction accuracy [44,45]. The major problems in MA evaluation that ML addresses are classification, regression, and prediction. ML algorithms can accomplish the conventional tasks of object recognition and classification. ML algorithms categorise future datasets using techniques based on previously classified training datasets. In ML classification, algorithms conduct analytical tasks that would take a human a lot of time to complete. ML algorithms, such as RF, SVM, MLPs, LSTM, DT, Naive Bayes, neural networks, XGBoost, and Adaptive Neuro-Fuzzy [ANF], have been used widely to classify or predict MA [2,12,41,46–51]. According to Ref. [52]; K-nearest neighbours, Naive Bayes, deep neural networks, and random forest (RF) are among the most frequently used ML algorithms for achieving the best model performance in MA evaluation.

3.1. Naive bayes

The Naive Bayes algorithm belongs to the category of supervised learning algorithms. It is a straightforward classification technique that allows for the rapid development of classification models. To make predictions, Naive Bayes employs the probabilities that each feature belongs to each class [47]. With a strong (naive) assumption that the features are conditionally independent, naive bayes classifiers are a form of straightforward "probabilistic classifier" in ML [53]. The assumption used in the Naive Bayes algorithm makes computing probability easier. It determines whether a data point will likely fall into a particular class. The Naive Bayes approach accurately and quickly predicts the class of the test data set in a multi-class prediction scenario. With parameters that are proportional to the number of features in a learning task, naive bayes classifiers are highly scalable. It is vital to keep in mind that the naive Bayes model incorporates the Bayes theorem. For this reason, the Bayes rule provides a decomposition of a conditional probability and is commonly employed within the bayesian learning family of learning approaches. Bayes rule is the equality.

$$P(z|w) = \frac{P(z)P(w|z)}{P(w)}$$

where P(w) denotes the prior probability, P(w |z) the posterior probability, and P(z |w) the likelihood. The Bayes theorem can be used to calculate the probability of z occurring, given that w has already occurred. Here, w stands for the evidence, and z stands for the hypothesis. The predictors are assumed to be independent and have a similar impact on the outcome in this situation. In other words, the existence of one feature has no bearing on the presence of another. It is hence referred to as naive. Z is the class variable (adherence) in measuring MA among NCD patients, indicating whether or not a patient is taking the prescribed medication. The attribute indicated by the feature w can include, among other features, the patient's age, gender, and medication frequency. In a prior research study [50], used the Naive Bayes classification algorithm to classify MA in patients with mobility impairments. The ML demonstrated its ability to distinguish between patients with movement disorders on and off medication with excellent model accuracy. With a remarkable accuracy of as high as 81% in one instance [47], classified a patient as being medication adherent or not in a study that employed the Naive Bayes algorithm among other ML algorithms.

3.2. K-nearest neighbours k-NN

The K-Nearest Neighbours (k-NN) method employs "feature similarity" or "nearest neighbours" to predict which cluster a new data point belongs to. The k-NN is named after the number of neighbours it discovers that is closest to a given one. One way to determine the likelihood that a data point belongs to a certain cluster is to look at the cluster to which the physically closest points belong. Assigning data to its nearest neighbour's cluster requires computation when using k-nearest neighbours for classification. Even though the k-NN technique can be applied to both classification and regression tasks, its most typical use is for the former. It is given to the class closest to 1 if k = 1. To classify K, its neighbours are polled in a plurality manner. Fig. 1 demonstrates the k-NN algorithm.

The procedure begins with the selection of the number K of neighbours, then moves on to the computation of the Euclidean distance between the K number of neighbours, then moves on to the selection of the K nearest neighbours based on the Euclidean distance, which is the distance between two points, then moves on to the counting of the number of data points in each category among the K neighbours, and finally moves on to the assignment of the new data points to the cluster that has the highest number of neighbours. Fig. 1 illustrates that, compared to category B, the three nearest neighbours are from category A; hence, this new data point belongs to group A.

3.3. Decision tree

Decision trees (DTs) are supervised learning suitable for classification problems because they can organise classifications at a granular level. Since DTs can handle numerical and categorical data, they can be used for classification and regression tasks. It functions similarly to a flow chart, breaking data points into two categories that are comparable to one another at a time as it progresses from the "tree stem" to the "branches" to the "leaves," where the categories become more finitely identical to one another. As a result, categories within categories start to develop, enabling organic classification with little human intervention. In a tree-structured classifier, the internal nodes represent the dataset's features, the branches represent the decision rules, and each leaf node represents the classification result. As the tree grows, the dataset is subdivided into progressively smaller groupings or nodes. A decision tree's output comprises leaf nodes and decision nodes. A decision node is characterised by two or more branches, whereas a leaf node represents the decision itself. The term "root node" refers to the node at the very top of the tree and corresponds to the most accurate prediction. In previous research, DT was used to determine which individuals with opioid use disorder (OUD) would eventually withdraw from their therapy [41] and construct ML models to classify MA in hypertensive patients with high classification performance [27].

3.4. Support Vector Machines

A supervised ML technique called the Support Vector Machine (SVM) analyses data in classification and regression tasks. SVMs are most frequently used for classification, although they can also be used for regression. The SVM is a statistical method for classifying data that does not lend itself to linear separation by mapping it to a high-dimensional feature space [54]. Each data point in an SVM method is represented as a point in N-dimensional space, with each feature denoting the value of a particular coordinate. As a result, they construct a hyperplane in multidimensional space to identify different classes. These support

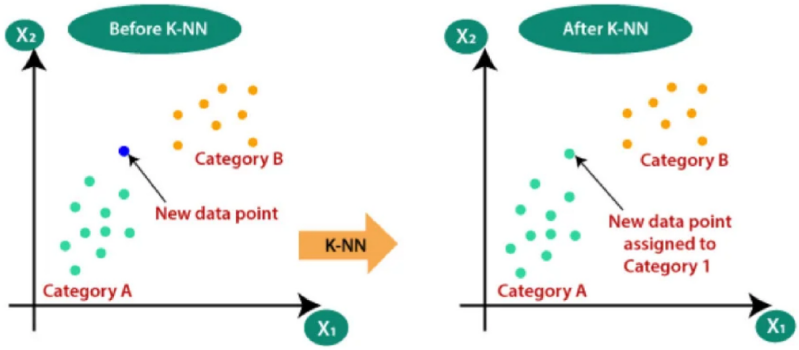


Fig. 1. k-NN algorithm procedure.

vectors are a representation of each observation's coordinates. The SVM represents the dataset as a collection of data points in space separated by a line or gap that is as large as is practically possible. Depending on which side of the line or separation the additional data points fall on, they are now mapped into the same area and classified as belonging to one of several classes. Based on the features of the newly acquired data, it is possible to deduce which classification scheme a new data point should belong to. The hyperplane or decision plane between any two classes is the best way to separate or classify them. It is a frontier approach to class separation.

Fig. 2 shows that the main goal is to get as much difference as possible within the available dataset. Margin refers to the difference in distance between the two points that are physically closest to one another. In the given dataset, the objective is to identify the hyperplane with the maximum possible margin between support vectors. Consequently, one starts by creating hyperplanes that best segregate the classes; the left-hand figure in Fig. 2 above illustrates three hyperplanes: black, blue, and orange. Next, SVM searches for the greatest marginal hyperplane. While the black accurately distinguishes between the two classes, the blue and orange show higher classification errors. Secondly, as can be seen in the right-hand figure of Fig. 2, the chosen right hyperplane is the one that has the greatest distance from the data points that are physically nearest to it.

3.5. Random forest (RF)

RFs are ensemble classifiers because they consist of numerous decision trees [47]. The RF approach is a decision tree development method in which many decision trees are first constructed using training data before new data is fitted into one of the trees as an "RF." For solving regression problems, RF builds a lot of decision trees, calculates the mean of all the decision trees, or uses majority voting when solving classification problems. One of the classes is chosen by each tree. The eventual classification of a sample is selected by employing a voting system that requires a simple majority. The class that is predicted to be the most popular among the classified samples is the one that receives the most votes from the trees. A new subset of samples is picked from the dataset, and each tree in the forest can grow to its maximum magnitude without being pruned in the process. A specific group of attributes is used at each tree branch node. The cardinality of this subset of features can be calculated by taking the square root of the total number of features present in the dataset. The number of selected distinct features does not change while the forest is developed.

3.6. Neural nets

Deep learning (DL) takes several forms, and one of them is represented by neural networks. Deep learning (DL) is a subfield of ML that uses artificial neural networks to extract deeper layers of data features. Artificial neural networks are a subfield of artificial intelligence inspired by biology and modelled after the human brain. Neural networks are modelled after how the human brain processes information and imitate the brain's fundamental functions. It handles a range of real-time tasks due to its capacity to do computations quickly and respond swiftly. A neural network is composed of many layers, each containing many units. With interconnected processing neurons (nodes), these layers are broken down into the input layer, hidden layer(s), and output layer to mimic the human brain. The artificial neural network's input layer is where data from the outside world enters. Data processing nodes analyse or classify data before sending it to the following tier. The input layer or other hidden layers provide the output for hidden layers. It is conceivable for ANN to contain a considerable number of hidden layers. Before being delivered to the hidden layer underneath it for additional processing, the output of each hidden layer is evaluated by the layer that preceded it. The results of the most recent data processing performed by the ANN are shown at the output layer. There could be one or many nodes in it. For

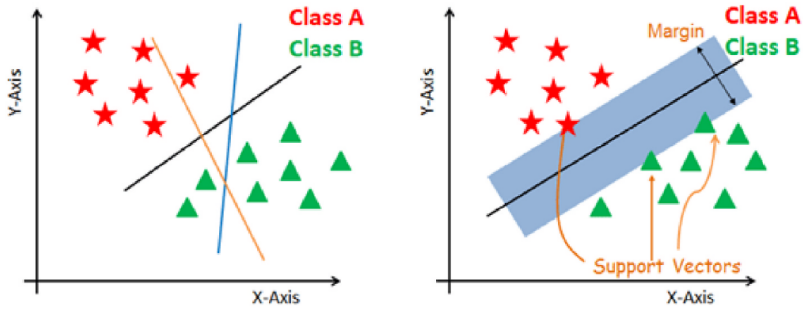


Fig. 2. SVM application.

example, if the classification problem is binary (yes/no), one output node in the output layer will return a value of either one or zero. This node will be present in the output layer. There may be more than one output node in the output layer if the problem is multi-class classification. The objective, such as identifying an object in an image or identifying the class of an object is performed by the neural network's final output neurons.

It is important to remember that each node, also called an artificial neuron, is connected to another and has a threshold and weight that go with it. Any node in the network whose output goes above a certain limit triggers activation and begins sending data to the subsequent network level [54]. If not, no data is transmitted to the following network layer. Three types of neural network applications for tackling real-world problems include pattern categorisation, prediction, and control and optimisation, which cover many scientific domains [55]. These networks can potentially be used in developing the next generation of computer systems and are now used for complex investigations in various fields, from engineering to medical sciences [56]. A neural network with many layers is said to be densely layered. Each unit in one layer of a neural network is connected to each unit in the layer below it in a neural network with numerous layers. The visual representation that is produced by a deep neural network is illustrated in Fig. 3.

Deep learning (DL) refers exclusively to the number of layers within a neural network. A "DL algorithm" is a neural network having more than three layers, including inputs and outputs [54]. A number represents the weight of the connections that are made between the different nodes. If one node stimulates another, the weight is positive; however, if one node inhibits another, the weight is negative. Nodes that have higher weight values can exert a more significant impact on other nodes. In theory, deep neural networks can transform any input into any output. However, they require significantly more training than conventional ML methods. They require millions of training data samples instead of the hundreds or thousands necessary for a more straightforward network or algorithm. Different neural networks can be differentiated by how information moves from the input node to the output node. These include feedforward neural networks and backpropagation algorithms. Data processing in feedforward neural networks only occurs in one direction, from the input node to the output node. In the hierarchical structure, every node in one layer is connected to every node below it.

A **feedforward network** uses a feedback loop to enhance predictions over time. Once the input layer has been established, weights are assigned. These weights assist in establishing the significance of each feature, with larger weights having a greater influence on the output than smaller ones. Then, after multiplying each input by its weight, the total is calculated. The output is then processed by an activation function, which determines the output afterwards. When the node's output reaches a certain threshold, it "fires," also known as "activates," which causes it to send information to a higher network layer. As a result, the input of one node becomes its output. A feedforward network, which transmits data from one layer to the next, is the name given to this neural network [54].

When considering the **backpropagation algorithm**, it is essential to remember that neural networks constantly train themselves by utilising corrective feedback loops to enhance their prediction analytics. This is a crucial consideration to keep in mind. In simpler terms, information travels from the input node to the neural network's output node via various paths. Only one proper path exists between the right input and output nodes. The neural network utilises a feedback loop to reach this conclusion. Each node in the path predicts the next node in the path, determines whether or not the prediction was accurate, and then assigns higher weight values to the node paths that produce more accurate predictions and lower weight values to the node paths that produce inaccurate predictions. Step 1 is repeated once the network nodes have produced a new prediction for the next data point using the higher-weight paths.

4. Feature selection

In many ML applications, FS is a key pre-processing step that helps solve problems with high dimensionality. FS is a typical ML process known as "subset selection." It involves the selection of a subset of features from the available data to apply a learning algorithm [57]. FS is the process of choosing the most important features and eliminating the unimportant, redundant, and noisy ones from the initial full feature set based on an evaluation criterion, reducing the feature set to those most important or relevant to the ML model [58,59]. Choosing the most significant features makes it feasible to reduce the complexity of prediction models in high-dimensional datasets. This is critical for reducing data noise and improving model precision and interpretability [60]. To determine if a patient follows their medical regimen, just the most significant features from a collection of medical variables should be considered. FS can be performed supervised or unsupervised, using regression or classification algorithms. As a result, FS, which entails identifying relevant features and removing unimportant features, is typically motivated by having a high-performing learning algorithm across the test set. The computation of feature importance scores, which indicates that some features are more important to the model than others, can serve as the foundation for FS. How much a feature adds to the model's prediction determines its relevance. It establishes how useful a particular feature is for prediction and the current model.

Most of the time, the importance of a feature is measured by a score, and the score's value goes up in proportion to how important the feature is. The existence of a feature significance score has many benefits. For

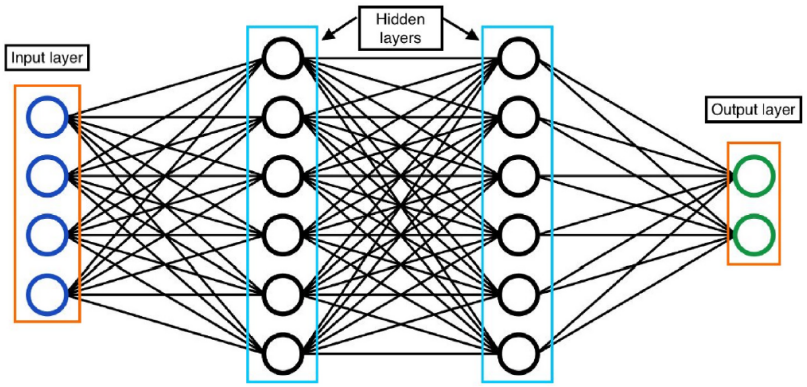


Fig. 3. Deep neural network architecture.

instance, it is possible to identify the relationship between independent features and dependent variables (targets). By analysing feature importance rankings, one can identify and remove unnecessary features. The model's performance might be accelerated or even improved by reducing the number of irrelevant features. Additionally, feature importance is often used for ML model interpretability. The results can illustrate why the ML model produces particular predictions and how features might be changed to vary those predictions. The advantages of FS include: reducing the number of features, improving model accuracy if the right subset is chosen, minimising overfitting and improving model generalisation, improving understanding of the features and their relationships to the response variables, lessening the complexity of a model and making it easier to interpret, limiting storage needs, and possibly assisting in data reduction [61]. There are two FS algorithms: supervised (for use with labelled data) and unsupervised (for use with unlabelled data). The supervised FS approaches are filter, wrapper, and embedded methods. The input and output under consideration establish the best FS strategy. For a dataset having numerical input and output, one can use Pearson's and Spearman's correlation coefficients, for instance, in an FS regression problem. The LR, Kendall's rank coefficient, and the ANOVA correlation coefficient are all relevant when the dataset has a numerical input variable and a categorical target variable. Classification-predictive modelling problems with categorical input variables can be handled using the chi-squared test or mutual information (MI), a powerful data-driven strategy.

4.1. Feature selection pipeline

An FS procedure typically consists of four steps: data pre-processing; subset generation, where candidate feature subsets are chosen based on particular search techniques; selecting the best or optimal feature subset through an evaluation function computation procedure, where the chosen subset is either specified or retained; and finally, result validation, where the performance is verified. The first phase in the FS pipeline is data pre-processing, which entails choosing the dataset that will be used for ML modelling. The dataset for the MA study might have been produced using real-time medical dose data, past EMRs, or questionnaires for the MA analysis. The dataset is then given a pre-processing data treatment, which includes removing duplicates (eliminating duplicate data in a data set by deleting an additional copy of a datapoint or feature and leaving just a single column or row of the duplication). This technique involves comparing the rows' columns to one another.

As part of the pre-processing data stage, missing values for the categorical variable could be replaced by the variable's modal value. In the case of discrete features, the most typical value should be used to fill in any missing values, whereas the average value should be used for continuous features. The likelihood of misclassification might rise as a result. To solve this problem, two solutions are suggested. If too many missing values are above the threshold, one option is to do away with the functionality altogether (which is user-specific). That's because there's a limit to how much information can be retrieved during the learning process. Second, the location of the missing value can be determined by employing a finite difference measure. Correlation measurements allow for a fast and precise feature selection when the missing values have been populated with their right counterparts. With imputation, a dataset can keep most of its original data while filling in the gaps left by missing values [62]. In other situations, variables or variable rows with a high percentage are deleted. Missing values can cause ML models to produce skewed results and reduce their accuracy [47]. conducted a study in which they eliminated features that were missing more than sixty percent of their values. This was done since the researchers determined that it was not possible to impute the missing values. Eliminating records that contain outliers or considering them missing values and replacing them with the proper imputation technique are two options for dealing with outliers [63]. If not addressed, these outliers have an adverse effect on model performance and must be handled cautiously. The dataset's excessively low or high values are known as outliers. These values may cause measurement errors in the model, biased parameter estimates, and incorrect findings. A cleaned candidate feature dataset is produced by the data-cleaning pre-processes.

The candidate feature dataset is the dataset that contains all features that are eligible for submission through the FS algorithm and have a chance to be selected as part of the final feature subset based on relevance. In this study, the initial full candidate feature set is named as a target population set of features that must be further refined into a feature subset or sample subset. The remaining steps involve applying various FS techniques, including the filter FS, wrappers, and embedded method. If the filter strategy is utilised, the next step is to choose the best or ideal feature subset by first rating features individually according to a particular criterion measure, like correlation-feature selection (CSF). In order to determine which features have the greatest rating, a cut-off value is employed. In the succeeding phase of the MA evaluation research, the selected features are input into an ML model in order to predict the variable target class, which may be MA or non-adherence. In the wrapper technique, the FS algorithm functions as a wrapper over the predictive model algorithm, choosing the best features based on the same model. The wrapper approach selects features by repeatedly selecting a subset of features and evaluating their importance based on their performance, such as the accuracy of a model generated using the pre-selected learning algorithm. This procedure is continued until all desirable features have been selected. Classification and FS can happen simultaneously with embedded FS methods since FS is incorporated into the learning algorithm. Among the FS techniques are the generation of subsets, the evaluation of subsets, criteria for determining when to halt the search, and validation procedures. Fig. 4 illustrates the FS process flow.

An FS procedure typically consists of four steps: data pre-processing; subset generation, where candidate feature subsets are chosen based on particular search techniques; selecting the best or optimal feature subset through an evaluation function computation procedure, where the chosen subset is either specified or retained; and finally, result validation, where the performance is evaluated. The first phase in the FS pipeline is data pre-processing, which entails choosing the dataset that will be used for ML modelling. The dataset for the MA study could have been produced using real-time medical dose data, retrieved from past EMRs, or collected through questionnaires for the MA analysis. The dataset is then subjected to data pre-processing, which includes removing duplicates (eliminating duplicate data in a data set by deleting an additional copy of a datapoint or feature and leaving just a single column or row of the duplication). This technique involves comparing the rows or columns to one another.

During the pre-processing stage, missing values could be filled in with the modal value of a categorical variable. The most frequent value should be used for discrete features to fill in any missing values. For continuous features, on the other hand, the mean score should be used. The likelihood of misclassification might rise as a result. To solve this problem, two solutions are suggested. If the number of missing values exceeds a certain threshold, one option (which is user-specific) is to remove the feature entirely. This is due to the fact that there is a limit to how much information can be retrieved during the learning process. Secondly, a finite difference measure can be used to identify the location of the missing value. Correlation measurements enable quick and precise feature selection when missing values are populated with their related counterparts. With imputation, a dataset can keep most of its original data while filling in the gaps left by missing values [62]. In other cases, features or variable rows with a high percentage of missing values are removed entirely. Missing values can cause ML models to produce skewed results and reduce their accuracy [47]. conducted a study in which they eliminated features that were missing more than sixty percent of their values. This was done since the researchers determined that it was not possible to impute the missing values.

Two ways to deal with outliers are to eliminate the records that have

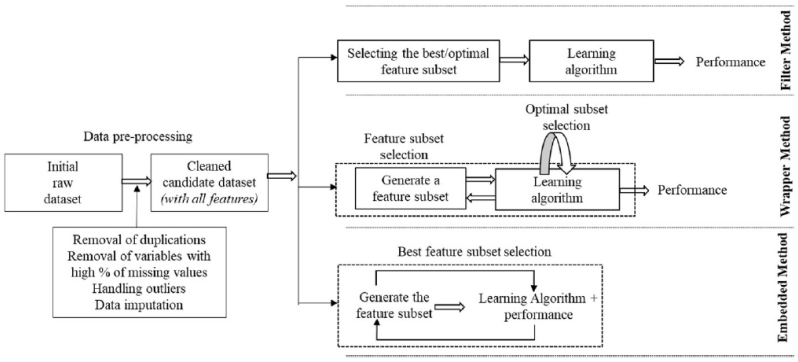


Fig. 4. Feature selection pipeline.

them or to treat them as missing values and fill them in with the right imputation technique [63]. These outliers have a negative impact on model performance if they are not addressed, so they must be handled with caution. Outliers are numbers that are abnormally low or high in the dataset. These values may cause measurement errors in the model, biassed parameter estimates, and incorrect findings. A cleaned candidate feature dataset is produced by the data-cleansing pre-processes.

The candidate feature dataset is the set of features that can be submitted to the FS algorithm and have a chance of being chosen as part of the final feature subset based on how relevant they are. In this study, the initial full candidate feature set can be considered a target population set of features that must be refined into a feature subset or sample subset. FS techniques, such as filter FS, wrappers, and the embedded method, are used in the remaining steps. If the filter strategy is used, the next step is to select the best or ideal feature subset by rating each feature individually using a specific criterion measure, such as correlation-feature selection (CSF). A cut-off value is used to determine which features have the highest rating. In the succeeding phase of the MA evaluation, the selected features are input into an ML model in order to predict the variable target class, which may be MA or non-adherence. In the wrapper technique, the FS algorithm serves as a wrapper around the predictive model algorithm, selecting the most important features based on the same model. The wrapper approach selects features by repeatedly selecting a subset of features and evaluating their importance based on their performance as a subset. The accuracy of the model produced using the preselected learning algorithm is used to assess the model's performance in each iteration. This procedure is continued until all desirable features have been selected. Classification and FS can happen simultaneously with embedded FS methods since FS is incorporated into the learning algorithm. Fig. 4 illustrates the FS process flow.

4.1.1. *Filter selection approach*

Filter methods pick features based on statistics rather than FS cross-validation performance. A chosen metric is employed to find unimportant features and recursive FS. Filter methods are frequently used as an initial pre-processing step without an induction algorithm. Filters assess feature significance by utilising data-intrinsic features. Because one uses the chosen metric to find irrelevant features, this FS component is called "filter-based". By relying on the dataset's attributes to determine a feature's significance, filter methods analyse and pick features independently of any learning algorithm [58,64]. Rather, features are selected depending on their performance in some statistical tests for correlation with the outcome variable. It is based on numerous assessments of the general features of the training data, including the distance between classes, the information, and other statistical dependencies, such as Pearson correlation, consistency, and entropy [65]. According to Microsoft (2021), a single method for figuring out feature significance is given in filter selection methods; this could include correlation approaches like Pearson correlation and chi-squared values. The component creates a dataset with the highest-scoring feature columns arranged by predictive power, and the predictive model is frequently built using the feature sets with the strongest scores.

There are two steps to the filtering process. In order to impact the choice of a feature subset, filter methods can either be univariate (where an ordered ranking list of features is formed) or multivariate (where the features are evaluated for relevance as a whole, identifying redundant and irrelevant characteristics). The classification model is created by first ranking each feature according to a particular criterion measure. Second, a threshold value is applied to determine which features have the highest overall rank. During model development, the remaining variables are deemed weak and irrelevant. The selected group of attributes is then transmitted to the induction classifier. Enhancing the classification process's efficacy and precision is possible by utilising the relevant feature subset. To get a smaller set of characteristics that other, more expensive FS methods may then process, filter methods are necessary for situations where the number of features is excessively large. It typically performs faster and better than the wrapper approach when working with high-dimensional datasets, and because it is independent of the induction algorithm, it produces higher generalisation [66]. A threshold is necessary to choose a subset since it prefers to select subsets with many attributes, sometimes even all [61]. Due to the absence of consideration for the relationships between the independent variables, it is probable that redundant features will be chosen. This approach produces a subset of highly generalisable features but is unsuitable for any specific learning algorithm because they are not biased toward any learning algorithm. The work flow for the filter approach is shown in Fig. 5 below.

Filter methods select features based on statistics rather than FS cross-validation performance. A chosen metric is employed to segregate unimportant features. Filter methods are frequently used as an initial pre-processing step without an induction algorithm. Filters assess feature significance by utilising data-intrinsic features. Because one uses the chosen metric to find irrelevant features, this FS component is called "filter-based". Filter methods analyse and select features independently of any learning algorithm by relying on the datasets' features to determine a feature's significance [58,64]. Rather, features are selected depending on their performance in some statistical tests for correlation with the outcome variable. It is based on numerous assessments of the general features of the training data, including the distance between classes, the information, and other statistical dependencies, such as Pearson correlation, consistency, and entropy [65]. According to

Fig. 5. Filter method workflow.

Microsoft (2021), a single method for figuring out feature significance is given in filter selection methods; this could include correlation approaches like Pearson correlation and chi-squared values. The approach creates a dataset with the highest-scoring feature columns arranged by predictive power, and the predictive model is built using the feature sets with the strongest scores.

There are two steps to the filtering process. In order to impact the choice of a feature subset, filter methods can either be univariate (where an ordered ranking list of features is formed) or multivariate (where the features are evaluated for relevance as a whole, identifying redundant and irrelevant features). The classification model is created by first ranking each feature according to a particular criterion measure. Secondly, a threshold value is applied to determine which features have the highest overall rank. During model development, the remaining variables are deemed weak and irrelevant. The selected group of features is then transmitted to the induction classifier. It is possible to enhance the classification process's efficacy and precision by utilising the relevant feature subset. Filter methods suit situations where the number of features is excessively large. It typically performs faster and better than the wrapper approach when working with high-dimensional datasets, and since it is independent of the induction algorithm, it produces higher generalisation [66]. A threshold is necessary to choose a subset since it prefers to select subsets with many attributes, sometimes even all [61]. It is likely that redundant features will be selected due to the lack of consideration for the relationships between the independent variables. This approach produces a subset of highly generalisable features but is unsuitable for any specific learning algorithm because it is not biased towards any learning algorithm. The work flow for the filter approach is shown in Fig. 5 below.

The relief method and MI are two common filtering techniques. The MI of two random variables is a method that can determine the degree to which the variables depend on one another. The MI shows the amount of information that a feature and a class label exchange and how one random variable affects another [67]. Since it shares the most information with the target class, its strongest attribute is its ability to distinguish between individuals from one class and those from another. The MI equals zero when the variables being explained are statistically independent. Examples of MI are as follows: (1) information gain, which determines the value of a feature by computing the information gain with respect to the class [68]; (2) symmetrical uncertainty, which determines the value of a feature by computing the symmetrical uncertainty with respect to the class and compensates for the bias of information gain; and (3) the chi-square statistic, which determines the value of a feature by computing the value of the chi-square statistic with respect to the class [69]. This method for weighing features is sensitive to the interplay of features. The Relief method evaluates a feature's quality based on how well its values can be distinguished across related instances [61]. The distinction between them can be seen in the scores given to each attribute. The Relief technique assigns a score to each feature, which can then be used to rank them. Relief's major goal is to rate features according to how well they cluster instances of the same class and how well their values distinguish between instances of different classes [68].

To achieve this, Relief selects a random instance from the acquired data and then searches for the closest instances of the same class and of other classes. The feature values from these instances are used to update the related scores for each feature. Because the weighting score of each feature is changed for each chosen case or instance, the relief method can be very useful for high-dimensional datasets. If it comes from a feature in an adjacent instance of the same class, the weight of a feature will decrease; if it comes from a feature in a different class, the weight will increase [59]. Relief searches for the two closest neighbours of a randomly selected instance, given by the notation $x_i = \{x_{1i}, x_{2i}, \ldots x_{ni}\}$, seeking one neighbour from the same class, marked as "nearest hit H", and one neighbour from a different class, denoted as ''nearest miss M''. Then, the values for x_i, M, and H are included in a revised version of the quality estimate for every single feature. This method is limited to two-class problems and does not account for redundancy. While it can manage discrete and continuous features, it cannot manage redundancy. On the other hand, the relief method is commendable since it may be used in all contexts, has little bias, takes feature interaction into account, and might pick up local dependencies that other methods miss. Dependency measures are another type of FS measure. Correlation feature selection, the chi-square test, linear regression, LR, and the Fisher exact test are just a few dependency-based techniques available.

4.1.1.1. Correlation feature selection (CFS). The correlation-based technique simplifies determining the significance of a feature by correlating it with the target or outcome variables. The Pearson method is used to assess the significance of a correlation between two continuous variables [58]. Analysis of feature importance correlation is generally applicable so long as it is possible to build predictive models and is not limited by specific representations, algorithms, or metrics [70]. Most of the time, the Pearson and Spearman correlation coefficient distributions are shown so that feature importance values and ranks can be compared. Pearson's correlation coefficient is used to choose linear regression features, while Spearman's rank coefficient is used to choose non-linear regression features. The Pearson correlation coefficient is applicable when variables are continuous and linearly correlated. It is also possible to use the Spearman correlation coefficient with ordinal categorical variables. A statistic known as the correlation coefficient is used to describe the strength and type of a relationship between two variables. A relationship's strength could fall anywhere between -1 and $+1$. When a perfect linear correlation exists between two variables, their correlation coefficient equals 1. In other instances, the correlation coefficient approaches one as the correlation strength increases.

A correlation value of 0 means that the two continuous variables do not have a linear relationship or that there is no correlation between the variables. Also, there is no correlation if the variables do not have a linear relationship with each other. If the coefficient is positive, then the variables are directly correlated; as the value of one variable increases, so does the value of the other variable. If the coefficient is negative, the variables have an inverse relationship, which means that as one variable's value increases, the value of the other variable tends to decrease. In the context of FS, correlation feature selection (CFS) is a basic filtering method that uses a correlation-based heuristic evaluation function to rank subsets of features. According to Ref. [68]; the CFS analyses a subset of data by considering the degree of duplication between each feature and its predictive value. The evaluation function favours subsets with features strongly correlated with the target class but uncorrelated among themselves. Certain variables were excluded from an earlier study by Ref. [12] that used ML to analyse MA because of their poor correlation with the target variable. In that study, the *t*-test of the correlation coefficient was used to determine whether or not the relationship between the features was statistically significant.

In addition, some techniques use multivariate, univariate, and bivariate FS. The univariate factor analysis (FS), based on univariate statistical tests, examines each feature independently to establish the extent to which it relates to the response or outcome variable. A regression analysis involving several features as predictors could be used to identify features based on univariate statistical tests and then examine each feature independently to determine the degree to which it relates to the outcome variable. Bivariate and multivariate analyses, which involve two or more variables, are distinguished from univariate studies by the number of variables they each involve [71]. The Fisher's exact test, chi-square tests, LR, and Bonferroni-adjusted chi-square tests are some of the statistical methods that could be utilised in the context of FS and in establishing the significance of features.

Linear regression is one of the most fundamental and widely used ML methods based on supervised learning. It is also capable of predictive analysis in addition to regression. Regression models predict a target

value based on independent factors. Linear regression is a statistical technique that can be used to characterise the relationships that exist between at least one predictor variable and at least one outcome variable (scalar response). When there is only one explanatory variable, simple linear regression is employed; multiple linear regression is utilised when there are numerous explanatory variables. A positive coefficient indicates that the mean of the dependent variable increases as the independent variable's value increases. A positive coefficient for the dependent variable indicates that it will increase in response to an increase in the independent variable.

4.1.1.2. Logistic regression. The family of supervised ML models includes the ML model known as logistic regression (LR). When the dependent variable is dichotomous, the LR approach is used for statistical analysis. Classification and other forms of predictive analytics frequently make use of LR. The LR method defines and explains the relationship between one dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables [72]. The probabilistic nature of the event is reflected in the range of possible values for the dependent variable, which goes from 0 to 1. In binary classification, a probability less than 0.5 will predict the value 0, whereas a probability larger than 0 will predict the value 1. A logit formula is used in LR to transform odds, defined as the ratio of the likelihood of success to the probability of failure. The odds of an outcome occurring in the presence of a particular event are contrasted with the odds of that outcome occurring in the absence of that event using the odds ratio [OR] [72]. If the OR exceeds 1, there is a greater likelihood that the event will result in the desired outcome. An event is linked to a lower likelihood of that outcome occurring if the OR is less than 1.

4.1.1.3. Chi-square tests. Researchers and analysts commonly use the chi-square test of independence (also known as Pearson's chi-square test of association) to determine the degree of relationship between two categorical variables. The chi-square statistic is commonly used to describe this type of association test. The chi-square test detects whether or not two variables are independent and requires that the variables be measured on an ordinal or nominal scale. Categorical data is a specific kind of data. There should be two or more distinct categories for each variable. (Laerd Statistics, 2018). One variable that satisfies this criterion is the use of medications, which may be divided into two categories: adherence and non-adherence.

4.1.1.4. Fisher's exact test. The Fisher exact test is a statistical process that is utilised when two nominal variables are present. Fisher's Exact Test is used to gauge if there is a statistically significant difference between the proportions of the categories in two group variables. Instead of using a mathematical formula to estimate the likelihood of a test statistic result, Fisher's exact study checks the likelihood of getting both the observed data and all data sets with larger extreme deviations under the null hypothesis that proportions are the same. On the other hand, normal statistical tests use the null hypothesis, which asserts that the proportions are the same [73]. When the number of people expected to participate in small, Fisher's exact test is more accurate than the chi-square or G-test of independence. When carrying out the Fisher's exact test, the unadjusted OR and a 95% confidence interval are used instead of the p-value.

4.1.1.5. Linear regression. One of the most basic and widely used ML methods based on supervised learning is linear regression. It is also capable of predictive analysis in addition to handling regression tasks. Regression models predict a target value based on independent factors. Linear regression is a statistical technique that can be used to characterise the relationships that exist between at least one predictor variable and at least one outcome variable (scalar response). When there is only one explanatory or independent variable, simple linear regression is employed; multiple linear regression is employed when there are multiple explanatory variables. A positive coefficient indicates that the mean of the dependent variable increases as the independent variable's value increases. A positive coefficient for the dependent variable indicates that it will increase in response to an increase in the independent variable.

4.1.1.6. Logistic regression. The family of supervised ML models includes the ML algorithm known as logistic regression (LR). When the dependent variable is dichotomous, the LR method is utilised for statistical analysis. Classification and other forms of predictive analytics frequently make use of LR. The LR method defines and explains the relationship between one dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables [72]. The probabilistic nature of the approach is reflected in the range of possible values for the dependent variable, which goes from 0 to 1. In binary classification, a probability less than 0.5 will predict the value 0, whereas a probability larger than 0.5 will predict the value 1. A logit formula is used in LR to transform odds, defined as the ratio of the likelihood of success to the probability of failure. The odds of an outcome occurring in the presence of a particular event are contrasted with the odds of that outcome occurring in the absence of that event using the odds ratio [OR] [72]. If the OR exceeds 1, there is a greater likelihood that the event will result in the desired outcome. An event is linked to a lower likelihood of that outcome occurring if the OR is less than 1.

4.1.1.7. Chi-square tests. Researchers and analysts commonly use the chi-square test of independence (also known as Pearson's chi-square test of association) to determine the degree of relationship between two categorical variables. The chi-square statistic is commonly used to describe this type of association test. The chi-square test detects whether or not two variables are independent and requires that the variables be measured on an ordinal or nominal scale. Categorical data is a specific data type; each variable should have two or more distinct categories (Laerd Statistics, 2018). One variable that satisfies this criterion is the use of medications by NCD patients, which may be divided into two target classes or categories: adherence and non-adherence.

4.1.1.8. Fisher's exact test. The Fisher exact test is a statistical procedure that is used when there are two nominal variables. Fisher's exact test is used to gauge if there is a statistically significant difference between the proportions of the categories in two group variables. Instead of using a mathematical formula to estimate the likelihood of a test statistic result, Fisher's exact study checks the likelihood of getting both the observed data and all data sets with larger extreme deviations under the null hypothesis that proportions are the same. On the other hand, normal statistical tests use the null hypothesis, which asserts that the proportions are the same [73]. When the number of people expected to participate is small, Fisher's exact test is more accurate than the chi-square or G-test of independence. When carrying out Fisher's exact test, the unadjusted OR and a 95% confidence interval are used instead of the p-value.

4.1.2. Wrapper method

Wrapper FS methods choose a group of features to use as a search issue by preparing, evaluating, and comparing one group of features to another group of features. This method makes it simpler to identify probable feature-feature interactions. Wrapper methods focus on feature subsets that will enhance the accuracy of the selection process results. In the wrapper technique, the FS algorithm functions as a wrapper over the predictive model algorithm, choosing the best features based on the same model. Unlike the filter method, the wrapper method builds models from scratch for each created subset. The algorithm's efficacy is

evaluated by how well it makes predictions. Instead of examining each feature separately, as filters do, wrappers analyse features in terms of how they relate. The analytics scientist uses a selection of features to train a model using wrapper techniques, and based on the conclusions drawn from the preceding model, it is decided whether to include or exclude certain features from the subset. The wrapper model assesses the goodness of the chosen subsets using the expected precision of a predetermined learning algorithm [59]. The wrapper techniques choose features as a consequence by repeatedly picking a subset of features and assessing the importance of those features based on their performance. For example, the accuracy of a model that was produced using a particular learning algorithm is one performance metric that can be used. The cycle is continued until either a specified stop condition is met or a subset of features that optimise the model's performance is found. Once one of these conditions is met, the cycle ends. ML algorithms make the comparison, which in most cases, makes the learning process computationally complex and expensive in terms of the time it takes to carry it out [58]. Wrapper strategies pick limited features to maximise the learning process's performance. Because of this, the possibility for these techniques to be generalised is restricted. Fig. 6 is an illustration of the workflow for the wrapper technique.

Even though the wrapper method could result in superior output, it still demands more processing resources [61]. For forward FS, backward feature removal, recursive feature elimination (RFE), and permutation feature importance, wrapper algorithms are frequently employed.

Forward selection is an iterative procedure that starts with no model features. This strategy, known as a "greedy search approach," begins with an empty set and incrementally adds features. The feature that improves the model the most is added at each iteration. This continues until adding a new feature does not affect how well the model performs [68]. With few representative features (the ideal subset), forward FS performs best [74]. This method's drawback is that once a feature has been included in the subset, it cannot be removed, which leads to subpar outcomes.

4.1.2.1. Backward elimination. To improve the performance of the model, backward elimination begins with all of the features. Then it eliminates the feature that is deemed to be the least significant or the feature with the highest error at each iteration, improving the model's performance [68]. The technique is repeated until there is no improvement in feature deletion.

Recursive feature elimination is a greedy optimisation approach to identify the feature subset with the highest performance. It creates models frequently and saves the most or least significant feature at each iteration. It builds the subsequent model using the remaining features, as long as they are still available. The features are then rated in reverse order of elimination. The features that received the highest ratings are maintained, while those that received the lowest are removed. Recursive FS is an FS that begins with all features and then does so without going backwards. This procedure is carried out several times until either the required number of features is acquired or it is determined that the performance cannot be improved further.

Permutation Feature Importance "Feature permutation importance" is a model-agnostic global explanation technique for understanding how an ML model behaves (Oracle, 2021). It performs an analysis and assigns a ranking to the significance of each feature based on its influence on the predictions generated by the trained ML model. The term "permutation of feature importance" refers to the decrease in a model's score generated by randomly selected feature values [75]. The model generates this reduction. The primary objective for determining which features are relevant for prediction is to observe how the ML model's predictions change as the value of a particular variable is altered. The decrease in model score indicates the model's reliance on the feature, as this procedure eliminates the link between the feature and the target. This method has the benefit of being model-independent and can be calculated multiple times with various feature permutations. The training set and a held-out testing or validation set can be used to determine the importance of permutation. Using a held-out set, one can determine which features are most responsible for the examined model's generalisation power. The likelihood of the model being overfitted increases when the training set contains significant features absent from the holdout set.

4.1.3. Embedded method

Learning algorithms that have FS incorporated are referred to as "embedded." Embedded FS approaches enable the simultaneous operation of FS and classification by incorporating the FS ML algorithm as a learning algorithm component. In other words, the embedded technique incorporates the FS process into the learning or model-building phases. It builds an optimized classifier by choosing features through the learning process. The features that will contribute the most are carefully extracted for each model iteration. The embedded technique, a hybrid of a filter and a wrapper, makes its feature selections dynamic during the learning process, basing those selections on how the classifier evaluates each feature. When compared to wrappers, this process results in lower overall processing costs. Additionally, it is less likely to overfit. This technique classifies dependencies more easily than wrappers since FS happens during the classification phase [59]. As a result, they are ideal and time-efficient because embedded techniques select features using the target learning algorithm rather than iteratively. Fig. 7 diagrammatically demonstrates the workflow of the embedded FS method.

Algorithms carry out embedded FS with their internal FS methods. The most popular embedded techniques include RIDGE regression, LASSO FS, decision tree FS, and RF FS. When building the model using criteria like entropy or gini, decision pruning and other tree-based learning techniques remove extraneous features [76]. Both LASSO and RIDGE regression have built-in penalization mechanisms to limit the possibility of overfitting [47]. utilised the Gini index throughout their research to determine which feature of each tree node reflected the optimal split. Each branch of the RF was validated using data collected from samples that were not selected for use in its development.

In **Lasso regression**, L1 regularisation is applied, resulting in a penalty proportional to the absolute magnitude of the coefficients. During the FS phase, the features with coefficients that were not precisely zero after the shrinkage stage were chosen to be incorporated into the model. The features with coefficients that were exactly zero were rejected. The interpretability of LASSO can be considerably enhanced by removing unnecessary features. It would prefer to pick one feature from each group of related features and ignore the others [77]. used the LASSO method to perform variable selection and regularisation in a study that analysed and compared three predictive analytic approaches to characterise medication non-adherence. This improved the final model's predictability and interpretability.

Ridge regression employs L2 regularisation, which imposes a

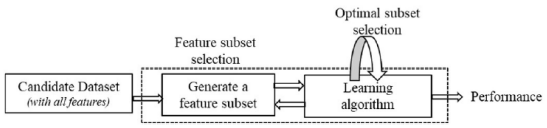


Fig. 6. Wrapper method workflow.

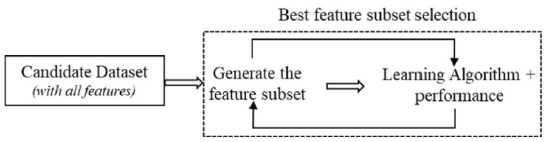


Fig. 7. An embedded filter method.

penalty equal to the square of the coefficient magnitude. In contrast to LASSO, which offers a limited number of functions, Ridge compresses model coefficients to values almost identical to zero but do not exactly equal it. Because LASSO reduces irrelevant coefficients to zero and cancels them out, it differs most significantly from the Ridge technique. Contrarily, Ridge produces non-zero coefficients that are better suited for feature interpretation.

Wrapper approaches evaluate the usefulness of selecting features by instructing a model to make inferences based on those selections. In contrast, filter techniques evaluate the relevance of features based on their connection with dependent variables. Since filter techniques do not need model training, they are much faster than wrapper methods. Wrapper techniques, on the other hand, need significant computational resources. Additionally, filter approaches use cross-validation, whereas wrapper approaches use statistical methods to evaluate a subset of features. Even though the filter method performs better and is more practical with more features, the wrapper method outperforms the other two, and the embedding method performs somewhere in the middle.

Embedded FS is performed by algorithms using their internal FS techniques. The most popular embedded techniques include RIDGE regression, LASSO FS, decision tree FS, and RF FS. When building the model using criteria like entropy or Gini, decision pruning and other tree-based learning techniques remove the extraneous features [76]. Both LASSO and RIDGE regression have built-in penalization mechanisms to limit the possibility of overfitting [47]. used the Gini index in their research to determine which feature of each tree node reflected the optimal split. Each branch of the RF was validated using data collected from samples that were not selected for use in its development.

In **Lasso regression**, L1 regularisation is applied, resulting in a penalty proportional to the absolute magnitude of the coefficients. During the FS phase, the features with coefficients that are not exactly zero after the shrinkage stage are chosen to be incorporated into the model. The features with coefficients that are exactly zero are rejected. The interpretability of LASSO can be considerably enhanced by removing unnecessary features. It prefers to choose one feature from each group of related features and disregard the others [77]. used the LASSO method to perform variable selection and regularisation in a study that analysed and compared three predictive analytic approaches to characterise medication non-adherence. This improved the final model's predictability and interpretability.

Ridge regression uses L2 regularisation, which imposes a penalty proportional to the square of the magnitude of the coefficient. In contrast to LASSO, which offers a limited number of functions, Ridge compresses model coefficients to values that are close to but not exactly equal to zero. The most significant difference between LASSO and the Ridge technique is that LASSO reduces irrelevant coefficients to zero and cancels them out. In contrast, Ridge yields coefficients that are non-zero and better adapted for feature interpretation.

Wrapper approaches assess the usefulness of selecting features by instructing a model to draw conclusions based on the selections. In contrast, filter techniques assess the significance of features in relation to dependent variables. Since filter techniques do not need model training, they are much faster than wrapper methods. Wrapper techniques, on the other hand, need significant computational resources. Additionally, filter approaches use cross-validation, whereas wrapper approaches use statistical methods to evaluate a subset of features. Even though the filter method performs better and is more practical with more features, the wrapper method outperforms the other two, and the embedding method performs somewhere in the middle.

5. Recent work on ML application to NCD MA incorporating FS and the importance of predictors

This section (Table 1) summarises recent articles on FS and the importance of predictors of MA among NCD patients (CVD, hypertension, diabetes, OUD, and others) in ML research.

6. Results

Two key thematic areas emerged from the outcomes of this systematic study. The primary thematic findings of the study are the FS methods used to assess the contribution of predictors to NCD MA in studies that employed ML approaches, as well as the potential features or predictors for NCD MA prediction.

6.1. Potential features that contribute to the prediction of NCD MA based on ML analytics

This study examined six research articles using ML to predict MA in hypertensive patients [22,35,48,78–80]. In China [27], used LR, DT, neural networks, and RF to classify MA in patients with hypertension. In that study, risk factors included age, location (urban or rural), educational level, marital status, occupation, lack of physical exercise, smoking and alcohol consumption, a balanced diet, dyslipidaemia, diabetes, and atrial fibrillation. All the developed ML classification models performed well. The DT model produced the highest precision and recall (0.941).

Gao et al. [79] used DT to analyse data from the medical records of hypertension patients treated at a Beijing community hospital. They identified the factors that make it challenging for hypertensive patients to take their medication as prescribed. The researchers discovered that people with hypertension older than 66.5 had poor MA. Concurrent diabetes, the types and frequency of medications, multiple medications, and more than 0.96 annual follow-up exams were linked with better adherence to antihypertensive therapy. Patients who visited the hospital more frequently were more compliant and had greater blood pressure control.

Aziz et al. [22] used artificial neural networks (ANN) and RF to identify factors influencing hypertension patients' adherence to treatment programmes. Support Vector Regression (SVR) and Self-Organizing Feature Maps (SOM) were also used to determine whether they could be useful in determining the adherence levels of hypertension patients in a tertiary hospital in Kuala Lumpur, Malaysia. The significance of the following features was established with the help of the RF feature importance mechanism: education level, marital status, general overuse, monthly income, and specific concern. The results show that the most accurate predictor of MA was a specific concern, which had a percentage increase in mean squared error (MSE) of 12.67% [89]. describe a specific concern scale as a measurement tool that evaluates the perceived likelihood of negative medication behaviour. Regarding the RF feature importance measure, "counselling regarding antihypertensive medication" showed the least amount of change in MSE.

Chu et al. [35] studied how MA risk factors interact and used classification and regression tree (CART) analysis to predict poor MA in older people with hypertension. The CART model reveals that many contributing features are responsible for low MA, including beliefs about health, the length of sickness, self-efficacy, and the amount of social support. A misclassification rate of only 26% was relatively low, indicating that the model accurately predicted MA [78]. used RF, LR, GB, and KNN to predict MA in patients with hypertension, focusing on healthcare usage features. The number of medication supplies on hand at the start of the prescription impacted the model. Higher values meant a higher chance of a good (adherent) outcome. Another significant factor was the past MA, where a higher-level past MA is associated with positive outcomes and vice versa.

Li et al. [80] employed CHAID to identify demographic profiles associated with nonadherence to antihypertensive medication using a population segmentation method based on ML. Using the 2016 New York City Community Health Survey data, the study identified and divided individuals into subgroups based on their non-adherence to antihypertensive medications. Age, diabetes, family income, health insurance coverage, race or ethnicity, and neighbourhood poverty have all

Table 1
Summary of the recent articles on FS and the importance of predictors of NCD MA in ML research.

Author	AIM	Data Collection	Potential Features/Variables Considered	ML Approach	Findings
[27]	Create ML models to classify MA in hypertensive patients.	As part of China's national stroke screening and intervention programme, people over the age of 40 in 31 provinces, autonomous regions, municipalities, and the Xinjiang Production and Construction Corps were screened from 2015 to 2017. Stratified cluster sampling in two stages was used during the screening process. At first, more than 200 screening regions were chosen. These regions were chosen based on the number of counties in each region and the population of each county. Then, one metropolitan city and one township were selected as the major sampling units based on geography and recommendations from a nearby hospital. Each major sampling unit's whole population above the age of 40 was surveyed using cluster sampling.	A person's age, location, whether they live in a city or a rural area, whether or not they have social health insurance, their level of education, marital status, job, level of physical inactivity, whether or not they smoke or drink, whether or not they eat a balanced diet, whether or not they take their blood pressure at home, dyslipidaemia, diabetes, and atrial fibrillation (AF) were all risk factors. Other risk factors include high blood pressure and high cholesterol.	LR DT Neural Net RF	The data from the screening showed that 88.9% of the 55,953 people diagnosed with hypertension took their medication as prescribed. Since 2013, people at a high risk of having a stroke have been watched to ensure they took their medicine as prescribed and checked their blood pressure regularly. As a result, more people than the general population took their medications on time. The experiment's classification models all performed well. The DT model was shown to have the highest precision and recall. The recall for the positive group in the DT model was 0.941 (those with low MA).
[35]	Analyse the interactions between MA risk factors and create a model to predict non-adherence in hypertensive older people.	For a cross-sectional study, 300 older people with hypertension were enrolled using a convenience sample. Between March 2019 and February 2020, people who were eligible for the study were picked up at outpatient cardiology and family medicine clinics at two different hospitals. Participants in this study were required to have high blood pressure and be at least 65 years old. Additionally, they needed to have been taking their blood pressure medication for at least a year. The participants were given a questionnaire to fill out during the in-person interviews that were conducted.	Features of the population and the disease, beliefs towards health, levels of self-efficacy, and social support.	CART	According to the CART model, a lack of MA can be attributed to a person's beliefs about health, the severity of their disease, their level of self-efficacy, and the support they receive from their social network. The model could predict the outcome accurately with a misclassification rate of only 26%, which was pretty low.
[78]	To examine the predictability of medication compliance in a sample of hypertension patients using a variety of predictive scenarios, with an emphasis on the factors that influence healthcare utilisation.	Patients with essential hypertension (ICD10 code i10) who were treated in Sweden's public primary, secondary, and specialised care institutions between the ages of 18 and 90 years old had their previous healthcare utilisation and demographic factors analysed.	Baseline predictors including demographic information of the patients (age/- gender) at the time of prescription, prescriber age, aggregate predictors (polypharmacy, ATC-code diversity, years under treatment, number of previous prescriptions) of current and past prescriptions, as well as visit frequency and visit types, such as outpatient, inpatient and emergency care visits.	RF LR GB KNN	The test set's AUCs range from 0.77 to 0.89, reflecting that model performance varies across data splits. In most cases, using historical data results in a slight improvement (an increase in absolute AUC of 1%–2%). Models performed poorly in patients with sudden PDC declines and index prescriptions (AUC test set: 0.56–0.66, recall: 0.58–0.63). Simulations show that the only explanation for observed PDC patterns was specific medication consumption behaviours. The model output was most strongly influenced by the quantity of supplies that were on hand at the beginning of the prescription; higher values of supplies on hand increased the likelihood of a successful (adherent) outcome. The

(*continued on next page*)

Table 1 (*continued*)

Author	AIM	Data Collection	Potential Features/Variables Considered	ML Approach	Findings
					participant's history with MA was also a significant factor. Higher values led to an increase in the positive model output, while lower values led to an increase in the negative model output. This was demonstrated by the large overlap that existed between the positive and negative labels for low previous adherence.
[22]	Determine, with the help of Ml, which factors influence the adherence of patients with hypertension to the treatment plans they have been prescribed.	In the year 2011, data were collected from a total of 160 adult patients at the outpatient clinics of the University Kebangsaan Hospital. All of these patients had been diagnosed with essential hypertension and were taking at least one antihypertensive medication for a period of more than a year at the time of the study.	A person's age, gender, level of education, concomitant diseases, the presence of others in their home, marital status, general overuse, monthly income, religion, ethnicity, specific concern, occupational field, general harm, length of time of antihypertensive medication intake, counselling for antihypertensive medications, the overall amount of antihypertensive medications taken per day, and aids in antihypertensive and medication intake are all factors that were taken into consideration.	RF ANN SVR	The RMSE values for the ML models generated using the selected variables were 1.42 for the ANN model, 1.53 for the RF model, and 1.55 for the SVR model. The RF feature importance mechanism was used to analyse the importance of the following features: level of education, marital status, general overuse, monthly income, and specific concern.
[79]	To spell out the factors that make it difficult for people with high blood pressure to comply with prescribed medication.	DT was used in a "hypertensive cohort" of individuals with high blood pressure from a community hospital in Beijing, China, to determine what prompted individuals to take their hypertension medications as directed. This "hypertensive cohort" was analysed to determine what prompted individuals to take their hypertension medications as directed. The Intelligent Chronic Disease Management System was used to collect data on 7638 hypertensive patients who were treated at the Fangzhuang Community Health Service Centre between June 1, 2014, and December 31, 2018. These patients were seen between the periods of June 1, 2014, and December 31, 2018.	Types of antihypertensive medications, body mass index, history of smoking, the overall number of hospital visits before the initial prescription, age, the total number of medication-using days in the year before enrolment, total outpatient follow-ups, and the presence of diabetes are all factors that were taken into consideration.	DT	Body mass index, smoking, the total number of hospital visits in the year prior to enrolment, the total number of days spent taking medication in the year before enrolment, age, diabetes, follow-up rates, and the number of times antihypertensive drugs were not taken as prescribed in the year prior to the first prescription are all factors that influenced compliance. In addition, the number of times antihypertensive drugs were not taken as prescribed in the year prior to the first prescription is also a factor. The antihypertensive MA prediction model had values of ROC-AUC equal to 0.81, sensitivity values equal to 0.78, and specificity values equal to 0.69.
[80]	Determine the characteristics of those who disregard medical advice and refuse to take antihypertensive medications.	Researchers used ML-based population segmentation to determine which demographic characteristics were associated with individuals who did not take their blood pressure medication as directed. The individuals were divided into different categories based on how well they could control their blood pressure with medication. They used information from the 2016 New York City Community Health Survey.	A person's age, the poverty level in their neighbourhood, the prevalence of diabetes, the income of their home, whether or not they have health insurance or their race.	CHAID	10% of adults in New York City did not take antihypertensive medication despite doctors' recommendations. Age, diabetes, family income, health insurance coverage, and race/ethnicity were shown to be important variables that can be used to predict nonadherence behaviours and split people with hypertension into subgroups.
[81]	To predict MA in persons at atherosclerotic CVD risk, use ML.	Forty persons at risk for atherosclerotic CVD were included in an eight-week feasibility trial. Participants were found through	Sociodemographic information such as age, sex, gender, ethnicity, the number of medications taken daily, having somebody to talk to,	SVR RF Regression GTB Regression AdaBoost Regression	When only the 10 factors with the greatest significant impact were taken into account, the model had an average error rate of 12.9%.

(*continued on next page*)

Table 1 (*continued*)

Author	AIM	Data Collection	Potential Features/Variables Considered	ML Approach	Findings
		recommendations from primary care doctors at one Southern California outpatient tertiary care facility. Participants needed to have at least one risk factor for atherosclerotic cardiovascular disease, such as being overweight or obese, suffering from hypertension, diabetes, metabolic syndrome, or having high cholesterol levels. They also had to get treatment for their illness, live independently, and agree to always carry the research phone with them. Every person was given a questionnaire containing several items to respond to. After the first data set was collected as a baseline, more data was collected using a pillbox that sent information to a server in the cloud.	knowing what to do in an emergency, participants' ability to take an active role in their treatment, and various clinical and psychosocial parameters such as knowledge of prescribed medication use, the number of comorbidities, and the presence of a special person in the patient's life who cares about their feelings was all taken into consideration.	DT Regression Linear Regression	There was a significant correlation between MA and the following factors: having a significant other in their lives, being told to take an active role in their daily medication, feeling confident about what to do in an emergency, and understanding how to take their medications. The results showed how important clinical and psychological factors are when predicting MA in people with a high chance of getting atherosclerotic CVD. The performance of linear regression was inferior to that of other methods of regression, specifically SVR and GTB regression, with SVR surpassing the other approaches by 16.6%.
[82]	Build a model of statin adherence using EMRs, then link that model to the biology and clinical outcomes of individuals who take statins.	They obtained the EMR data from the Military Health System, which maintains administrative information for active-duty, retired, and dependent US military personnel eligible for medical benefits for themselves or their families. The study asked questions of people who have prescribed a statin for the first time in 2005 or 2006. As a starting point, this study used non-negative matrix factorization to gather billing, lab, and pharmacy claims data from the two years before the first statin prescription.	Gender, age, follow-ups in days, days, supply, smokers, race, number of labs, number of concomitant drugs, strength, having diabetes or established CVD, atrial fibrillation, congestive heart failure (HF), women's health maintenance, sinus disease and allergies, physical medicine and rehabilitation, treatment for insomnia, pain, eye disease, pain treatment; antibiotics, asthma/chronic obstructive pulmonary disease, musculoskeletal pain.	RF	Patients who adhered to their statin therapy were more likely to be younger, female, smokers, have longer follow-up periods, and interacted with the healthcare system more frequently than patients who did not. Additionally, those who continued their statin therapy had higher starting doses and more supply days of the medication. Patients with diabetes or established CVD had greater rates of statin adherence. The well-known negative effects of statins on musculoskeletal symptoms may lead to disorders like chronic pain associated with poor treatment adherence. A c-statistic of 0.736 was established by using the patient's statin prescription, estimated disease risk, and EMR data to predict that the patient will not take their statins as prescribed. Following the administration of the initial refill, the c-statistic rose to a value of 0.81.
[8]	The study analysed patient determinants in medication persistence to predict lipid-lowering therapy persistence for one to two years after discharge from the initial hospitalisation for CVD.	In order to carry out a proportional hazard model analysis, a sub-cohort consisting of 313,207 patients who had their initial hospitalisation due to CVD were taken from a web-based CVD risk assessment and management decision support system for primary care in New Zealand. Patients in the sub-cohort underwent their first risk assessment after January 1, 2011, and no later than December 31, 2014, they should have experienced their first ischemic CVD episode. In total, 2480 patients were enrolled in the trial.	Age, gender, history of diabetes, level of complications and comorbidities, ethnicity, level of deprivation, body weight, number of hospital days, kind of CVD diagnosis, in-patient procedures, newness to therapy, refill history. The number of days between the most recent pre-hospitalisation dispensing date and the hospital admission date is all considered.	LR SVM ANN BRT	According to proportional hazard modelling, there was a high association between persistence and age, a history of diabetes, the severity of complications and comorbidities, obesity, the number of hospital days remaining, the type of CVD diagnosis, in-patient procedures, and the age of the therapy. Additionally, there was a high association between persistence and the number of hospital days remaining. When easily accessible criteria were considered, the BRT proved to be the most accurate predictor, with a c-statistic of 0.811 (0.799–0.824) for a 1-

(*continued on next page*)

Table 1 (*continued*)

Author	AIM	Data Collection	Potential Features/Variables Considered	ML Approach	Findings
		Drug persistence after one and two years was predicted using ML models.			year prognosis and 0.793 (0.772–0.824) for a 2-year prediction. People between the ages of 30 and 54 had a significantly higher risk of not taking their medications, whereas women had a slightly higher risk of not taking their medications. When compared to Europeans, Indian patients demonstrated a significantly reduced non-persistence rate. Only patients from the poorest areas had a higher likelihood of not adhering to treatment than patients from the wealthiest areas.
[77]	Compare predictive analytic tools for characterising medication non-adherence and determining when each method should be used.	An agreement based on payment of a set cost for each service provided in a 5% Medicare sample was used to identify 11,969 Medicare recipients who had been hospitalised for an acute myocardial infarction and had subsequently filled a statin prescription at the time of discharge or shortly afterwards. These patients had been treated for their condition in an acute care setting. This took place between 2007 and 2012. The claims for prescription medications were broken down into three categories in the dataset: Part A, Part B, and Part D. Patients were only considered for inclusion in the study if they met all of the following criteria: they had to have been discharged from the hospital without receiving hospice care; they needed to have had Part D coverage for at least nine months before hospitalisation and for one year after discharge; they needed to have survived for at least one year after leaving the hospital; and they needed to have filled a prescription for statins within the first 30 days after leaving the hospital. These individuals had previously been given a prescription for statins and were expected to receive a new one within the first thirty days after their prior one had run out. The focus of the study was placed on the patients' initial admission, even if they had previous admissions that qualified them for the study.	Age, being female, prior statin use, race, number of CVD medications prescribed at discharge, geographic region, having a coronary artery bypass graft (CABG) procedure during their index admission, dually eligible for Medicaid, any hospitalisation in the prior six months, chronic kidney disease, diagnosed with diabetes, non-acute institutional care in previous 6 months, cerebrovascular disease, diagnosed with hypertension, diagnosed with dyslipidaemia, congestive HF, length of stay during the index admission, rural residence, prior MI, prior CABG.	LR RF LASSO	There was not much difference between the three data analysis methods, despite the fact that the LASSO regression model selected more than 90% of all the viable variables (the C-index ranged from 0.664 to 0.673). Previous statin use was the most significant predictor of lipid-lowering MA across all models (OR = 3.65, 95% CI: 3.34–3.98; OR = 3.55).
[47]	Determine whether patients take their medications as prescribed and adhere to their dietary and exercise regimens.	Data from 90 cases that were obtained retrospectively were used to evaluate the classifiers. The University Hospital of Ioannina's Department of Cardiology provided the information. Patients with HF who were at least 18 years old, receiving the	General information, including the patient's age, gender, and caretaker; allergies and medical conditions, including the severity of the patient's post-myocardial infarction and HF (KILLIP classification) and the extent of the patient's HF	RF Simple CART RT LMTs RBF Network, RF SVM BayesNet	The first classification problem (global adherent) had the best detection accuracy at 82%, while the second classification problem (medication adherent) had the best detection accuracy at 91%. It was shown by combining several classifiers with different

(*continued on next page*)

Table 1 (*continued*)

Author	AIM	Data Collection	Potential Features/Variables Considered	ML Approach	Findings
		best care possible, and who had recently been hospitalised, admitted as an emergency, or undergone specialised consultation for decompensated HF were included in the study. Additionally, within the previous year, patients had to have at least one electrocardiogram (ECG) indicative of HF.	(NYHA class); the patient's smoking habit; the patient's drinking habit; and the presence or absence of comorbidities; drugs, including the active ingredients, dose, and frequency of use; and any other relevant information. Biological information connected to heart failures, such as a left bundle branch block or intraventricular delay.	MLP Naive Bayes	FS metrics that factors representing the patient's health and medication therapy might function as predictors of adherence. The results imply that understanding medicine adherence affects predictions of global adherence.
[83]	Utilising SVM, determine whether patients with HF have features that are predictive of MA.	A questionnaire regarding their MA was completed by seventy-six patients who went to an outpatient clinic at a hospital in Bucheon, South Korea, between January and April of 2010. An SVM model was developed to determine the factors that most accurately predict MA.	Ejection fraction, gender, amount of medications taken daily, medication knowledge, HF functional class, age, education level, monthly income, number of years married, and gender of the spouse. The Mini-Mental Status Examination is available in Korean under the abbreviation MMSE-K.	SVM	The best two classification models for MA in HF patients used predictors that included two subsets of features drawn from the following: gender, spouse, age, education, monthly income, number of medications taken daily, knowledge about medications, NYHA functional class, ejection fraction, and MMSE-K. The most accurate detection rate was 77.63%.
[2]	Find out how well the patient's drop history from the drug event monitoring system can be used to predict how well the patient will be able to take their prescription.	Over the course of three years, a total of 165,223 records pertaining to injection disposal were compiled from the 5915 different HealthBeacon units. The dataset that was used was far larger than any other dataset that had been described in the relevant literature before. It included data from more than 8000 Smart Sharps Bin (SSB) equipment and 342,174 injection records from years ago. The proposed model was based on injection (drop) data from a patient's medical history that was collected by an SSB connected to the internet. An Internet-connected SSB was utilised to monitor compliance with injection protocols.	Only historical information regarding a patient's injections or drops, such as the dates on which they were given (beginning with the earliest drop and ending with the most current drop). The most important categories of features were the historical dips, the weekday, the prescription count, and the continental location.	Extremely Randomised Trees RF XGBoost LSTM RNN Gradient Boosting MLP Ensemble learning	When the WEKA tools were applied to the whole training dataset, the factors that proved to be the most relevant were, in a nutshell, classified as follows: historical drops, medication frequency, nation, and scheduled medication event day. According to the findings of this study, the most important element was the patient's previous involvement with dropping out of treatment. Therefore, historical drop data was considered an important variable in predicting drug adherence. The importance of all the features was added up, and historic drops were found to be the dominant feature, accounting for 94% of the feature importance. This suggests that historical drop data contains the most useful predictive information, with "day of the week" and "medication" coming in second and third, respectively. The accuracy with which ML algorithms can identify patients who are likely to stop taking their prescriptions if they are given injectable medications at home was demonstrated by this study. With an AUC of 0.8902 and an accuracy of 82.90% on the unseen testing set, the best ML model, LSTM, can predict patients' adherence to self-injections at home with a chance greater than 50% based on a big dataset collected from SSBs. This is possible because LSTM has an accuracy of 82.90% on the unseen testing set.
[84]	Assess the degree to which ML algorithms can recognise the	At the Sichuan Provincial People's Hospital, a study was	Duration of unadjusted hypoglycaemic treatment,	ANN QUEST	The highest AUCs for the testing set for diabetic nephropathy,

(*continued on next page*)

Table 1 (*continued*)

Author	AIM	Data Collection	Potential Features/Variables Considered	ML Approach	Findings
	risks of complications and insufficient glycaemic control in people with type 2 diabetes but do not adhere to their treatment programmes.	carried out on the problems and the blood glucose prognosis of type 2 diabetes patients who did not comply with their treatment programmes from the beginning of 2010 through the end of the year. These T2D patients hadn't had their levels of glycosylated haemoglobin A tested or their treatment plans for hyperglycaemia changed in the prior 12 months.	gender, age, hereditary history, number of hypoglycaemic drugs and the duration of T2D.	BN Discriminate CHAID CRT Ensemble models	diabetic peripheral neuropathy, diabetic angiopathy, diabetic eye disease, and glycosylated haemoglobin A, respectively, were .902 ± 0.040, 0.859 ± 0.050, 0.889 ± 0.059, 0.832 ± 0.086, and 0.825 ± 0.092. Both univariate analysis and ML techniques produced the same result. The duration of T2D, the length of uncorrected hypoglycaemic treatment, the kind of insulin used, the number of hypoglycaemic medications used, and the overall cost of hypoglycaemic therapy were the most relevant risk variables for complications and poor glycaemic control. Risk variables for non-adherence to T2D glycaemic treatment included the quantity of hypoglycaemic medications taken and the duration of T2D.
[12]	Evaluate a variety of ML techniques, and then choose a model that can predict the likelihood that a patient will not adhere to the medication they have been prescribed.	The acquisition of the EMRs took place between April 2018 and March 2019. In addition, 401 patients with type 2 diabetes who were treated at the outpatient clinic of the Sichuan Provincial People's Hospital in China were asked to fill out a questionnaire in person. For data collection, individual questionnaires were used.	Demographic features (age, gender, occupation, family history), diabetes-related information (self-glycaemic monitoring, previous blood glucose level, diabetes course, medication regimen duration, medication regimen, HbA1c and blood glucose test results at the time). Information about a person's health, such as how they use Chinese traditional medicine if they've had surgery, if they have other health problems, and if they're taking other medicines simultaneously. Type 2 diabetes needed careful attention to several factors, including physical activity, diet (including eating patterns, whether or not the patient's previous fasting blood glucose was under control, and whether or not a realistic diet plan was followed), and mental health. Details concerning the conformance requirements (how many medications should be taken, how many were prescribed, and how many were left).	Bayesian Network Neural Net SVM LR K-NN LSVM RF C 5.0 model Tree-AS CHAID Quest C&R Tree Ensemble model	Sixteen different variables could be employed in the modelling process, and 30 different ML approaches were used to generate 300 models. The algorithm with an AUC of 0.866 ± 0.082 was the one that proved to be the strongest among these. When assessing whether a patient will take their medications, the most important factors to consider are how long the patient has been taking them, their previous blood sugar level, and their eating patterns or habits. The most important features of a patient's MA were the length of time they took their current medications, their previous blood glucose level, and their eating habits.
[85]	Use ML to establish adherence levels for the best hospitalisation risk discrimination and to examine the relationship between oral hypoglycaemic MA and hospitalisation avoidance.	Between the years 2007 and 2011, there were 33,130 people with type 2 diabetes evaluated. The patients were located using administrative claims data from the Pennsylvania Medicaid programme.	Claims and encounter data for Medicaid, including outpatient, inpatient, long-term care, and professional services; claims for prescription drugs, including information such as fill date, quantity dispensed, days of supply, and prescriber information.	RSV ST	The greatest discriminating adherence scores for every hospitalisation varied from 46% to 94%. The complexity of the patient's medications as well as their general state of health, were used to calculate these levels.
[23]	Establish important risk factors for early termination from treatment for OUD.	The researchers examined 2,381,902 cross-sectional treatment episodes for those 12 and older discharged from national treatment facilities for substance use disorders in the	Potential factors included treatment characteristics (the primary source of payment, service setting, and referral source, length of stay in treatment, and days waited to enter treatment), substance	RF, CART Bagged CART, and Boosted CART	The results obtained with RF (74% ROC-AUC) were significantly better than those obtained with alternative methods. Characteristics of the service setting, geography, the primary

(*continued on next page*)

Table 1 (*continued*)

Author	AIM	Data Collection	Potential Features/Variables Considered	ML Approach	Findings
		US between January 2015 and December 2019.	use history, socioeconomic status (e.g. geographic region, primary income source and employment status, education, living status), and demographic characteristics (age, race, ethnicity, sex and diagnosed psychological problems).		funding source, and the referral source were the most important risk variables. Race, ethnicity, and gender all had much less impact on prediction. When determining the treatment setting and length of stay, the employment status of the patient and the delay in admission (measured in days) were two of the most influential parameters. Their relevance rose as the length of time between treatments reduced. Notably, as therapy lengthened, the importance of the referral source grew.
[86]	To build an ML model for predicting each person's chance of discontinuing the OUD therapy programme.	The database of the Houston Emergency Opioid Engagement System contains information on 715 people who were addicted to opioids and had joined the programme more than ninety days before the index date. The data was collected through the use of self-reported questionnaires in addition to a health examination performed by a nurse practitioner or physician assistant.	Previous overdose and relapse and improvement in reported quality of life.	LR Radial Basis SVM RF AdaBoost Gentle Boost Logit Boost Robust Boost Total Boost	The best model (RF) had a sensitivity of 0.81 and a specificity of 0.65 for predicting the number of people who drop out after 90 days. After one hundred and twenty days, these values reached 0.86 and 0.66, respectively. The research found that the participants' overall quality of life had improved, and it also found things like a history of overdose or relapse that made them more likely to drop out of treatment.
[41]	The purpose of the study was to design and assess a two-stage clinical decision-making framework that was based on ML. The framework was used to determine which OUD patients would stop taking their medication in less than a year.	As part of the research, a retrospective observational examination of claims from all Massachusetts payers was carried out between 2013 and 2015. Participants in the trial consisted of 5190 people who had commercial insurance, started buprenorphine therapy between January and December 2014, and had not used the drug for at least a year before beginning treatment. Participants in the study were required to have buprenorphine therapy. It was required that the therapy have been completed for a minimum of two months without the need for buprenorphine prescriptions.	A patient's age, gender, income, type of insurance, patient location (such as whether a county is urban or rural), patient's primary prescriber's speciality and training, whether or not the primary prescriber had formal training in addiction treatment and was a pain specialist, patients' early indications of treatment adherence, and patients' early signs of treatment adherence, were all taken into consideration. In addition, patients' early indications of treatment adherence were also taken into consideration. PDC within the first two months, PDC within the first two months, and PDC within the first three months.	LR DT RF XGBoost SVM ANN	48.7% of patients who started buprenorphine treatment discontinued it in the first year. The AUC-ROC statistic, often known as the C-statistic, varied from 0.55 to 0.59 for the first-stage models. According to the C-statistic, the models' ability to differentiate between individuals was statistically significantly improved when two- and three-month PDC information regarding patients' adherence to treatment was integrated. This improvement was statistically significant ($p < 0.001$).According to the study, the C-statistic ranged from 0.65 to 0.71 with the addition of a two-month PDC to the previously collected patient-level baseline data. Following the incorporation of a patient-level, three-month PDC as well as additional patient-level covariates, the C-statistic of the models improved from 0.69 to 0.74. With a value of 0.74, the DT model exhibits the highest C-statistic gain. Age, gender, assigned county urbanicity, insurance type, and a proxy measure of income status were all associated with treatment discontinuation.
[87]	Classify people in the United States who have advanced breast cancer and are not taking the tamoxifen that has been recommended to them.	By searching via the IBM MarketScan Commercial Claims and Encounters database and the Medicare Claims database, 3022 women aged 18 or older were identified as having been diagnosed with metastatic	The patient's age, gender, insurance type, site of care, and medical history were recorded at the beginning of each session, regardless of whether the patient was an inpatient or an outpatient. The diagnosis, the clinical	LR RF Boosted LR Feedforward Neural Net	All models' prediction accuracy was only moderate. The results of the LR were evaluated as having a 0.31 specificity (95% CI, 29 to 33), 94% sensitivity (95% CI, 89 to 92), and an AUROC of 0.64. Despite failing to predict

(*continued on next page*)

Table 1 (*continued*)

Author	AIM	Data Collection	Potential Features/Variables Considered	ML Approach	Findings
		breast cancer between the years of 2012 and 2017.	procedures, the co-morbidities, the pre-treatment procedures, the treatment, and any further features of the interactions involving health care.		nonadherence (positive predictive value 48%), the model successfully predicted adherence (negative predictive value 89%). Age ≥55 and pre-treatment procedures or therapies were among the top predictive indicators identified by variable relevance.
[21]	To build ML models to predict patients with Crohn's disease's nonadherence to azathioprine and help carers streamline the intervention process.	Patients with Crohn's disease who were prescribed azathioprine at any point between September 2005 and September 2018 were invited to participate in a cross-sectional study designed to be conducted at a single location. Some research data came from electronic medical records, and some from questionnaires.	Medication concern belief, education, anxiety, depression, medication necessity belief, medication knowledge	BPNN SVM RF LR	With LR, BPNN, and SVM, respectively, the average classification accuracy and AUC were 81.6%, 85.9%, and 87.6%, respectively. Nonadherence to AZA treatment was found to be associated with the following four risk factors: belief in medication concern (OR = 3.130, $p < 0.001$), education (OR = 2.199, $p < 0.001$), anxiety and depression (odds ratio = 1.190, p 0.001), belief in medication need (odds ratio = 0.004, $p < 0.001$), and belief in medication understanding (OR = 0.805, p = 0.013).
[51]	Developing an accurate model for predicting the degree to which Chinese kidney transplant patients would adhere to immunosuppressive treatment regimens in order to enhance and plan for future individualised interventions (RTPs).	In China, 1191 RTPs were the subject of a retrospective, multicentre, cross-sectional study from October 2020 to February 2021. The Basel Assessment of Adherence to Immunosuppressive Medicine Scale was utilised to evaluate the patients' level of adherence to the medication they were prescribed (BAASIS). Beliefs about Medicines Questionnaire (BMQ), Perceived Social Support Scale (PSSS), and General Self-Efficacy Scale (GSES) were some of the elements from the integrated theory that were used to construct the models. Other elements included the Health Belief Model (HBM), Theory of Planned Behaviour (TPB), and Beliefs about Medicines Questionnaire (BMQ).	Age, marital status, HBM-perceived barriers, using a pill box after transplantation, PSSS family, drug side effects prior to transplantation, TPB attitudes, time after transplantation, household income, and drug side effects after transplantation were all factors that were taken into consideration.	LR RF MLP SVM XG Boost	38.5% of the patients in the "derivation cohort" did not take their immunosuppressants the way they were supposed to. The SVM model outperformed the other four models with an average AUC of 0.75, a sensitivity of 0.59, with a specificity of 0.73. According to the findings of the study of the SHAP, the most important factors in the prediction model were the recipient's age, marital status, perceived challenges faced by the HBM, use of a pillbox following transplantation, and PSSS-family support.
[18]	To examine the accuracy of non-adherence prediction utilising information from remote, real-time drug dose measurements.	Four thousand one hundred eighty-two clinical trial participants were recorded using a smartphone video app as they took their prescriptions. The smartphone-based app validates MA in clinical studies using computer vision technology and offers real-time adherence monitoring. Participants in the study had a variety of ailments, including congestive HF and chronic obstructive pulmonary disease.	The patient's age, gender, and race, as well as any previous indicators of whether or not they adhered to the plan or not, and the length of time they had been taking the prescription.	XGBoost	It was established that observed adherence was the most accurate indicator of whether an individual will adhere to a plan in the future. The classification models accurately predicted experiment adherence (area under the curve = 0.83), future week adherence (area under the curve = 0.87), and subsequent day adherence (area under the curve = 0.87).
[88]	Determine which smoker subgroups adhere to nicotine replacement therapy and create profiles for each of those groupings.	Recommended Usual Care (R–UC, n = 315) and Abstinence-Optimized Treatment (A-OT, n = 302) were compared using data from a randomised controlled trial with two arms. A secondary classification tree	Motivation to quit, confidence in quitting, stress-related cravings, age of first cigarette, smoking bans at work, number of smokers in the household, the belief that medicine is most likely	GUIDE Classification Tree	Various nicotine addiction variables predicted treatment adherence in various treatment contexts. In the A-OT study, participants who expected to require more medication four weeks after stopping smoking and between four and sixteen

(*continued on next page*)

Table 1 (*continued*)

Author	AIM	Data Collection	Potential Features/Variables Considered	ML Approach	Findings
		analysis was carried out using these data. A-OT consists of eleven counselling sessions, seven to eleven automated medication reminders, three weeks' worth of pre-quit mini-lozenges, and twenty-six weeks' worth of a nicotine patch in addition to mini-lozenges, among other aspects. According to the findings of the analysis, there are subsets of smokers that respond favourably to nicotine patches. Mini-lozenges were employed in both treatment scenarios, and the A-OT participants' subgroups demonstrated high adherence to them. Primary care clinics run by two different health systems in Wisconsin were utilised in order to recruit participants for the study. It was required of participants that they be at least 18 years old, that they smoke at least 5 cigarettes per day, that they speak English, that they plan to stay in the area for at least a year, and that they are reachable by phone. In addition to this, they had to agree to give up smoking completely within the next month.	harmful, and hours spent around smokers.		weeks after quitting smoking were more dependent on smoking at the beginning and more vulnerable to smoking signals. People who were initially less dependent on smoking and those who were exposed to and sensitive to smoking cues had a greater likelihood of using medications. A higher level of adherence was linked to more self-motivation and confidence to quit smoking and the belief that medications to help people stop smoking are safe and easy to use. Negative beliefs about the acceptability and safety of medication were found to predict poor adherence.
[10]	Recommends an MA ML model for fibromyalgia treatment that could make up for the unreliability of information from online discussion forums for health.	Transfer learning is a subfield of ML that involves applying information and knowledge obtained from one data set to another. Using only a small amount of trustworthy survey data, the initial phase of this endeavour consisted of developing prediction adherence models for the condition that was the focus of the investigation. Using information from health social forums, these algorithms were then used to forecast MA. The predictive models were produced using the ensemble ML technique (RF). This study illustrates the transfer learning methodology by analysing data from the PatientsLikeMe social health forum and the Medical Expenditure Panel Survey.	The kind of drug, the length of time it has been used, the daily dosage, the out-of-pocket cost, the region, the age, and the sex of the patient.	Transfer Learning RF Ensemble ML	With an accuracy of 73.2% and an F1 score of 73.9, a model that was trained on MEPS patients who had fibromyalgia was able to predict the adherence of other MEPS patients accurately. The proposed approach was evaluated on fibromyalgia patients from PatientsLikeMe, and the MEPS-trained model accurately predicted adherence with an F1 score of 79.7. A more straightforward MEPS model was built after the daily consumption, and out-of-pocket expenses of the two predictors were eliminated from the model. After excluding these two variables from consideration, the accuracy of the PatientsLikeMe dataset as well as its F1 score, improved significantly.
[62]	Create a model that is capable of predicting MA in older adults who have chronic diseases by using SVM in conjunction with traditional statistical methods. After that, identify the variables that have the strongest effect on predicting MA in older patients suffering from chronic conditions.	The research examined 293 older adults with chronic diseases receiving care at a single tertiary institution. These patients were all over the age of 65. The self-report of Morisky was used to evaluate drug adherence. Data were collected through in-person interviews with respondents who were, on average, 73.8 years old. Patients' MA was investigated using a self-reported questionnaire. Participants were separated into two groups for the study: those who followed their	Factors that were taken into consideration included: gender, age, self-efficacy, depression, health literacy, knowledge of medication, perceived health status, activities of daily living, medication side effects, daily pill counts, number of drug types, length of time since diagnosis, occupation, monthly income, spouse, and level of education.	SVM LR	With sixteen predictor variables and both LR and SVM, the accuracy of classification was 71.1% and 97.3%, respectively. It was discovered that the best indicator of drug adherence was self-efficacy. When self-efficacy was the sole variable utilised, the accuracy was 72.4%. Multiple sclerosis was linked to the number of different medication types, the number of tablets taken daily, and the length of time since diagnosis, according to the findings of this study. People who used drugs on a more regular basis were

(*continued on next page*)

Table 1 (*continued*)

Author	AIM	Data Collection	Potential Features/Variables Considered	ML Approach	Findings
		prescribed treatment plans (120) and those who did not (173).			more likely to comply with their prescriptions.MA was also affected by things like age, health literacy, education level, and drug knowledge among older adults with long-term illnesses. When it comes to the accuracy of classification, SVM outperforms LR because it selects the best possible classifier to maximise the geometric margin between samples and minimise empirical classification errors. LR, on the other hand, only considers the geometric margin between samples.
Julius et al. (2021)	Method for evaluating the risk that a patient will not take their medication at the initial clinic session before the prescription is prepared. The prognostic procedure was described using the ANF prediction model, which uses non-clinical patient data as input factors.	In order to collect non-clinical data from outpatients being treated for chronic diseases, a standardised questionnaire centred on the concept of health beliefs was utilised. This was done to evaluate how well the proposed ML framework could forecast when patients fail to take their medication as prescribed.	Information about the patient that is not directly related to their clinical condition, such as the patient's perceptions, beliefs, and behaviours, not only does the designed framework include a search module with the title "Extending EMR through the Creation of Patient Profiles, Medication Prescriptions, Drug Information, Appointments, and Adherence Level Databases," which not only assists the patient in locating information, making complaints, and reporting adverse drug reactions from the comfort of their own room but also serves as a repository of data for research on MA, the module was designed to help the patient.	ANF	A taxonomy of intervention techniques, design requirements, and guidelines for successful integration and deployment in healthcare institutions were also offered. The framework's different components, features, and functionalities were also provided. This would be a true prognosis and prediction tool for enhancing patient adherence because it would be simple to determine the patient's level of non-adherence.

been identified as significant features that care providers can use to predict non-adherence behaviours and classify people with hypertension into subgroups. Age was the most significant factor in nonadherence to antihypertensive treatment. Different determinants of antihypertensive drug use were observed in people of different ages. Concurrent diabetes and race or ethnicity were the most significant predictors of MA in young adults aged 18 to 44. Diabetes, having health insurance, and living in a poor neighbourhood were significant predictors of MA in people aged 45 to 64. All the middle-aged diabetic persons living in low-income areas followed their doctors' advice and adhered to antihypertensive medication taking. The strongest predictors among the elderly aged 65 and older were household income and neighbourhood poverty. Several factors influenced people's behaviours within each age group. Diabetes strongly predicts how well young and middle-aged people take their medicines. However, the neighbourhood poverty level significantly impacted MA among middle-aged and older people. In both [79,80] research, hypertension patients with concurrent diabetes had better drug adherence.

By and large, amongst all the six above-reviewed articles that applied ML to analyse MA, the features that contribute to the prediction of hypertension MA emerged as medication-taking follow-up reminders, having concurrent diabetes, types of medications used, and the frequency of medication use, when patients use different kinds of antihypertensive medication, patients with more hospital visits, an educational level, marital status, general overuse, monthly income or household income, specific concern, health beliefs, disease duration, self-efficacy, social support, the number of available supplies at the beginning of the prescription, higher values of available supply, increasing the likelihood of past positive adherence, neighbourhood poverty, health insurance coverage, race, and ethnicity.

6.1.1. Potential features that contribute to the prediction of CVDs and statin adherence in research that applied ML analytics

Six studies were found that used ML analytics to identify potential features contributing to the prediction of MA in CVD patients [8,47,77,81–83]. [83] employed SVM to identify predictors of MA in patients with heart failure (HF). Data were collected using a self-reported questionnaire from 76 patients who visited outpatient clinics at a hospital in Bucheon, South Korea, between January and April 2010. Five predictors (gender, daily frequency of medication, medication knowledge, functional classification of the extent of HF using the NYHA functional class, and spouse) and seven predictors (age, education, monthly income, ejection fraction, cognitive function measured using the Mini-Mental Status Examination-Korean, medication knowledge, functional classification of the extent of HF), respectively, were used in the two models that best classified MA in HF patients. The most accurate model yielded a detection rate of 77.63%. In both the five- and seven-variable SVM models, there was a definite preference for medication knowledge and the functional classification of the degree of HF as predictors of MA in HF patients. In particular, knowledge about medications as a single

feature was also a significant predictor.

In their research [77], used three different models - LR, LASSO regression, and RF-generated models - to find the best way to predict medication nonadherence when taking chronic lipid-lowering medications. In each of the models, prior statin use was the most important factor in determining whether or not someone would adhere to the use of statins in the future. Even though the LASSO regression model selected more than 90% of all available predictors, the discrimination of all three analytical techniques was only moderate (the C-index ranged from 0.664 to 0.673). Compared to patients who had just started taking a statin after a myocardial infarction, individuals who had previously been prescribed statins had significantly greater odds of adhering. In addition, the research found significant relationships between adherence and factors such as race, the number of cardiac medications given out at discharge, the region, and having coronary artery bypass graft (CABG) surgery during the period of the index stay. Diabetic patients and those recently hospitalised had a lower likelihood of adherence. In addition to the LR and LASSO regression models, the RF demonstrated that prescriptions for statins within the past six months were the most accurate predictor of lipid-lowering prescription medications.

Karanasiou et al. [47] identified HF patients as adhering to or not adhering to medication using RF, basic CART, RT, LMTs, the RBF network, RF, SVM, BayesNet, MLP, and naïve Bayes. The results imply that medication knowledge affects HF patients' overall adherence and how well it will be predicted. In addition, it predicts whether the individual suffering from heart failure will comply with all areas of patient management, such as the medicine, food, and exercise that have been prescribed. The study considered the following features: age, gender, carers, allergies, and medical condition (such as the severity of post-myocardial infarction HF), the extent of HF, the patient's smoking and alcoholism habits, the presence or absence of comorbidities, the dose and frequency of medication intake, biological data related to HF disease, and clinical examinations (left bundle branch block or intraventricular delay; left ventricular ejection fraction). The results show that the patient's health and medication use can predict adherence.

Hu et al. [8] investigated how long patients stayed on their prescribed medications after being admitted to the hospital for the first time due to CVD to predict how well they adhered to lipid-lowering therapy. The models were developed using boosted regression trees (BRT), linear regression, SVM, and ANN. Three hundred and thirteen thousand, two hundred and seven (313,207) patients from a sub-cohort of a web-based cardiovascular disease risk assessment and treatment decision support system for primary care in New Zealand were examined. The evaluation was done with the help of a proportional hazard model. The proportional hazard model indicated a strong association between adherence and the following features: age, a previous diagnosis of diabetes, the severity of complications and co-morbidities, obesity, the number of hospital days spent, the type of CVD that was diagnosed, in-patient procedures, and being new to therapy. BRT performed the best when predicting cases that might be known soon after discharge, with a c-statistic of 0.811 for a 1-year forecast and 0.793 for a 2-year prediction. Individuals between the ages of 30 and 54 were significantly less likely to take their medications as directed, but the compliance rates for women were marginally higher. In comparison to European patients, Indian patients had a lower MA. It was also discovered that patients from poorer communities had less MA than their counterparts from wealthier communities. Findings also reveal that patients' perceptions of their health may have a significant bearing on how well they can comply with the requirements of their prescriptions.

Lucas et al. [82] used the RF to make an EMR-based model for statin adherence. They linked it to biological and clinical outcomes in statin therapy patients. They obtained the EMR data from the military health system, which maintains administrative data on United States armed forces members eligible for health benefits for themselves or their families. The most significant features were age, the number of times the patient interacted with a healthcare system, the length of time a patient was followed up, the features of the statin prescriptions a patient was given, intestinal disease, hypertension treatment, hypertension and heart disease, and women's health/menopause. Patients who adhered to their statin therapy were more likely to be younger, female, smokers, have longer follow-up periods, and interact with the healthcare system more frequently than patients who did not adhere to statin therapy. Additionally, those who continued their statin therapy had higher initial doses and more supply days of the medication. Patients with diabetes or pre-existing CVD had greater rates of statin adherence. The adverse effects of statins on musculoskeletal symptoms, such as chronic pain, have been extensively reported. These effects may contribute to conditions that are associated with significantly reduced adherence.

Mirzadeh et al. [81] used SVR, RF regression, GTB regression, AdaBoost regression, DT regression, and linear regression to identify MA in patients at risk for atherosclerotic CVD. Participants were recruited through recommendations from primary care doctors at one Southern California outpatient tertiary care facility. The adherence prediction model showed an average error rate of 12.9% when applied using the top 10 variables. Results show that having a spouse, being told to take an active role in their care, feeling confident in their abilities to handle health emergencies, and having medication knowledge were all strongly related to MA. The findings highlighted the significance of clinical and psychological variables in predicting MA. The performance of the SVR and GTB regressions was superior to that of the linear regression. The SVR model had the best overall performance, with an error rate of 16.6%.

Six articles using different ML methods were reviewed to ascertain the features significantly predicting MA in patients taking statins or having CVDs. The features that emerged are age, education, monthly income, ejection fraction, cognitive function measure, medication knowledge, functional classification of the extent of HF, gender, daily frequency of medication, spouse, prior statin use, race, number of CVD medications prescribed at discharge, having a CABG procedure during their index admission, diabetes history, complication and comorbidity level, obesity, days spent in the hospital, CVD diagnosis type, and inpatient procedure.

6.1.2. Potential features that contribute to the prediction of MA amongst diabetes patients in research that applied ML analytics

Gu et al. [2] investigated the accuracy with which drop history data obtained from a drug event monitoring system may be used to predict an individual patient's level of adherence. The ML techniques used in the study were randomised trees, RF, XGBoost, LSTM RNN, gradient boosting, MLP, and ensemble learning. Most of the data used in the modelling process came from previous patient injection (drop) data obtained via a connected smart, sharp bin (SSB). The study used an Internet-connected SSB to track injection protocol adherence. The extra trees classifier was used to determine the most important ML model-building features. The most important features were the patient's residence, the number of times injections had been administered to them in the past, and the day they were due to receive their medication. Historic patient injections accounted for 94% of the overall significance of the features, making them the most significant feature. It was clear from this that historical drop data was the most useful information for prediction. The day of the week the medication was taken and the number of times it was taken each day was the second and third most important features. The best ML model, LSTM, effectively predicted patients' adherence to self-injection at home, with an AUC of 0.8902 and an accuracy of 82.90% on the unseen testing set.

Fan et al. [84] used ANN, QUEST, the Bayesian network [BN], discriminate, CHAID, CRT, and ensemble models to assess the ability of ML techniques to predict MA among T2D patients. The study was conducted at the Sichuan Provincial People's Hospital between January and December 2010 on the complications and blood glucose prognosis of type 2 diabetes patients who did not adhere to their treatment plans. The most significant risk factors for complications and insufficient glycaemic

control were the duration of type 2 diabetes, the period of uncontrolled hypoglycaemic medication, the type of insulin used, the quantity of hypoglycaemic drugs used, and the total cost of hypoglycaemic therapy. Non-observed risk factors for managing T2D-related hyperglycaemia included the amount of hypoglycaemic medication administered and the duration of time a person had T2D. Relatedly [12], used neural networks and BN, among other ML techniques, to build models for predicting the likelihood of a patient not adhering to their treatment plan. It was determined that the length of time a patient has been on their current medication regimen, previous blood sugar readings, and eating habits are the most significant predictors of MA.

Lo-Ciganic et al. [85] leveraged random survival forests (RSFs) and survival trees (STs) to investigate the correlation between oral hypoglycaemic MA and avoiding hospitalisation. They also used ML to predict the likelihood of hospitalisation based on the best levels of adherence. The highest possible ratings for determining compliance ranged from 46% to 94% for each hospital stay. Historical patient injections, medication frequency, country, and day of the week for the scheduled medication schedule, the number of hypoglycaemic medications, the duration of type 2 diabetes, the duration of the current medication regimen, the former blood glucose level, eating habits, the patient's general health, and the complexity of the prescriptions the patient receives were found to be the most significant predictors of MA in diabetic patients. Furthermore, the amount of hypoglycaemic medications used is an important predictor.

6.1.3. Potential features that contribute to the prediction of MA among patients with opioid use disorders (OUDs) in research that applied ML analytics

In three previous studies [23,41,86], ML models were developed to predict MA in patients with OUDs [86]. used LR, radial basis models, SVM, RF, AdaBoost, gentle boost, logit boost, robust boost, and total boost to create models for predicting an individual's likelihood of discontinuing an OUD rehabilitation program. The research found specific dropout risks factors, such as previous medication overdoses, drug abuse, and relapse. The most significant factor was whether the person had previously overdosed. Overdoses were more common among those who dropped out of the programme than among those who stayed. This indicates that those with a protracted drug abuse history may require additional assistance to continue outpatient treatment. The second most significant factor was the rise in living quality. Only 64% (after 90 days) or 59% (after 120 days) of individuals who stayed in the programme reported an increase in their quality of life, compared to more than 94% of those who left. Last but not least, reported relapses since joining the programme occurred 1.8–2 times more frequently in the group that dropped out. After 90 days, the strongest model (RF) had a sensitivity of 0.81 and a specificity of 0.65 in predicting dropout; after 120 days, these values increased to 0.86 and 0.66, respectively.

Stafford et al. [90] used RF, CART, bagged CART, and boosted CART to identify the key risk variables for early termination of OUD therapy. The developed ML models aimed to determine which features were most important for predicting whether or not a patient would discontinue treatment before it was completed. The results achieved with RF (74% ROC-AUC) were much better than those achieved with other ML approaches. Notably, they found that the treatment provider's location, the main source of treatment funding, and the source of treatment referral were better at predicting treatment dropout than individual-level factors. The most important risk factors were the setting where medical care was given, the location, the main source of funding, and the source of referral. Race, ethnicity, heroin use, injection drug use, and gender all had significantly less predictive power and were given poor significance ratings. The type of service (rehab, ambulatory, or detox) was the most important feature that determined the outcome. When it was taken out of the model, there was a more than 5% drop in the accuracy of the results.

Hasan et al. [41] designed and evaluated a two-stage clinical decision-making framework based on ML to identify patients who were likely to discontinue OUD medication within the next twelve months. The study's sample population consisted of individuals with commercial insurance through "Massachusetts All Payer" who had claims submitted between 2013 and 2015. Multiple ML models (LR, DT, RF XGBoost SVM, and ANN) performed exceptionally well during the initial two cycles of buprenorphine treatment or OUD therapy. Patients' ages, genders, incomes, the median household income in their assigned county, how urban their assigned county was, insurance type as a supply-side indicator, and insurance statuses were all important in predicting treatment dropout. However, PDC in the early treatment phase was the most important factor. Additional information regarding the extent to which patients adhered to their prescribed treatment regimens throughout the initial treatment period, as measured by a two-month and three-month PDC, substantially improved the statistical ability of the models to differentiate between groups. The DT model obtained the highest C-statistic gain, 0.74, out of all the models. There was a correlation between treatment discontinuation and several demographic factors, including age, gender, the extent of urbanisation in the assigned county, insurance type (an indicator of the supply side), and income level.

6.1.4. Potential features that contribute to the prediction of MA among other NCDs in research that applied ML analytics

Yerrapragada et al. [87] analysed tamoxifen non-adherence among commercially insured metastatic breast cancer patients in the United States using real-world data and ML methods. Among the ML techniques used were LR, boosted LR, RF, and feedforward neural network models. These ML models were trained using clinical procedures, comorbidity, treatment, and health care encounter data. The age of 55 and the treatments received before therapy (lymphatic nuclear medicine, radiation oncology, and arterial surgery) were two of the most significant predictors of survival. In another study [91], developed ML models that could predict when Crohn's disease patients will fail to take their azathioprine as prescribed. Back-propagation neural networks (BPNN) and SVM were used to build ML models, which were then compared to LR models. The average classification accuracy for SVM, BPNN, and LR were 87.7%, 85.9%, and 81.6%, respectively. According to the multivariate analysis, education, anxiety, depression, and beliefs about medication increased the likelihood of a patient not taking AZA as prescribed.

Julius et al. (2021) employed a predictive method to predict patients' medication non-adherence levels during the initial clinic visit before prescriptions were issued. Their study revealed an adaptive neuro-fuzzy (ANF) predictive model incorporating patient non-clinical data as input variables. Non-clinical observations from the study include perception, belief, and behavioural tendencies. The framework developed includes a search module captioned "Extending EMR through the creation of patient profile, medication prescription, drug information, appointment, and adherence level databases," which not only allows patients to conduct information searches, file complaints, and report adverse drug reactions from the comfort of their rooms but also serves as a dataset repository for research. There was also a taxonomy of intervention methods, design requirements, and guidelines for effective integration and deployment in healthcare institutions. The various components, features, and functionalities of the framework were also provided. The patient's medication non-adherence would be instantly observable, making this a valuable prognosis and prediction tool for improving MA.

Kim et al. [88] identified a group of smokers who adhered to nicotine replacement therapy using the GUIDE classification tree. Adherence was noticeably higher in people who needed nicotine more frequently and were more easily influenced to smoke. A higher level of adherence was linked to a stronger desire and confidence to quit smoking and the belief that medicines to help people stop smoking are safe and easy to use. Poor adherence was found to be linked to having negative thoughts about how acceptable and safe medication is [10]. proposed an MA-ML model

for treating fibromyalgia to account for data inaccuracy from social health forums. With an F1 score of 73.9, a model that was trained on fibromyalgia patients in the Medical Expenditure Panel Survey (MEPS) could accurately predict adherence for other MEPS patients. Medication type, number of years of therapy, frequency of use, dosage, out-of-pocket expenses, location, age, and gender are among the variables examined. The proposed method was tested on PatientsLikeMe fibromyalgia patients, and the MEPS-trained model correctly predicted adherence with an F1 score of 79.7. After removing daily usage and out-of-pocket costs, a modified version of the MEPS model was created. The precision of the PatientsLikeMe dataset and its F1 score were both improved as a consequence.

In a separate study that also employed XGBoost [18], investigated the accuracy with which remote real-time drug administration measurements can predict non-adherence. A smartphone application monitored clinical trial participants as they took their medication. The study's subjects suffered from various illnesses, including chronic obstructive pulmonary disease and congestive HF. It was discovered that empirically observed adherence is the most accurate indicator of whether or not a person will continue to adhere to prescribed medication in the future. The classification models made accurate predictions for participants' adherence to the experiment (AUC = 0.83), the following week (AUC = 0.87), and the following day (AUC = 0.87), respectively.

Lee et al. [62] developed a predictive model for MA in older patients with chronic conditions. They used SVM with more conventional statistical methods to identify the features that significantly predicted MA in the patients. LR and SVM classification accuracy with various predictors was 71.1% and 97.3%, respectively. Self-efficacy was found to be the best predictor of MA. According to the results of both the LR and SVM analyses, self-efficacy is an important factor in determining MA in the Korean population of older adults. People with more prescriptions filled were more likely to adhere to the treatment regimens laid out for them. The prevalence of MA in older people with chronic diseases was affected by various factors, including age, health literacy, depression, education level, and understanding of medications. Regarding classification, SVM is preferable to logistic regression (LR) because it chooses the best classifier to maximise the geometric margin between samples and minimise empirical classification errors. SVM is more accurate than LR even though it utilises fewer variables. This is because it builds an ideal classifier that minimises empirical classification errors while differentiating between patient samples.

Zhu et al. [51] employed LR, RF, MLP, SVM, and XG Boost to individualise future interventions better to predict immunosuppressive drug adherence in Chinese kidney transplant patients. In China, 1191 RTPs were the subject of a retrospective, multi-centre, cross-sectional study from October 2020 to February 2021. This research aimed to establish whether or not there was a connection between not adhering to the RTP and aspects of the integrated theory. In particular, ML techniques were utilised to develop the models and establish which features had the greatest influence on non-adherence. The BAASIS was used to evaluate the patient's level of adherence to the treatment regimen. The models were built by using the earmarked data in conjunction with the HBM, the TPB, the BMQ, the PSSS, and the GSES. The SHAP method was utilised to determine the relative importance of each feature used in predicting the chance of IM nonadherence in RTPs. The SHAP analysis results indicate that the most important factors in the prediction model are the respondent's age, marital status, reported difficulties in the HBM, use of a pillbox following transplantation, and support from family members in the PSSS. The SVM model outperformed the other four models in terms of performance. It had a sensitivity of 0.59, a specificity of 0.73, and an AUC of 0.75.

6.2. Methods utilised in research employing ML techniques to determine the significance of NCD MA predictors

Several FS and screening techniques emerged. These various techniques include multivariate analysis yielding ORs and p-values, univariate analysis for continuous variables using t-tests, variance analysis, or the Wilcoxon signed-rank test, chi-square test, or fisher's exact test for categorical variables, the SHAP, correlation analysis, the change in mean squared error (MSE), the Akaike information criterion (AIC), the forward, backwards, and stepwise methods of LR, forward and backward feature selection through cross-validation, the RFE, permutation of OOB predictor observations through cross-validation RF permutation importance (yielding the OOB error rate), predictor importance (PI), Mann–Whitney U, Fisher exact test, q-values (p-values adjusted for Benjamini-Hochberg (BH) false discovery rate), Waikato environment for knowledge analysis (WEKA), cox proportional hazard modelling yielding hazard ratios [HRs]), the loss in predictive accuracy by withholding a particular variable and calculating the change in the AUROC.

6.2.1. Multivariate feature analysis

Wang et al. [21] conducted a multivariate analysis with BPNN, SVM, RF, and LR to identify risk factors for not taking AZA medications as prescribed. That study reported the odds ratios (ORs) in their order of magnitude, accompanied by significance values in terms of p-values ($p < 0.001$) [77]. compared predictive analytic approaches to characterise medication non-adherence that leveraged LR, LASSO, and RF. They adopted multivariate LR analysis, a common classification algorithm in data science and ML, to predict the outcome variable (MA) depending on multiple independent variables. In their multivariate LR analysis results [77], reported the most important variables reflected through the ORs. Their backward stepwise selection LR model found that factors such as race, the number of CVD medicines taken at discharge, geographic location, and the presence of a CABG procedure during the index admission were all statistically significant ($p < 0.001$).

In order to establish which factors had a significant impact on MA modelling [12], used different kinds of LR analysis: forward, backwards, and stepwise. In a different multivariate analysis, the Cox proportional hazard models with 95% confidence intervals for HRs were used to estimate the influence of different factors on the amount of time a patient spends using medications [8]. This was determined through the proportional hazard model, a type of survival model in which variables such as age, diabetes history, obesity, and the number of hospital days spent were strongly associated with drug persistence (the dependent variable). An increase of one unit in a covariate was shown to have a multiplicative effect on the hazard rate when analysed using the proportional hazard model.

6.2.2. Univariate and bivariate feature analysis

Using univariate analysis, it was also feasible to determine the importance of the variables that predicted MA or non-adherence. In their study [84], used t-tests, variance analysis, and the Wilcoxon signed-rank test as techniques for performing univariate analysis on continuous data. The chi-square and fisher's exact tests were used to analyse the categorical variables. It was established that the data was statistically significant when the p-value was less than 0.05 [79]. used the chi-square and Wilcoxon signed-rank tests to differentiate characteristic variables from binary and continuous variables. In another study [21], used the *t*-test, the Mann–Whitney *U* test, and a one-way analysis of variance on continuous variables. On the other hand, they used fisher's exact test or the chi-square test on categorical data. Once more, p-values that were lower than 0.05 were regarded as being statistically significant [86]. estimated the PI or ranks by permuting OOB predictor observations in the RF classifier model to determine how the significance of the predictors should be ranked. In addition, the Mann–Whitney *U* test was used to compare continuous variables between the OUD therapy program's maintained group and the dropped-out group;

and the Fisher exact test was utilised to analyse dichotomous variables.

A Benjamini-Hochberg (B–H) false discovery rate in other analyses and significant factors were chosen based on their adjusted p-values translating to q-values. The expected fraction of significant tests that are incorrectly identified as significant is referred to as the false discovery rate, abbreviated as FDR. Adjusting the rate helps account for the fact that small p-values (less than 5%) can arise by accident, which may lead someone to reject the true null hypothesis incorrectly. This can be prevented by adjusting the rate. Individuals who use the B–H strategy have a better chance of avoiding type I errors, also known as false positives. In related findings, Bonferroni-adjusted chi-square tests were used to identify all variables with a statistically significant (small p-value) relationship with the response variable in an M-based population segmentation strategy that identified population profiles associated with not taking blood pressure medication as prescribed. This approach was used to identify population profiles associated with a higher likelihood of not taking hypertension medication as prescribed [80]. The probability of making a type I error arises as multiple statistical tests are carried out, which is why the Bonferroni correction adjusts the probability values (p). The study by Ref. [82] used a bivariate method to model the link between the features in an EHR dataset. Specifically, this was done in order to reduce the number of features that needed to be evaluated in building prediction models for statin adherence. The correlation between variables was also useful for integrating correlated variables into meaningful groups or features that could facilitate the clinical interpretation of correlated data elements. In order to establish the nature of the correlation that exists between the two groups of data, the Pearson method was applied. In the research carried out by Ref. [84]; the level of statistical significance for a variable was established at 0.9 (1−α).

6.2.3. Waikato environment for knowledge analysis (WEKA)

WEKA was utilised by Ref. [2] to minimise the number of dimensions or features included in the feature vectors and select the most relevant for the prediction of MA using ensemble learning and DL models. Highly correlated features were identified and eliminated by utilising the information gain-based FS technique and the "InfoGainAttributeEval" attribute evaluator provided by WEKA. After that, it was used to determine and select the features that represented the most accurate prediction information. This attribute evaluator enables the computation of the information gain, or entropy, for each dataset feature pertaining to the target variable (MA or non-adherence). Entropy is another name for information gain. Using this method to sort through the myriad ways the dataset variables could be combined, they identified the dataset features that were most relevant to the ML analytics task. Thereafter, the extra trees classifier was then used to identify the most important features based on the feature importance score, which varied from 0 to 1. The scores provide information about a factor that can be used to determine the target adherence class. These scores indicate the importance of each feature in terms of the precision with which accurate predictions can be made about the target variable.

6.2.4. Filter methods and mutual information (MI) measures

Karanasiou et al. [47] employed a filtering technique to determine the most important features. Since the filter FS method was applied, the FS procedure was executed independently of the induction algorithm. Specifically, by comparing the IG of each variable to the target variable, the IG, a well-known filter model and approach, was used for FS. This was done so that the IG could be determined more accurately. Since Karanasiou et al. (2016 also employed tree-based ML methods, the IG assisted in building decision trees from a training dataset by evaluating the information gain for each variable and selecting the variable with the highest information gain. Entropy is thus reduced, and the dataset is successfully segmented into groups for classification. For example, the J48 algorithm, which builds a decision tree using the C4.5 tree inductive process, first splits each tree node according to the feature with the maximum normalised information gain and then recurses on smaller sub-lists. This ensures that the algorithm generates the most accurate results possible. A lead node will be generated to select that class if all the samples belong to the same class. If the features do not satisfy the criteria for splitting, a decision is made at a higher level using the expected class values. This method is also called into action whenever an instance of a class that had been hidden from view is uncovered. The features of multidimensional datasets were evaluated using the gain ratio, an alternate approach to ranking features. This technique removed low-ranking features from the evaluation to generate new, more reduced datasets.

6.2.5. Backward feature elimination

Backward feature elimination is an advanced technique for FS that enables the selection of an optimal number of features by removing those that do not significantly affect the target variable or prediction of outcome [83]. used the data from a questionnaire survey to find the most efficient way to identify MA predictors in HF patients. This was achieved by applying SVM to all possible feature combinations (2^{11-1}) of each of the 11 features that were included in the dataset. They found the feature combinations that provided the maximum level of accuracy. Similarly, the RF variable importance technique was used for FS in a study that used RF, ANN, SVR, and self-organizing feature maps (SOM) to discover features linked with hypertension patients' adherence levels. This study aimed to identify the most significant predictors of patients' adherence to their prescribed treatment regimen. Specifically, the FS technique was applied to a trained RF model using the backward elimination procedure to identify significant MA-related factors [22]. The variables were then ordered according to the increase in their mean squared error (MSE), and backward elimination was used to calculate FS scores. The model's RMSE was determined after going through the process of eliminating each variable in ascending order. If the RMSE value increases, then using a backward elimination technique to eliminate a variable reveals its significance to the level of adherence in hypertension patients. A SOM was used to exhibit higher-dimensional nonlinear data and feature distribution in the unified distance matrix map. This was done to understand better the link between the variables that relate to the adherence levels among hypertension patients. All ML algorithms were designed with these important variables in mind, which were uncovered in the FS stage.

Similarly [22], determined the significance of each variable by calculating the average change in MSE across all decision trees. When building the model, the most important variables were the patient's age, length of stay in the index hospital, and the number of statin prescriptions filled in the previous six months. Based on the LR and LASSO regression models, the most important predictor of lipid-lowering medication was the number of statin prescriptions filled in the last six months. In a study by Ref. [51]; RFE, another strategy for selecting predictors backwards, was used. This FS method involves going through the dataset one feature at a time and removing any redundant or irrelevant features based on how well the model performed in the cross-validation test. In applying the RFE approach, the process started by developing a model that included all the predictors and assigning a significance score to each. The least significant predictors were removed from the model, and the importance ratings were performed systematically.

Lucas et al. [82] calculated how much their model's accuracy would decrease if a specific variable was removed and then examined the change in AUROC by weighing the relative importance of the baseline variables. This analysis was conducted to determine how much the accuracy of their model would decline if a particular variable was omitted. The most significant predictors of MA were identified and incorporated into an ML model. These predictors included but were not limited to the length of follow-up, age, and statin prescription features. Additionally [23], highlighted how altering a model by including or excluding a significant variable illustrates the significance of the feature. In their

study, when an important variable, "treatment setting", was removed, the accuracy of the test model dropped by more than 5%. It thus emerged that this was the variable that had the most impact. In the study by Ref. [87]; they chose the ML model with the highest AUROC to depict the varying significance of baseline features in classifying tamoxifen non-adherence among commercially insured metastatic breast cancer patients in the United States.

6.2.6. Out-of-bag (OOB) error

Three studies used the OOB error rate to rank the importance of variables in ML-based analytics of MA [21,22,86]. The importance of the variables was ranked in a study by Ref. [22]; who used ML to identify features linked with hypertensive patients' adherence levels. The most significant variable had the largest increase in the mean error percentage, while the least significant variable had the smallest increase. The order in which the variables were ranked guided the backwards elimination process. The error rate was computed at the end of each phase of elimination. If there was an increase in the error rate, then the variable that had been omitted was considered relevant or critical to MA [21]. built ML models that could accurately predict which patients with Crohn's disease would fail to take their medication as prescribed. The "FS" function found in the MATLAB Statistics and ML Toolbox was used in their study to determine the significance of each feature [86]. also estimated the predictor importance (PI) by shuffling the observations of the OOB predictor within the RF classifier model. In the same study, the Mann–Whitney U and Fisher exact tests were used to compare the groups of participants who remained in the OUD therapy programme and those who dropped out. The Mann–Whitney *U* test was used for continuous factors, and the Fisher exact test was used for dichotomous factors.

6.2.7. Akaike's information criterion (AIC) minimisation

Zullig et al. [77] developed a LASSO model and concurrently used Akaike's Information Criterion (AIC), among other approaches. The AIC is a metric for evaluating the accuracy of predictions and, by extension, the validity of statistical models for any given data set and set of features. The LASSO model used the AIC minimisation feature, which enabled approximately 90% (25 of the 28 variables) selection of all relevant features. The majority of these variables were also chosen using the LR model. Regarding the magnitude and direction of the relationships, the LASSO model generated comparable outcomes to the LR model. The OR values for the candidate features provided evidence of their relevance. Patients prescribed statins during the preceding six months had a considerably higher likelihood of continuing to take them (OR 3.55). An OR calculates the relationship between a variable and the likelihood of an event occurring (in this case, MA or non-adherence).

6.2.8. SHapley additive exPlanations (SHAP)

The SHAP method was used in three previous studies [41,51,78] to determine how significant MA predictors were for patients with NCDs. SHAP calculates how much each model's input variable contributes to the overall prediction accuracy. The SHAP method was used in the study by Ref. [51] to assess the accuracy with which predictors might predict the likelihood of Chinese RTPs not taking immunosuppressants. The RFE method was used to eliminate irrelevant features, followed by the SHAP method to rank the predictors in order of significance. The SHAP library's beeswarm and bar plots were largely used to illustrate the contribution or significance of each feature to the model's prediction. The beeswarm plot displayed a summary of how the top features of the datasets influence the model's output. Fig. 8 illustrates the application of SHAP concepts by displaying the SHAP values for each feature and sample. The SHAP value is displayed along the horizontal axis, and each line in the graph denotes a different feature. The SHAP values quantify the amount and direction of each feature that impacts the predicted MA. In each instance, the given explanation is represented by a single dot on each feature row. Thus, all the little dots on the plot represent a single observation or data point. In the SHAP plots, colour is used to display the original value of a feature. The colour shows if that data point has a higher or lower value when compared to other data points. Data points in red represent high feature values, while those in low blue values. For instance, as depicted in Fig. 8 below, it could be observed that lower values of age have positive SHAP values (the points extending towards the right are increasingly blue), and higher values of age have negative SHAP values (the points extending towards the left are increasingly red). This indicates that age negatively predicts the risk of non-adherence to medication. Also, results presented in Fig. 8 show that an increased number of HBM-perceived barriers is associated with barriers that negatively impact the prediction, while a decreasing number has a positive impact. It is also critical to note that on the beeswarm, the features are ordered by their effect on prediction, with the first being the most important and the last being the least important. The importance of each feature can be determined using the SHAP algorithm by comparing the relative magnitude of the mean absolute SHAP values for the various features. The most important features were a person's age, marital status, and how the HBM views the limitations imposed on them. This is illustrated in Fig. 8. After transplantation, the adverse effects of the medications are the least important factor.

Galozy et al. [78] also utilised Beeswarm SHAP plots to examine the extent to which predictors had an impact on model output. The SHAP framework made compiling local model explanations for the test set predictions possible. This was done to identify which factors were most relevant for predicting MA in a group of hypertension patients. According to the results of their beeswarm plot, the most significant factor in determining model output was the number of available supplies at the

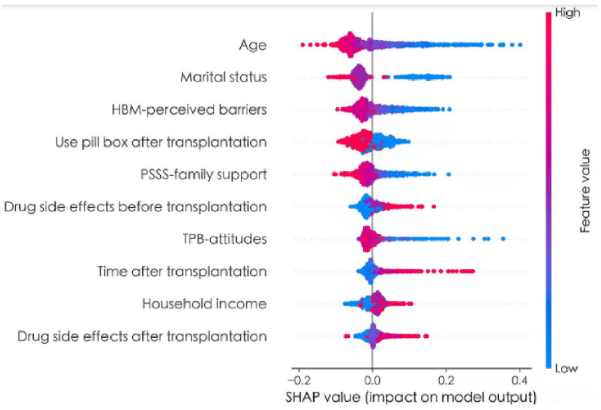


Fig. 8. The SHAP values of predictors (Source: Adopted from [51].

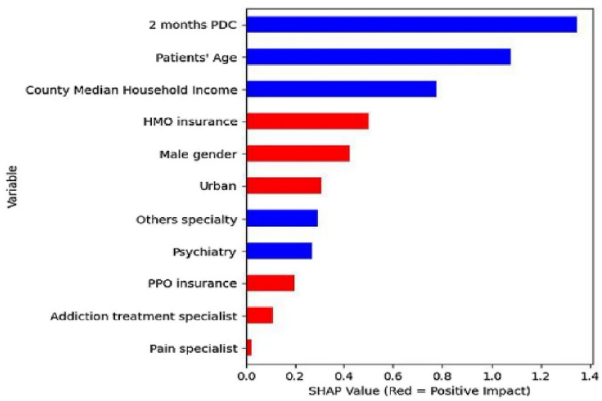


Fig. 9. Shap values (source: Extracted from [41].

study, when an important variable, "treatment setting", was removed, the accuracy of the test model dropped by more than 5%. It thus emerged that this was the variable that had the most impact. In the study by Ref. [87]; they chose the ML model with the highest AUROC to depict the varying significance of baseline features in classifying tamoxifen non-adherence among commercially insured metastatic breast cancer patients in the United States.

6.2.6. Out-of-bag (OOB) error

Three studies used the OOB error rate to rank the importance of variables in ML-based analytics of MA [21,22,86]. The importance of the variables was ranked in a study by Ref. [22]; who used ML to identify features linked with hypertensive patients' adherence levels. The most significant variable had the largest increase in the mean error percentage, while the least significant variable had the smallest increase. The order in which the variables were ranked guided the backwards elimination process. The error rate was computed at the end of each phase of elimination. If there was an increase in the error rate, then the variable that had been omitted was considered relevant or critical to MA [21]. built ML models that could accurately predict which patients with Crohn's disease would fail to take their medication as prescribed. The "FS" function found in the MATLAB Statistics and ML Toolbox was used in their study to determine the significance of each feature [86]. also estimated the predictor importance (PI) by shuffling the observations of the OOB predictor within the RF classifier model. In the same study, the Mann–Whitney U and Fisher exact tests were used to compare the groups of participants who remained in the OUD therapy programme and those who dropped out. The Mann–Whitney *U* test was used for continuous factors, and the Fisher exact test was used for dichotomous factors.

6.2.7. Akaike's information criterion (AIC) minimisation

Zullig et al. [77] developed a LASSO model and concurrently used Akaike's Information Criterion (AIC), among other approaches. The AIC is a metric for evaluating the accuracy of predictions and, by extension, the validity of statistical models for any given data set and set of features. The LASSO model used the AIC minimisation feature, which enabled approximately 90% (25 of the 28 variables) selection of all relevant features. The majority of these variables were also chosen using the LR model. Regarding the magnitude and direction of the relationships, the LASSO model generated comparable outcomes to the LR model. The OR values for the candidate features provided evidence of their relevance. Patients prescribed statins during the preceding six months had a considerably higher likelihood of continuing to take them (OR 3.55). An OR calculates the relationship between a variable and the likelihood of an event occurring (in this case, MA or non-adherence).

6.2.8. SHapley additive exPlanations (SHAP)

The SHAP method was used in three previous studies [41,51,78] to determine how significant MA predictors were for patients with NCDs. SHAP calculates how much each model's input variable contributes to the overall prediction accuracy. The SHAP method was used in the study by Ref. [51] to assess the accuracy with which predictors might predict the likelihood of Chinese RTPs not taking immunosuppressants. The RFE method was used to eliminate irrelevant features, followed by the SHAP method to rank the predictors in order of significance. The SHAP library's beeswarm and bar plots were largely used to illustrate the contribution or significance of each feature to the model's prediction. The beeswarm plot displayed a summary of how the top features of the datasets influence the model's output. Fig. 8 illustrates the application of SHAP concepts by displaying the SHAP values for each feature and sample. The SHAP value is displayed along the horizontal axis, and each line in the graph denotes a different feature. The SHAP values quantify the amount and direction of each feature that impacts the predicted MA. In each instance, the given explanation is represented by a single dot on each feature row. Thus, all the little dots on the plot represent a single observation or data point. In the SHAP plots, colour is used to display the original value of a feature. The colour shows if that data point has a higher or lower value when compared to other data points. Data points in red represent high feature values, while those in low blue values. For instance, as depicted in Fig. 8 below, it could be observed that lower values of age have positive SHAP values (the points extending towards the right are increasingly blue), and higher values of age have negative SHAP values (the points extending towards the left are increasingly red). This indicates that age negatively predicts the risk of non-adherence to medication. Also, results presented in Fig. 8 show that an increased number of HBM-perceived barriers is associated with barriers that negatively impact the prediction, while a decreasing number has a positive impact. It is also critical to note that on the beeswarm, the features are ordered by their effect on prediction, with the first being the most important and the last being the least important. The importance of each feature can be determined using the SHAP algorithm by comparing the relative magnitude of the mean absolute SHAP values for the various features. The most important features were a person's age, marital status, and how the HBM views the limitations imposed on them. This is illustrated in Fig. 8. After transplantation, the adverse effects of the medications are the least important factor.

Galozy et al. [78] also utilised Beeswarm SHAP plots to examine the extent to which predictors had an impact on model output. The SHAP framework made compiling local model explanations for the test set predictions possible. This was done to identify which factors were most relevant for predicting MA in a group of hypertension patients. According to the results of their beeswarm plot, the most significant factor in determining model output was the number of available supplies at the

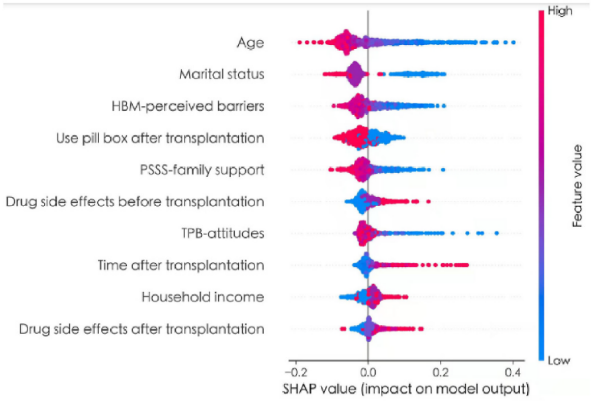


Fig. 8. The SHAP values of predictors (Source: Adopted from [51].

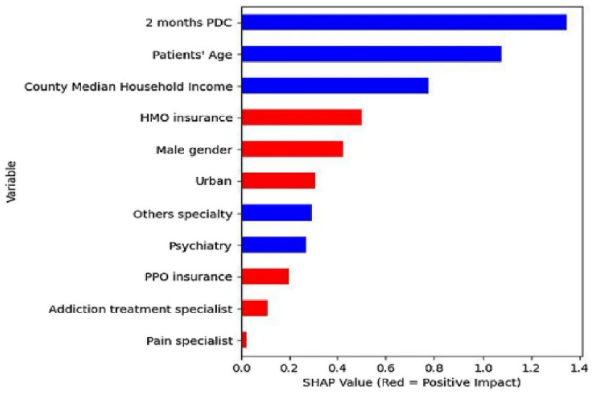


Fig. 9. Shap values (source: Extracted from [41].

start of the prescription. The plot indicated that higher values of accessible supply increased the likelihood of positive (adherent) output. Additionally [41], produced variable significance plots by utilising the SHAP values that were derived from the XGBoost model. SHAP presents an alternate way of calculating the relevance of each feature to make it more straightforward to understand. One can determine the importance of each feature by determining its average absolute value and comparing it to the SHAP value. This will result in a standard bar graph similar to Fig. 9. In this graph, the blue bar represents a negative influence, while the red bar shows a positive effect. According to Ref. [41]; Fig. 9 is a bar graph that depicts, in descending order of importance, the average effect of several features on OUD treatment dropout. The features are listed in the order in which they are most important. The bar graph also shows, in red and blue, whether these variables have a favourable (red) or unfavourable (blue) relationship with terminating OUD treatment. According to the SHAP bar plot findings, the presence of PDC throughout the early treatment period, specifically the first two months of treatment, negatively correlates with the patient's decision to stop treatment. This suggests that the PDC value should be increased when there is a higher likelihood that a patient will continue receiving treatment.

6.2.9. *Other feature importance measurement and selection approaches*

Koesmahargyo et al. [18] collated all of the data in one place after determining the importance of each feature using cross-validation. This was done to examine how each predictor factor fits into the different types of adherence models. The average increase in accuracy after adding a feature was used to assess its significance across all splits. When processing the testing dataset [10], used PI to determine the importance of each predictor. PI is the ratio of times a predictor is traversed to the total number of times all predictors are traversed in the RF model. In the RF model, decision trees were built by identifying the predictors leading to the greatest impurity drop. This is done to achieve optimal results. The frequency with which a predictor is chosen indicates its relative entropy compared to other predictors. This is true even though mechanisms are in place to prevent different nodes from repeatedly selecting the same predictor. As a result, if the PI is high, it suggests that the predictor is very important to the model. To establish whether a patient should be placed in the right or left branch, the selected predictor needs a reference value at each node in the network. The various possible values for a certain predictor are investigated using an iterative mechanism. Medication knowledge, medication types, age, educational status, number of medications taken daily, and income level were initially chosen because they were shown to affect MA in an earlier study on NCD patients or because a group of experts deemed they would have clinical value. Several variables were omitted from a distinct ML-based model of MA because there was either insufficient data for them or a poor correlation with the variable of interest [12]. [18] removed features with more than sixty percent missing values. This was done since the data structure did not allow these features to be populated.

7. Discussion

MA is affected by different features, as evidenced by this current research. These features include patient-level sociodemographic and socioeconomic factors, medication-prescriber-level features, medication information, healthcare system features, and disease-related features. Some of these factors are modifiable, and some are not. The non-modifiable factors include patient-level factors such as age, gender, race and marital status. To alter modifiable risk factors such as HBM-perceived hurdles and self-efficacy, medication knowledge, and TPB, NCD patients must be educated on why they need these medications and how to administer them correctly. Using various ML approaches for feature selection and screening, it has been determined that various features significantly impact predicting MA in multiple NCD conditions. The features that influence the prediction of MA across different NCDs and research articles are age, education level, gender, race or ethnicity, income level or household income, spouse, medication knowledge, disease duration, duration of the current medication treatment or intake, drug side effects, living in rural areas or cities, geographic region of care, past MA or PDC, health beliefs, social support or aids in medication intake, comorbidities, depression, medication type, insurance type or status, daily frequency of medication or daily intake or daily pill counts, job type, smoking history, smoking history, and follow-up rates.

Medication knowledge, for example, was identified as a significant predictor in Ref. [83] SVM model for identifying predictors of MA in HF patients; similarly [21], multivariate analyses identified medication knowledge as one of the four risk factors associated with predicting non-adherence to azathioprine in Crohn's disease patients. Another prior study discovered that medication knowledge is significantly related to MA in elderly chronic disease patients [62]. Multiple studies have found that comorbidities are a strong predictor of MA. For example [81], used comorbidities to predict MA in people at risk for atherosclerotic CVD [8], used comorbidities to predict the persistence of lipid-lowering therapy. Comorbidities labelled as "concurrent diabetes" were identified as a factor significantly impacting MA in hypertensive patients in a study by Ref. [79]. Sociodemographic and socioeconomic variables are predominantly the other features cutting across the various research works and NCD conditions of modelling MA or non-adherence. For instance, 'age' was a significant contributor to MA prediction across multiple studies [8,41,51,62,79,80,83]. These results resonate with findings in past studies that have shown not only age as the determinant of MA but also other variables such as gender and income status [92–94] and whether a patient stays in an urban or rural area [95,96], as also revealed in our study.

However, contradictory findings surfaced. In one study by Ref. [23]; race, ethnicity, and gender are insignificant predictors of who will discontinue OUD treatment; however, race or ethnicity is a significant factor that can be used to predict non-adherence to antihypertensive drugs in the study by Ref. [80]; and gender to predict MA in HF patients [83], as well as patients who will discontinue OUD [41]. The contrasting findings regarding gender being a weaker predictor in one study and a strong predictor in another study predicting MA among patients with the same condition (OUD) demonstrate that results may be susceptible to differences in the study setting, time scope, and population involved [8]. reported medication non-persistence among the younger age group (30–54) [87], identified patients of older age (>55 years) as one of the most predictive factors of medication non-adherence, and [82] found the younger to be more adherent to statin therapy than the older age group. Although Stafford et al. (2020) discovered that age did not affect model prediction accuracy, numerous studies have shown that young people are at risk of not adhering to a prescribed treatment plan [97–100]. It was discovered that various NCD conditions might have distinct features that contribute significantly to the prediction of MA. It is important to remember that if an elderly patient has other health problems, limited mobility, or is socially isolated, taking their medicines as prescribed and paying close attention to their restrictions and medical follow-up can be challenging [101]. discovered that individuals over 66 had poor MA. Surprisingly, older adults were less likely to take their prescribed medications. This could be because older people have poor memories, usually live alone, and, if they live alone, do not have someone to help them take their medications. This assertion is consistent with our research findings, which also showed the importance of self-efficacy, the presence of a spouse, and social or family support as key predictors of MA and better health outcomes. According to Refs. [102,103], marital status is associated with clinical outcomes, and patients receiving less support from their families are likelier to discontinue their medication.

Comorbidities, such as having both hypertension and diabetes, are associated with treatment adherence, as revealed by our findings. This is consistent with the findings of past research, which reported the same result, namely that people with hypertension and diabetes have better adherence (Tan et al., 2021; [104]. This could be because these patients

are conscious of the severity of their diseases, the consequences of failing to take their prescriptions as prescribed, and the importance of maintaining appropriate medication use. On the contrary, research has also shown that patients with multiple comorbidities are more likely to fail to take their medications as directed because their treatment plans are more complex. In the long term, this can have serious consequences [105]. According to our findings, the types of medications used and the frequency with which they are used can affect MA. Evidence suggests that those who take only one or once-daily hypertension medicine perform better on the MA (Kamran et al., 2014). This could imply that it is likely simpler to remember taking medication prescribed to be taken once a day than medication prescribed to be taken multiple times per day. Adherence is also influenced by education, as individuals with low levels of education frequently struggle to understand how to take their prescriptions correctly [106]. The patient's level of education is linked to their general medication knowledge, which has been widely cited as a major predictor of MA across various NCDs and research articles. This shows that patients with more knowledge may have a greater awareness of themselves, increasing the likelihood that they will engage in self-care.

A number of approaches to FS and methods for establishing the relevance or importance of a feature in modelling MA or non-adherence were found. These are mechanisms that are utilised to reduce the number of features present in the feature vectors applied in ML-based MA prediction research. Reported are multivariate, univariate, and bivariate analysis; proportional hazard modelling, ANOVA; t-tests; the Wilcoxon signed-rank test; the Chi-square test; Fisher's exact test; the Mann–Whitney test; the Bonferroni-adjusted chi-square test; and correlation analysis such as the Pearson correlation. As a general rule, the threshold for statistical significance is usually set at a p-value that is lower than 0.05. It is critical to note that prior to creating the ML models for predicting or classifying MA amongst NCD patients, FS and screening could be carried out using the traditional statistical methods of multivariate, univariate, and bivariate analysis. Thus, statistical analytic methods are useful in ML modelling of MA as precursors to help identify significant predictors of MA. The ML will ultimately be built based on a reduced feature set with improved model accuracy. This aligns with the findings of [21]; who reveals that the strong performance of ML models may be credited to the FS technique, which combines an ML algorithm and a univariate analysis to build a collection of eight-dimensional vectors with decreased dimensionality. In order to discover the combinations that offered the most accurate model classification, forward and backward elimination was used in the process. The SOM was applied to help visualise the relationship between the factors determining adherence levels. Since ML models are thought to be black boxes [107], demonstrated that SOM can be used to visualise higher-dimensional nonlinear data and feature distribution in the unified distance matrix map. This allows for a better understanding of the link between the factors influencing adherence levels. RFE, another form of backward selection method, helped remove the least important predictors in modelling MA. When the MSE is used, a significant variable has a bigger proportion of its MSE value growing than a variable not considered [108].

In the ML approaches, the cross-validated change in AUROC is computed and used to characterise the importance of baseline features in the classification of non-adherence. The screening of the variables reveals a decrease in predictive power when an important component is left out, and the results are analysed to see how the AUROC has been affected. Akaike's Information Criterion (AIC) is another FS criterion that can be used to estimate prediction error in a LASSO model, and feature importance is shown through ORs. The OOB error rate for ranking the importance of variables in ML-based analytics of MA is also commonly used. WEKA additionally assists in finding and eliminating interrelated features by utilising the information-gain (IG)-based FS technique. This is done so that the predicted features with the highest level of accuracy might be chosen. This technique aids in sifting through the myriad of possible ways in which the dataset features could be merged in order to identify the select few that are most relevant to the ML modelling task. Other filter methods and MI procedures were utilised in addition to the IG. The gain ratio, symmetrical uncertainty, relief, 1-R, and chi-squared FS were some of the analyses that were carried out. These methods ensure that the chosen features are not dependent on the learning process and assess the value of the retrieved attributes. The SHAP also proved to be a popular and reliable technique to analyse and rank the importance of the predictors of MA. What makes SHAP more insightful are the SHAP beeswarm and bar graph plots, which allow analysts to visualise an information-dense summary of how the top features in the datasets impact the model's output as well as ordering features by their impact on prediction, the first being the most important and the last being the least important one.

Our research also revealed that other variables were incorporated into the ML model either because they had previously been shown to affect MA in NCD patients in previous research or because a group of knowledgeable professionals deemed them therapeutic relevant. Others were thrown out because more than 50% of the values for the features were missing, and the data structure made it impossible to fill in the missing values. By and large, the value proposition of using ML to carry out FS and analyse feature importance is evident and cannot be doubted. Applying a combination of ML algorithms, multivariate, univariate, and bivariate analysis, and visualisation techniques such as SOM and SHAP demonstrates the suitability of these approaches for selecting significant variables, analysing and ranking feature importance, making predictions, and visualising the results. It has been shown that using ML-based FS and variable significance screening procedures to reduce the number of features makes it easier to build ML models for predicting or classifying MA in patients with NCD. In summary, the output of the FS could be a weighting of each feature in a dataset, a ranking of the dataset features based on how important they are, or a subset of the full candidate feature set chosen based on a subset evaluation strategy like the one used in wrapper methods, which involves choosing a feature subset with the best predictive power or grouping together features that have individually shown to be highly predictive.

8. Strengths and limitations

Given the paucity of research that has used ML techniques to analyse predictors of MA among NCD patients or conducted FS and feature importance screening for modelling MA, Consequent to this, this review detects, collates, and analyses significant implicit literature hidden in a fragmented piece of literature about FS and its importance in the development of ML models for predicting or classifying MA among NCD patients with conditions such as CVD, hypertension, diabetes, cancer, and chron's, among other diseases. This single study collates and articulates the various ML FS and features important screening techniques and metrics. The research has certain limitations. This review was limited to studies that only reported FS and the importance of leveraging ML. Thus, the sample size of the study is relatively small.

9. Conclusion

In conclusion, ML models help predict which patients are more or less likely to take their medicines as prescribed. Identifying features that significantly predict the target variable is a precursor to improved model accuracy in predicting or classifying MA among NCD patients. FS and feature importance ranking are thus indispensable in determining which features to include or exclude in model building. A combination of features ultimately decides the risk of non-adherence to medication. Adherence or non-adherence is a multifaceted issue that can influence various socio-demographic and economic patient characteristics, disease-related features, behavioural and condition-related factors, and healthcare system-related factors. MA or non-adherence is a complex problem, so it is critical to identify the important features in predicting

MA. It is possible to use ML techniques to figure out how these risk factors interact with one another, how relevant they are for predicting MA, and how they differ amongst risk groups with low MA. FS is increasingly important as the size and complexity of the average dataset continue to grow exponentially. Therefore, ML-based FS and feature importance screening are superior to traditional statistical and non-statistical methods when determining the most salient features to predict MA in patients with NCD. If researchers or analysts have a solid understanding of FS approaches, their models will perform better and have a greater understanding of the data's structure and properties. This systematic literature review study informs future research practitioners to undertake a quantitative empirical study incorporating ML-based FS methods shown in this current study to determine the impact of predictors on MA among NCD patients in the context of an African-based EHRs dataset. Since this review study revealed a lack of published work on applying ML to predict or classify MA amongst NCD patients, an African-based dataset must be the primary focus of future research interests.

Data availability

N/A.

Ethical approval

N/A.

Authors' contributions

All authors contributed equally to the writing of this manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

N/A.

References

[1] World Health Organisation. Noncommunicable diseases. https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/ncd-mortality#:~:text=Noncommunicable diseases (NCDs) kill 41,- and middle-income countries; 2022.
[2] Gu Y, Zalkikar A, Liu M, Kelly L, Hall A, Daly K, Ward T. Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data. Sci Rep 2021;11(1):18961.
[3] Rudemiller NP, Crowley SD. The role of chemokines in hypertension and consequent target organ damage. Pharmacol Res 2017;119:404–11.
[4] Schulz M, Krueger K, Schuessel K, Friedland K, Laufs U, Mueller WE, Ude M. Medication adherence and persistence according to different antihypertensive drug classes: a retrospective cohort study of 255,500 patients. Int J Cardiol 2016;220:668–76.
[5] Harding JL, Pavkov ME, Magliano DJ, Shaw JE, Gregg EW. Global trends in diabetes complications: a review of current evidence. Diabetologia 2019;62:3–16.
[6] Ismail LMN, Selim MA, El-Khashab SO. Factors affecting medication adherence among patients with rheumatic disorders. J Nurs Educ Pract 2017;7(8):7–18.
[7] Lester CA, Mott DA, Chui MA. The influence of a community pharmacy automatic prescription refill program on Medicare Part D adherence metrics. Journal of Managed Care & Specialty Pharmacy 2016;22(7):801–7.
[8] Hu F, Warren J, Exeter DJ. Predicting lipid-lowering medication persistence after the first cardiovascular disease hospitalization. Methods Inf Med 2020;59(02/03):61–74.
[9] Kronish IM, Ye S. Adherence to cardiovascular medications: lessons learned and future directions. Prog Cardiovasc Dis 2013;55(6):590–600.
[10] Haas K, Miled Z Ben, Mahoui M. Medication adherence prediction through online social forums: a case study of fibromyalgia. JMIR Medical Informatics 2019;7(2):e12561.
[11] Martin LR, Williams SL, Haskard KB, DiMatteo MR. The challenge of patient adherence. Therapeut Clin Risk Manag 2005;1(3):189–99.
[12] Wu X-W, Yang H-B, Yuan R, Long E-W, Tong R-S. Predictive models of medication non-adherence risks of patients with T2D based on multiple machine learning algorithms. BMJ Open Diabetes Research and Care 2020;8(1):e001055.
[13] Cea-Calvo L, Marín-Jiménez I, de Toro J, Fuster-RuizdeApodaca MJ, Fernández G, Sánchez-Vega N, Orozco-Beltrán D. Different associations of intentional and non-intentional non-adherence behaviors with patient experience with healthcare and patient beliefs in medications: a survey of patients with chronic conditions. Patient Prefer Adherence 2020:2439–50.
[14] Lee H, Park JH, Floyd JS, Park S, Kim HC. Combined effect of income and medication adherence on mortality in newly treated hypertension: nationwide study of 16 million person-years. J Am Heart Assoc 2019;8(16):e013148.
[15] Zullig LL, Blalock DV, Dougherty S, Henderson R, Ha CC, Oakes MM, Bosworth HB. The new landscape of medication adherence improvement: where population health science meets precision medicine. Patient Prefer Adherence 2018:1225–30.
[16] Zakeri M, Sansgiry SS, Abughosh SM. Application of machine learning in predicting medication adherence of patients with cardiovascular diseases: a systematic review of the literature. 2020.
[17] Scarton L, Nelson T, Yao Y, Segal R, Donahoo WT, Goins RT, DeVaughan-Circles A, Manson SM, Wilkie DJ. Medication adherence and cardiometabolic control indicators among American Indian adults receiving tribal health services: protocol for a longitudinal electronic health records study. JMIR Research Protocols 2022;11(10):e39193.
[18] Koesmahargyo V, Abbas A, Zhang L, Guan L, Feng S, Yadav V, Galatzer-Levy IR. Accuracy of machine learning-based prediction of medication adherence in clinical research. Psychiatr Res 2020;294:113558.
[19] Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319(13):1317–8.
[20] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104–16.
[21] Wang L, Fan R, Zhang C, Hong L, Zhang T, Chen Y, Liu K, Wang Z, Zhong J. Applying machine learning models to predict medication nonadherence in Crohn's disease maintenance therapy. Patient Prefer Adherence 2020:917–26.
[22] Aziz F, Malek S, Ali AM, Wong MS, Mosleh M, Milow P. Determining hypertensive patients' beliefs towards medication and associations with medication adherence using machine learning methods. PeerJ 2020;8:e8286.
[23] Stafford IS, Gosink MM, Mossotto E, Ennis S, Hauben M. A systematic review of artificial intelligence and machine learning applications to inflammatory bowel disease, with practical guidelines for interpretation. Inflamm Bowel Dis 2022;28(10):1573–83.
[24] Babel A, Taneja R, Mondello Malvestiti F, Monaco A, Donde S. Artificial intelligence solutions to increase medication adherence in patients with non-communicable diseases. Frontiers in Digital Health 2021;69.
[25] Bohlmann A, Mostafa J, Kumar M. Machine learning and medication adherence: scoping review. JMIRx Med 2021;2(4):e26993.
[26] Bernardo WM, Nobre MR, Jatene FB. Evidence-based clinical practice. Part II–Searching evidence databases. Rev Assoc Méd Bras 2004;50(1):104–8. 1992.
[27] Li X, Xu H, Li M, Zhao D. Using machine learning models to study medication adherence in hypertensive patients based on national stroke screening data. In: 2021 IEEE 9th international conference on bioinformatics and computational biology (ICBCB); 2021. p. 135–9.
[28] Aldeer M, Javanmard M, Martin RP. A review of medication adherence monitoring technologies. Applied System Innovation 2018;1(2):14.
[29] Pirri S, Lorenzoni V, Turchetti G. Scoping review and bibliometric analysis of Big Data applications for Medication adherence: an explorative methodological study to enhance consistency in literature. BMC Health Serv Res 2020;20(1):1–23.
[30] Krousel-Wood M, Thomas S, Muntner P, Morisky D. Medication adherence: a key factor in achieving blood pressure control and good clinical outcomes in hypertensive patients. Curr Opin Cardiol 2004;19(4):357–62.
[31] Kleczyk EJ, Evans D. Leveraging predictive analytics to derive patient adherence drivers. Pharmaceutical Management Science Association; 2021.
[32] Anderson LJ, Nuckols TK, Coles C, Le MM, Schnipper JL, Shane R, Jackevicius C, Lee J, Pevnick JM, Pharm-Dc Group Choudhry Niteesh K Md, P. D. O. D. M. D. S. C. M. D. A systematic overview of systematic reviews evaluating medication adherence interventions. Am J Health Syst Pharm 2020;77(2):138–47.
[33] Gil-Guillen VF, Balsa A, Bernárdez B, Valdés y Llorca C, Márquez-Contreras E, de la Haba-Rodríguez J, Castellano JM, Gómez-Martínez J. Medication non-adherence in rheumatology, oncology and cardiology: a review of the literature of risk factors and potential interventions. Int J Environ Res Publ Health 2022;19(19):12036.
[34] Verbrugghe M, Verhaeghe S, Lauwaert K, Beeckman D, Van Hecke A. Determinants and associated factors influencing medication adherence and persistence to oral anticancer drugs: a systematic review. Cancer Treat Rev 2013;39(6):610–21.
[35] Chu H-Y, Huang H-C, Huang C-Y, Chu C-C, Su C-T, Tsai I-L, Hu H-LS, Guo S-L. A predictive model for identifying low medication adherence among older adults with hypertension: a classification and regression tree model. Geriatr Nurs 2021;42(6):1309–15.
[36] Comer S, Cunningham C, Fishman MJ, Gordon FA, Kampman FK, Langleben D, Nordstrom B, Oslin D, Woody G, Wright T. National practice guideline for the use of medications in the treatment of addiction involving opioid use. Am Soc Addicit Med 2015;66:39–42.
[37] Kassahun T, Gesesew H, Mwanri L, Eshetie T. Diabetes related knowledge, self-care behaviours and adherence to medications among diabetic patients in Southwest Ethiopia: a cross-sectional survey. BMC Endocr Disord 2016;16:1–11.

[38] Ronquest NA, Willson TM, Montejano LB, Nadipelli VR, Wollschlaeger BA. Relationship between buprenorphine adherence and relapse, health care utilization and costs in privately and publicly insured patients with opioid use disorder. Subst Abuse Rehabil 2018:59–78.

[39] Saloner B, Daubresse M, Alexander GC. Patterns of buprenorphine-naloxone treatment for opioid use disorder in a multi-state population. Med Care 2017;55 (7):669.

[40] Weidt F, Silva R. Systematic literature review in computer science-a practical guide. Relatórios Técnicos Do DCC/UFJF 2016;1(8).

[41] Hasan MM, Young GJ, Shi J, Mohite P, Young LD, Weiner SG. A machine learning based two-stage clinical decision support system for predicting patients' discontinuation from opioid use disorder treatment: retrospective observational study. BMC Med Inf Decis Making 2021;21(1):1–21.

[42] Nagaraj P, Deepalakshmi P, Ijaz MF. Chapter 9 - optimized adaptive tree seed Kalman filter for a diabetes recommendation system—bilevel performance improvement strategy for healthcare applications. In: Bhoi AK, de Albuquerque VHC, Srinivasu PN, G.B.T.-C, S.C.T. for the A. of H. D. Marques, editors. Intelligent data-centric systems. Academic Press; 2022. p. 191–202. https://doi.org/10.1016/B978-0-323-85751-2.00010-4.

[43] Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eD octor: machine learning and the future of medicine. J Intern Med 2018;284(6):603–19.

[44] Al'Aref SJ, Maliakal G, Singh G, van Rosendael AR, Ma X, Xu Z, Alawamlh OAH, Lee B, Pandey M, Achenbach S. Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the CONFIRM registry. Eur Heart J 2020;41(3):359–67.

[45] Emanuel EJ, Wachter RM. Artificial intelligence in health care: will the value match the hype? JAMA 2019;321(23):2281–2.

[46] Daghistani T, Alshammari R. Comparison of statistical logistic regression and random forest machine learning techniques in predicting diabetes. J Adv Inf Technol 2020;11(2):78–83.

[47] Karanasiou GS, Tripoliti EE, Papadopoulos TG, Kalatzis FG, Goletsis Y, Naka KK, Bechlioulis A, Errachid A, Fotiadis DI. Predicting adherence of patients with HF through machine learning techniques. Healthcare Technology Letters 2016;3(3): 165–70.

[48] Li Y, Shan B, Li B, Liu X, Pu Y. Literature review on the applications of machine learning and blockchain technology in smart healthcare industry: a bibliometric analysis. J Healthc Eng 2021;2021:11. https://doi.org/10.1155/2021/9739219. 9739219.

[49] Mohite P. Leveraging machine learning framework to predict buprenorphine/ naloxone treatment discontinuation. Northeastern University; 2020.

[50] Tucker CS, Behoora I, Nembhard HB, Lewis M, Sterling NW, Huang X. Machine learning classification of medication adherence in patients with movement disorders using non-wearable sensors. Comput Biol Med 2015;66:120–34.

[51] Zhu X, Peng B, Yi Q, Liu J, Yan J. Prediction model of immunosuppressive medication non-adherence for renal transplant patients based on machine learning technology. Front Med 2022;333.

[52] Delpino FM, Costa ÂK, Farias SR, Chiavegatto Filho ADP, Arcêncio RA, Nunes BP. Machine learning for predicting chronic diseases: a systematic review. Publ Health 2022;205:14–25.

[53] Al-Aidaroos KM, Bakar AA, Othman Z. Naive Bayes variants in classification learning. In: 2010 international conference on information retrieval & knowledge management (CAMP); 2010. p. 276–81.

[54] IBM. What is a neural network?. https://www.ibm.com/topics/neural-networks; 2021.

[55] Malekian A, Chitsaz N. Concepts, procedures, and applications of artificial neural network models in streamflow forecasting. In: Advances in streamflow forecasting. Elsevier; 2021. p. 115–47.

[56] Ghosh-Dastidar S, Adeli H. Spiking neural networks. Int J Neural Syst 2009;19(4): 295–308.

[57] Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinf 2006;7:1–16.

[58] Blessie EC, Karthikeyan E. Sigmis: a feature selection algorithm using correlation based method. J Algorithm Comput Technol 2012;6(3):385–94.

[59] Bouchlaghem Y, Akhiat Y, Amjad S. Feature selection: a review and comparative study. E3S Web of Conferences 2022;351:1046.

[60] Michel P, Ngo N, Pons J-F, Delliaux S, Giorgi R. A filter approach for feature selection in classification: application to automatic atrial fibrillation detection in electrocardiogram recordings. BMC Med Inf Decis Making 2021;21(4):1–17.

[61] Sánchez-Maroño N, Alonso-Betanzos A, Tombilla-Sanromán M. Filter methods for feature selection–a comparative study. Lect Notes Comput Sci 2007;4881:178–87.

[62] Lee SK, Kang B-Y, Kim H-G, Son Y-J. Predictors of medication adherence in elderly patients with chronic diseases using support vector machine models. Healthcare Informatics Research 2013;19(1):33–41.

[63] Seliem MM. HandlingOutlier data as missing values by imputation methods: application of machine learning algorithms. Turkish Journal of Computer and Mathematics Education (TURCOMAT) 2022;13(1):273–86.

[64] Hopf K, Reifenrath S. Filter methods for feature selection in supervised machine learning applications–review and benchmark. 2021. *ArXiv Preprint ArXiv: 2111.12140*.

[65] Akhiat Y, Chahhou M, Zinedine A. Feature selection based on graph representation. In: 2018 IEEE 5th international congress on information science and technology (CiSt); 2018. p. 232–7.

[66] Akhiat Y, Chahhou M, Zinedine A. Ensemble feature selection algorithm. Int J Intell Syst Appl 2019;11(1):24.

[67] Asnaoui Y, Akhiat Y, Zinedine A. Feature selection based on attributes clustering. In: 2021 fifth international conference on intelligent computing in data sciences (ICDS); 2021. p. 1–5.

[68] Kumari B, Swarnkar T. Filter versus wrapper feature subset selection in large dimensionality micro array: a review. Int J Comput Sci Inf Technol 2011;2(3): 1048–53.

[69] Gheyas IA, Smith LS. Feature subset selection in large dimensionality domains. Pattern Recogn 2010;43(1):5–13.

[70] Rodríguez-Pérez R, Bajorath J. Feature importance correlation from machine learning indicates functional relationships between proteins and similar compound binding characteristics. Sci Rep 2021;11(1):14245.

[71] Sandilands DD. Bivariate analysis. Encyclopedia of Quality of Life and Well-Being Research; 2014. p. 416–8.

[72] IBM. A glimpse inside the mind of a data scientist: must-have skills? Daily challenges? Find out what actual data scientists really think about their critical role in data science. https://www.ibm.com/downloads/cas/W6GEX9LL; 2016.

[73] McDonald, J. (n.d.). Fisher's Exact Test. https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Biological_Statistics_(McDonald)/02%3A_Tests_for_Nominal_Variables/2.07%3A_Fisher's_Exact_Test.

[74] Venkatesh B, Anuradha J. A hybrid feature selection approach for handling a high-dimensional data. In: Innovations in computer science and engineering: proceedings of the sixth ICICSE 2018; 2019. p. 365–73.

[75] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[76] Njoku UF, Abelló Gamazo A, Bilalli B, Bontempi G. Impact of filter feature selection on classification: an empirical study. Proceedings of the 24rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP): Co-Located with the 24th International Conference on Extending Database Technology and the 24th International Conference on Da 2022:71–80.

[77] Zullig LL, Jazowski SA, Wang TY, Hellkamp A, Wojdyla D, Thomas L, Egbuonu-Davis L, Beal A, Bosworth HB. Novel application of approaches to predicting medication adherence using medical claims data. Health Serv Res 2019;54(6): 1255–62.

[78] Galozy A, Nowaczyk S. Prediction and pattern analysis of medication refill adherence through electronic health records and dispensation data. J Biomed Inf 2020;112:100075.

[79] Gao W, Liu H, Ge C, Liu X, Jia H, Wu H, Peng X. A clinical prediction model of medication adherence in hypertensive patients in a Chinese community hospital in Beijing. Am J Hypertens 2020;33(11):1038–46.

[80] Li Y, Jasani F, Su D, Zhang D, Shi L, Yi SS, Pagán JA. Decoding nonadherence to hypertensive medication in New York City: a population segmentation approach. Journal of Primary Care & Community Health 2019;10:2150132719829311.

[81] Mirzadeh SI, Arefeen A, Ardo J, Fallahzadeh R, Minor B, Lee J-A, Hildebrand JA, Cook D, Ghasemzadeh H, Evangelista LS. Use of machine learning to predict medication adherence in individuals at risk for atherosclerotic cardiovascular disease. Smart Health 2022;26:100328.

[82] Lucas JE, Bazemore TC, Alo C, Monahan PB, Voora D. An electronic health record based model predicts statin adherence, LDL cholesterol, and cardiovascular disease in the United States Military Health System. PLoS One 2017;12(11): e0187809.

[83] Son Y-J, Kim H-G, Kim E-H, Choi S, Lee S-K. Application of support vector machine for prediction of medication adherence in heart failure patients. Healthcare Informatics Research 2010;16(4):253–9.

[84] Fan Y, Long E, Cai L, Cao Q, Wu X, Tong R. Machine learning approaches to predict risks of diabetic complications and poor glycemic control in nonadherent type 2 diabetes. Front Pharmacol 2021:1485.

[85] Lo-Ciganic W-H, Donohue JM, Thorpe JM, Perera S, Thorpe CT, Marcum ZA, Gellad WF. Using machine learning to examine medication adherence thresholds and risk of hospitalization. Med Care 2015;53(8):720.

[86] Gottlieb A, Yatsco A, Bakos-Block C, Langabeer JR, Champagne-Langabeer T. Machine learning for predicting risk of early dropout in a recovery program for opioid use disorder. Healthcare 2022;10(2):223.

[87] Yerrapragada G, Siadimas A, Babaeian A, Sharma V, O'Neill TJ. Machine learning to predict tamoxifen nonadherence among US commercially insured patients with metastatic breast cancer. JCO Clinical Cancer Informatics 2021;5:814–25.

[88] Kim N, McCarthy DE, Loh W-Y, Cook JW, Piper ME, Schlam TR, Baker TB. Predictors of adherence to nicotine replacement therapy: machine learning evidence that perceived need predicts medication use. Drug Alcohol Depend 2019;205:107668.

[89] Gatt I, West LM, Calleja N, Briffa C, Cordina M. Psychometric properties of the belief about medicines questionnaire (BMQ) in the Maltese language. Pharm Pract 2017;15(1).

[90] Sunday Julius M, Rita Alo U, Uchenna Onu F, Ihuoma Akobundu C. Machine learning framework to predict patient non-adherence to medication using non-clinical data: a prognosis approach. Proceedings of the 9th International Conference on Computer and Communications Management 2021:98–103.

[91] Wang D, Li J, Sun Y, Ding X, Zhang X, Liu S, Han B, Wang H, Duan X, Sun T. A machine learning model for accurate prediction of sepsis in ICU patients. Front Public Health 2021;9:754348.

[92] Ciesielski T, Iyengar R, Bothra A, Tomala D, Cislo G, Gage BF. A tool to assess risk of de novo opioid abuse or dependence. Am J Med 2016;129(7):699–705.

[93] Pear VA, Ponicki WR, Gaidus A, Keyes KM, Martins SS, Fink DS, Rivera-Aguirre A, Gruenewald PJ, Cerdá M. Urban-rural variation in the socioeconomic determinants of opioid overdose. Drug Alcohol Depend 2019;195:66–73.

[94] Samples H, Williams AR, Olfson M, Crystal S. Risk factors for discontinuation of buprenorphine treatment for opioid use disorders in a multi-state sample of Medicaid enrollees. J Subst Abuse Treat 2018;95:9–17.
[95] Edmond MB, Aletraris L, Roman PM. Rural substance use treatment centers in the United States: an assessment of treatment quality by location. Am J Drug Alcohol Abuse 2015;41(5):449–57.
[96] Pullen E, Oser C. Barriers to substance abuse treatment in rural and urban communities: counselor perspectives. Subst Use Misuse 2014;49(7):891–901.
[97] Belaiche S, Décaudin B, Dharancy S, Noel C, Odou P, Hazzan M. Factors relevant to medication non-adherence in kidney transplant: a systematic review. Int J Clin Pharm 2017;39:582–93.
[98] Brorson HH, Arnevik EA, Rand-Hendriksen K, Duckert F. Drop-out from addiction treatment: a systematic review of risk factors. Clin Psychol Rev 2013;33(8):1010–24.
[99] Gokoel SRM, Gombert-Handoko KB, Zwart TC, van der Boog PJM, Moes DJAR, de Fijter JW. Medication non-adherence after kidney transplantation: a critical appraisal and systematic review. Transplant Rev 2020;34(1):100511.
[100] Lappan SN, Brown AW, Hendricks PS. Dropout rates of in-person psychosocial substance use disorder treatments: a systematic review and meta-analysis. Addiction 2020;115(2):201–17.
[101] Morisky DE, Ang A, Krousel-Wood M, Ward HJ. Predictive validity of a medication adherence measure in an outpatient setting. J Clin Hypertens 2008;10(5):348–54.
[102] Chisholm-Burns MA, Spivey CA, Wilks SE. Social support and immunosuppressant therapy adherence among adult renal transplant recipients. Clin Transplant 2010;24(3):312–20.
[103] Ladin K, Daniels A, Osani M, Bannuru RR. Is social support associated with post-transplant medication adherence and outcomes? A systematic review and meta-analysis. Transplant Rev 2018;32(1):16–28.
[104] Zhang Y, Li X, Mao L, Zhang M, Li K, Zheng Y, Cui W, Yin H, He Y, Jing M. Factors affecting medication adherence in community-managed patients with hypertension based on the principal component analysis: evidence from Xinjiang, China. Patient Prefer Adherence 2018:803–12.
[105] Ghembaza MA, Senoussaoui Y, Kendouci Tani M, Meguenni K. Impact of patient knowledge of hypertension complications on adherence to antihypertensive therapy. Curr Hypertens Rev 2014;10(1):41–8.
[106] Oosterom-Calo R, Van Ballegooijen AJ, Terwee CB, Te Velde SJ, Brouwer IA, Jaarsma T, Brug J. Determinants of adherence to heart failure medication: a systematic literature review. Heart Fail Rev 2013;18:409–27.
[107] Kihato PK. Analysis and visualization of metabolic syndrome using self organizing maps (SOM). 2013.
[108] Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. Pattern Recogn Lett 2010;31(14):2225–36.

3.2 Chapter Summary

The article making up Chapter three (3) provided an overview of the empirical research findings in existing studies, specifically highlighting the various feature selection techniques features that have a significant impact on ML-based classification and prediction of MA. These features include sociodemographic and economic characteristics, past adherence behaviour, medication information, behavioural, disease-related, and healthcare system-related variables. The chapter additionally discusses the practical implications of these findings for healthcare practitioners, highlighting how recognising the significance of various features can guide focused treatments and enhance adherence rates. In addition, the chapter examines the interpretability and reliability of the selected features, guaranteeing that the model's results are not only precise but also robust. This feature selection review chapter directly informs the generation of a dataset useful for measuring MA in the forthcoming segment - chapter 4. By and large, this foregoing chapter establishes a solid foundation for the development of datasets and ML-based classification models in later chapters by providing an in-depth awareness of the feature selection techniques and the most important features that impact MA among patients with NCDs.

CHAPTER 4

Data wrangling and generation for ML models in medication adherence analytics (Paper 3)

4.1 Introduction

The fourth chapter delves into the pragmatic aspects of preparing a dataset useful for building robust ML models that classify MA in patients with NCDs. In addition to publishing a patient-level and medical claims dataset for MA analytics, the chapter carefully describes the methodically arrayed procedures and techniques used in data wrangling (DW). Thus, a pragmatic approach to developing and generating the dataset was taken by including a logical set of essential DW components. The chapter begins by emphasising the importance of DW, which is a critical procedure involving the cleansing, data conversion, profiling, and organisation of unprocessed data into a suitable format for analysis. The main data sources for this study consist of patient-level data and medical claims data acquired from a Zimbabwean medical aid services provider, encompassing the entirety of the year 2022. This dataset comprises pharmacy refill records, patient-level data, and medical claims data, offering a comprehensive perspective on patient adherence behaviour and associated information. This dataset, which contains a variety of different types of data, offers a valuable basis for conducting extensive MA analytics. The steps involved in data importation and first inspection are outlined in depth, with a focus on ensuring that all data is successfully loaded and initially checked for consistency and completeness. The subsequent section provides a detailed explanation of the data pre-processing procedures, which encompass the management of missing values, rectification of inconsistencies, and elimination of duplicates in order to guarantee the reliability of the data.

The chapter explores data enrichment strategies, which involve integrating supplementary data from external sources to enhance the understanding of patient adherence behaviour and related data. As a result, another critical aspect of this chapter is the examination of feature engineering, which involves the creation of new features from the original data to enhance the ML models' prediction capabilities. Some examples of this include

computing the rates or thresholds for medicine refills, analysing patterns in the timing of refills, and producing overall measurements of healthcare usage. These specifically designed features capture the subtle details of how the patients follow their prescription regimen and offer extra information for the models. The process involved enhancing the current dataset by adding new features and expanding the original raw dataset with data acquired from external sources that are related to the existing data points. This also involves gathering additional data regarding the current data points that are already stored, as well as regrouping the column values of the existing variables in the dataset. The chapter continues by discussing data integration techniques, with a specific emphasis on combining different data sources while ensuring the accuracy and consistency of the data. Data profiling was conducted to get insight into the statistical characteristics of the dataset, detecting any possible concerns that require attention prior to model construction. Data preprocessing procedures, such as rescaling numerical data and encoding categorical variables, were utilised to prepare the data for ML algorithms. A dimensionality reduction technique, specifically the RF feature importance scoring mechanism, was employed to determine the importance of each feature, hence improving the performance of the model.

The chapter also addresses the difficulties faced during data wrangling, including handling various data formats, guaranteeing data privacy and security, and managing extensive datasets. The text provides a comprehensive explanation of methods used to protect the privacy of sensitive patient data and ensure adherence to ethical guidelines. The chapter ends with an analysis of the significance of data quality and its influence on the efficacy of ML models. The importance of technical validation, which involves a complete analysis of the generated dataset to guarantee the accuracy and quality of the data. Accordingly, at the end of the data wrangling process, the dataset's quality was verified by evaluating the performance of an RF model using the generated dataset. The prepared dataset, which is available at
https://data.mendeley.com/datasets/zkp7sbbx64/2, accompanies this published article.

Paper status: In press

Journal Name: Data Intelligence

Journal impact factor: 1.3 (Q1)

Student's contribution: Conceptualisation, planning, methodology, coding, data curation, validation and original manuscript writing

Other author's contributions: Data acquisition, supervision, editing, and revision

Percentage of student's contribution: 90%

DATA PAPER

Data Wrangling and Generation for Machine Learning Models in Medication Adherence Analytics: A practical Standpoint using Patient-Level and Medical Claims Data

Wellington Kanyongo[1,2], Absalom El-Shamir Ezugwu[3†], Tsitsi Moyo[4], Jean Vincent Fonou Dombeu[5]

[1]School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg Campus, Pietermaritzburg, 3201, South Africa

[2]Department of Computer Science, Faculty of Science Engineering, Bindura University of Science Education, Bindura, Zimbabwe

[3]Unit for Data Science and Computing, North-West University, 11 Hoffman Street, Potchefstroom, 2520, South Africa

[4]Cimas Medical Aid Society, Borrowdale Office Park, Borrowdale Road, P. O Box 1243, Harare, Zimbabwe

[5]School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg Campus, Pietermaritzburg, 3201, South Africa

Keywords: Dataset; Dataset generation; Diabetes; Hypertension; Non-communicable disease (NCD), Patient(s)

Citation: Kanyongo W., Ezugwu A.E.S., Moyo T., et al.: Data Wrangling and Generation for Machine Learning Models in Medication Adherence Analytics: A practical Standpoint using Patient-Level and Medical Claims Data. Data Intelligence XX(XX), XX-XX (2024). doi: https://doi.org/10.1162/dint_a_XXXX

Submitted: March 24, 2024; Revised: April 15, 2024; Accepted: May 7, 2024

ABSTRACT

Even though data wrangling (DW) accounts for more than half of the machine learning (ML) process, there is a dearth of research on data wrangling and dataset preparation. Consequently, we present a valuable state-of-the-art dataset prepared based on medical and pharmaceutical claims, as well as patient-level data collected from the transactional processing system of a Zimbabwean health insurance company. We were given two initial raw sub-datasets (one for the diabetics and one for the hypertensive patients) that had been retrieved and divided based on the International Classification of Diseases 10th Revision Code (ICD-10). Since adherence metrics measure the percentage of patients covered by prescription claims for the same medication in the same therapeutic class, within the measurement year, the created dataset only comprised data for 2022, which spanned from January 1 until December 31, 2022. We acknowledge, however, that there is no universally approved compendium of DW stages; yet certain building blocks and functionalities with considerable overlaps are commonly accepted and can be considered typical of DW. In light of this, we take a pragmatic approach to generating a dataset,

† Corresponding author: Absalom El-Shamir Ezugwu (E-mail: Absalom.ezugwu@nwu.ac.za; ORCID:).

© 2024 Chinese Academy of Sciences. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

incorporating nine essential DW building blocks and tasks, employed iteratively and incrementally, akin to an agile approach to dataset generation. Thus, we used some of the DW tasks iteratively and incrementally throughout the DW process without restricting them to a single stage. Following a rigorous DW procedure, the created dataset was then used to build an ML model, which then demonstrates a generally good degree of performance, as evidenced by the 81% accuracy. The outcomes, techniques, and insights proffered in this article inform future researchers, data scientists, and analysts on how to conduct data mining, dataset enrichment, and DW. The created dataset can stimulate further study in areas such as evaluating medication adherence (MA) and performing ML tasks such as classifying, predicting, and clustering MA behaviours.

Specification table.

Subject	Computer Science & Health Informatics
Specific subject area	Medication adherence analytics or evaluation, machine learning, non-communicable diseases
Type of Data	Tabulated Microsoft Excel file (dataset with numerical and categorical data)
How the data were acquired	Anonymised raw datasets of medical claims and patient-level data of diabetes and hypertension patients from January 1, 2022, to December 31, 2022, were collected from Cimas Medical Aid Society, a Zimbabwean medical insurance company that offers a variety of medical aid packages. We acquired CSV data files extracted from Cimas' transaction processing system via its ICT department.
Data format	Comma-Separated Values (CSV)
Description of data collection	Previously, no dataset encompassing all the variables under consideration was generated. We obtained data from Cimas, which encompasses the timeframe of January 1, 2022, to December 31, 2022. We targeted patients in Cimas's transactional processing system (TPS), which collects and manages member medical claims and patient-level data. Cimas' TPS has big medical and pharmacy claim data. The datasets included individuals aged 16–101 with essential diabetes (ICD-10 Code: E10-) and hypertension (ICD-10 Code: i10-) who were covered by medical aid. We obtained two Python-importable CSV files for patients with diabetes and hypertension. Cimas provided additional data (data on diabetes and hypertension complications and comorbidity status) that were linked to the first two datasets. We added medical aid schemes and hospital cover types subscribed to by health insurance members from Cimas' brochures and websites to the initial two datasets.
Data source location	Institution: Cimas City: Harare Country: Zimbabwe
Data accessibility	Repository name: Mendeley Data Data identification number (DOI number): 10.17632/zkp7sbbx64.2 All data can be accessed at the following link: https://data.mendeley.com/datasets/zkp7sbbx64/2

incorporating nine essential DW building blocks and tasks, employed iteratively and incrementally, akin to an agile approach to dataset generation. Thus, we used some of the DW tasks iteratively and incrementally throughout the DW process without restricting them to a single stage. Following a rigorous DW procedure, the created dataset was then used to build an ML model, which then demonstrates a generally good degree of performance, as evidenced by the 81% accuracy. The outcomes, techniques, and insights proffered in this article inform future researchers, data scientists, and analysts on how to conduct data mining, dataset enrichment, and DW. The created dataset can stimulate further study in areas such as evaluating medication adherence (MA) and performing ML tasks such as classifying, predicting, and clustering MA behaviours.

Specification table.

Subject	Computer Science & Health Informatics
Specific subject area	Medication adherence analytics or evaluation, machine learning, non-communicable diseases
Type of Data	Tabulated Microsoft Excel file (dataset with numerical and categorical data)
How the data were acquired	Anonymised raw datasets of medical claims and patient-level data of diabetes and hypertension patients from January 1, 2022, to December 31, 2022, were collected from Cimas Medical Aid Society, a Zimbabwean medical insurance company that offers a variety of medical aid packages. We acquired CSV data files extracted from Cimas' transaction processing system via its ICT department.
Data format	Comma-Separated Values (CSV)
Description of data collection	Previously, no dataset encompassing all the variables under consideration was generated. We obtained data from Cimas, which encompasses the timeframe of January 1, 2022, to December 31, 2022. We targeted patients in Cimas's transactional processing system (TPS), which collects and manages member medical claims and patient-level data. Cimas' TPS has big medical and pharmacy claim data. The datasets included individuals aged 16–101 with essential diabetes (ICD-10 Code: E10-) and hypertension (ICD-10 Code: i10-) who were covered by medical aid. We obtained two Python-importable CSV files for patients with diabetes and hypertension. Cimas provided additional data (data on diabetes and hypertension complications and comorbidity status) that were linked to the first two datasets. We added medical aid schemes and hospital cover types subscribed to by health insurance members from Cimas' brochures and websites to the initial two datasets.
Data source location	Institution: Cimas City: Harare Country: Zimbabwe
Data accessibility	Repository name: Mendeley Data Data identification number (DOI number): 10.17632/zkp7sbbx64.2 All data can be accessed at the following link: https://data.mendeley.com/datasets/zkp7sbbx64/2

1. INTRODUCTION

Real-world data is unstructured and complex, making analysis challenging without data wrangling (DW). If DW is not given the attention it needs, it will become a major bottleneck for executing advanced data analytics, particularly in big data and machine learning (ML) modelling. Essentially, the ML model results will be good as far as the data quality is good [1]. In fact, DW is usually the first step as researchers and data analysts wrangle the data into an analytics-ready form [2]. This is because ML models require data to be in a specific form; thus, raw datasets often cannot be used. ML systems. It is worth noting that a high-quality dataset is the product of rigorous DW processes. First used in 2011, the term "data wrangling" describes a series of steps involved in transforming raw data into an asset suitable for analysis [3]. DW is the laborious process of transforming, reformatting, and combining data to make it more palatable for various downstream applications, such as visualization and analysis [4]. In our paper, DW refers to the iterative procedure of collecting, integrating, manipulating, enriching, and merging raw data. Data analysts can spend up to 80% of their time on DW and dataset preparation, according to [5]; however, this can impede them from undertaking more effective knowledge discovery tasks. This resonates with [6], who submit that data scientists spend 50–80% of their valuable time on the mundane task of collecting and preparing structured or unstructured data before it can be explored for useful analysis. Thus, it is often agreed that DW is the most tedious and time-consuming phase for data analysts [7]. Data analysts often use data sets collected for various purposes. Indeed, new insights are often obtained by merging data sets that were produced independently of each other, for example, by integrating data from outside an organisation with internal data resources. As a result, through the process of DW, there is a need to discover data sources, clean, integrate multiple data sets, eliminate duplications, restructure data and transform fields into a form that is suitable for an intended analysis. According to [1] there can be many activities for DW, such as data transformation, data cleaning, data integration, data optimisation and data conversion, which are used to convert the raw data to quality data.

Our article on data generation supports the call made in [8] to increase the availability of medical research data. Similarly, [9] admit the dearth of research on dataset preparation, even though data generation is critical for ML modelling. According to [10], African institutions have limited technical capacity to extract, document, archive, and share data, as well as the methodological capacity required to interpret complex data, putting them at a disadvantage. In that respect, the absence of harmonised data sets and information systems is a major obstacle to the successful management of public health emergencies [11]. To resolve this, long-term investments and research will be required. The limited availability of datasets for evaluating medication adherence (MA) among NCD patients, and more so, in the context of Africa and a developing country, Zimbabwe, motivated this research undertaking. The article generates a state-of-the-art dataset that is useful for examining MA among patients with diabetes and hypertension in a developing country. Cimas Medical Aid Society, a corporation offering medical insurance coverage and healthcare services, provided the data used for dataset generation in this article. Cimas health insurance products cover insured individuals' medical and surgical expenses. Our efforts cumulatively aid in addressing the scourge of medication non-adherence, which is reported to be at least 50% by the World Health Organisation. The value of this current data descriptor and the resultant dataset are articulated in Section 1.1.

1.1 VALUE OF THE DATA

- This article generates a state-of-the-art dataset that is valuable for examining MA among patients with diabetes and hypertension in the setting of a developing nation. The data can be used to assess variations in MA levels among non-communicable disease (NCD) patients based on individual characteristics. This then alerts the class of patients who have the corresponding traits and need specific care to combat the epidemic of medication non-adherence.
- Researchers, data scientists, and analysts interested in data preparation, mining, and enrichment can find valuable resources in this article. This, in turn, can lead to the creation of additional datasets for assessing MA and conducting research on NCD, which currently have limited datasets available for MA evaluation.
- Researchers can build, train, test, evaluate, and distinguish various ML models for MA assessment using the generated dataset. This information can then be used to inform the creation of artificial intelligence-based MA evaluation applications that are long-lasting.
- This paper serves as a data descriptor, providing a full explanation of the generated MA dataset. Thus, this data descriptor and the generated dataset have the potential to spur further research encompassing the ML tasks of classifying, predicting, and clustering MA behaviours.
- We have meticulously a comprehensive set of organised processes and techniques that we used to prepare the dataset. As a result, this gives a road map for future research activity in terms of dataset production. Thus, researchers, data scientists, and analysts interested in conducting DW, data mining, and enrichment can find valuable resources in this article.

2. EXISTING DATASETS AND METHODS

This section discloses the various published datasets and data articles on MA among NCD patients. Researchers created a variety of datasets and data articles using MA-related data obtained from questionnaires, electronic health records, and real-time electronic medication measurement data. [12] collected and published a dataset to evaluate MA, among other research tasks. The data were gathered via a cross-sectional self-administered survey of outpatients with chronic illnesses receiving medication at Alex Ekwueme Federal Teaching Hospital, Abakaliki (AE-FUTHA), Nigeria. Thus, the dataset was created using a medication non-adherence assessment questionnaire prepared in Microsoft Word and distributed to the chronic disease outpatients. The questionnaire was used to collect demographic data as well as data on the number of prescriptions prescribed, the quantity of tablets administered each day, the times of day when medication is taken, and the reasons for taking medications at the specified times. Their dataset, organised in Microsoft Excel, was intended to aid in the development of MA intervention systems, as well as the design of intelligent non-adherence prediction models.

[13] developed a dataset based on data collected via a cross-sectional survey questionnaire from outpatients with non- communicable chronic diseases at three tertiary healthcare facilities in Nigeria's southeast. This multidimensional patient dataset is used as a tool for online evaluation and prediction of medication non-adherence, with an intervention system and ML. The dataset also contains demographic

data, prescription regimens, a phone call with SMS to give persuasive adherence intervention messages, and explanations for missing pills. A questionnaire was developed and validated for reliability, with factors influencing medication non-adherence based on the World Health Organisation and other literature. On a 5-Likert scale, categorisation variables such as belief level, adherence level, behavioural pattern, and knowledge level could be converted into fuzzy variables. Leveraging patients' multidimensional datasets improves their MA by providing in-depth insights into why patients do not take their prescribed medications, allowing for the easy identification of clusters of patients who are likely to be non-adherent across the domain of chronic diseases, and supporting clinical decision-making with patient monitoring and management.

[14] created a dataset for detecting nonadherence clusters in neurological patients. The data were gathered from 429 patients who visited the outpatient clinic or stayed in the neurological ward at Jena University Hospital in Germany. The dataset comprises demographic information such as medical history, the medication the patient takes on a daily basis, and the person in charge of the patient's daily medications. Adherence was measured using the German Stendal Adherence to Medication Score (SAMS), an adaptation of the validated German Essen Compliance Score. The SAMS questionnaire contains 18 questions that constitute a cumulative scale (0–72), with 0 indicating complete adherence and 72 indicating complete non-adherence. Patients who take less than 80% of their prescription medication are classified as non-adherent, with (SAMS = 0) signifying fully adherent, moderately non-adherent (SAMS 1–10), or non-adherent (SAMS > 10). The various types and clusters of nonadherences using the German Stendal Adherence to Medication Score (SAMS) were identified.

[15] created a dataset entitled "Working data for adherence to anti-hypertensive therapy." The published dataset was submitted to an online repository (Mendeley Data) and saved as an SPSS file with the *.sav* extension. The SAV file is a data format created by the Statistical Package for the Social Sciences (SPSS), a statistical analysis tool used among scholars, health researchers, and data miners. The SAV is saved in a proprietary binary format and contains both a dataset and a dictionary representing the dataset. It saves data as rows and columns. Similarly, another study created and published an SPSS- or SAV-formatted dataset on antihypertensive MA in severely hypertensive patients after referral to secondary care [16]. That paper collected data on severely hypertensive individuals at the beginning and end of 12 months of access to a secondary healthcare centre in a southern Brazilian city. They employed the Morisky Green Levine Scale (MGLS) to assess non-adherence to antihypertensive medications. The Morisky Green Levine (MGL) adherence scale is a 4-item self-reported assessment of medication non-adherence among patients with chronic illnesses [17]. Demographic and clinical information was gathered at the start of the secondary health care programme. Adherence status, blood pressure, and antihypertensive medicine therapy were evaluated at the start and after a 12-month follow-up.

[18] published a dataset titled "MA among individuals with cardiovascular disease: A development and validation of a multivariable predictive model - Dataset." The dataset comprises information on individuals' sociodemographic factors, clinical and medical data, MA, and health behaviour data. They gathered the data from consecutive patients with cardiovascular disease who visited an outpatient clinic at Debretabor

General Hospital in Ethiopia. The data is used to create multivariable prediction models and predictors for MA outcome indicators. [19] compiled and generated a dataset using data obtained from remote real-time assessments of medicine administration. They assessed the dataset's accuracy in dynamically predicting non-adherence. The researchers conducted a study involving a substantial number of clinical trials (n = 4,182), where participants were monitored using a smartphone application that recorded videos of them taking their prescribed medication. Using ML-based classification models, the patient's primary diagnosis, demographics, and past indication of observed adherence or non-adherence were used to predict: (1) adherence rates ≥ 80% throughout the clinical trial; (2) adherence ≥ 80% for the subsequent week; and (3) adherence on the subsequent day.

[20] employed AdhereR, a package for the popular open-source statistical environment R that helps researchers compute EHD-based adherence estimates and visualise individual prescription histories and adherence patterns. This is a collection of two packages for the R open-source statistical environment that implement several methods of computing MA and medication history visualisations, both interactive (for data exploration) and high-quality (for publication). AdhereR allows for a flexible and comprehensive assessment of EHD-based MA using consumer-grade hardware and software. It determines persistence by identifying therapy sessions for each patient and computing the start, end, and length of each episode.

Smart Sharp Bins (SSB) have the potential to track MA. The Smart Sharps Bin is the world's first networked injection management solution. The Smart Sharps Bin includes sensors to track disposal and facilitates at-home injections by providing injection reminders, friendly interventions, and convenient sharps disposal. The SSB, designed and manufactured by HealthBeacon Ltd., is ideal for use in the home. It digitally records the disposal of used hypodermic needles, passively monitors injection disposal data, and uploads it to a cloud-based database for long-term storage and further analysis, allowing the support team to control drop disposals and provide follow-up services as needed [21]. [21] created ensemble learning and deep learning drug adherence models based on the SSB-generated data. They used an internet-connected "smart sharps bin" (SSB) to monitor adherence to injection protocols. The dataset encompasses data from approximately 8,000 SSB units and 342,174 historical injection records.

MA measurement data can be collected using smart pill bottles, which encourage patients to take their medications as prescribed. [22] used smart pill bottles to track medication adherence. Participants in the trial received electronic pill bottles as part of the intervention, and those assigned to the intervention arm received up to one SMS daily. Following that, the reinforcement learning (RL) prediction algorithm was utilised to tailor each message based on daily pill bottle adherence statistics. In MA assessment, computerised pillbox monitoring is preferable to pill counts and patient self-reports. Smart pill bottles, for example, have a mechanism that automatically tracks the time of medication intake, allowing for an objective assessment of tablets taken on time [23]. The smart pill bottles have various other functions that could promote MA, such as issuing a reminder at the scheduled dose time and sending messages to designated helpers if a person is more than an hour late [24].

MA estimation utilising electronic healthcare data (EHD) has been crucial to clinical practice and research in a variety of illnesses. Several authors employed electronic health information, such as

medical insurance claims data, to create datasets for analysing MA. [25] gathered data on Medicaid claims and encounter data for outpatient, inpatient, long-term care, and professional services, as well as prescription drug claims with details such as fill date, quantity dispensed, days of supply, and prescriber information; [26] retrieved data on all patients who were treated at the Fangzhuang Community Health Service Centre using the Intelligent Chronic Disease Management System; [27] collected EHR data from the MilitaryHealth System about patients who had been prescribed statin.

This article generates a unique dataset using electronic records for NCD patients in Zimbabwe, a developing country. It is actually the first state-of-the-art dataset for evaluating MA in Zimbabwean patients with diabetes and hypertension. In contrast to previous African data articles and published datasets [12, 13, 18], our current data article generates a dataset with over 20,000 unique rows, making it suitable for ML models. This is because large datasets are essential for ML algorithms, allowing them to find patterns and relationships in the data. The dataset we acquired for this study using electronic records is unique in the context of NCD patients in a developing country because it includes predictors that are absent from most MA assessment datasets. These variables include the occurrence of hypertension or diabetes-related complications among the NCD patients on medication, members' annual contributions to their medical aid coverage, the total number of claims made annually by patients with medical insurance, and the total amount of medical supplies or units given to a patient with an NCD for a variety of illnesses over a year. This data article not only publishes the dataset but also meticulously documents the systematically arrayed procedures and techniques used to prepare it. Thus, we adopt a pragmatic approach to creating and generating the dataset by incorporating a systematic array of critical DW building blocks.

3. DATA WRANGLING TASKS AND PIPELINES

While such tools may share certain common DW-supporting functionality, there is a lack of standards in terms of requirements, representations, and outputs in their operational offerings [28]. DW, also known as data preparation, cleaning, remediation, and munging [29], is the digital art of moulding and organising raw information objects into usable formats. Practitioners employ a variety of tools and procedures, both manual and automated, but approaches differ from project to project based on the setup, purpose, and features. In traditional data warehousing environments, DW has been carried out using Extract-Transform-Load platforms, with significant manual involvement in specifying, configuring, or tuning many of them [30]. Previous research suggests that DW activities, such as profiling, formatting, integration, transformation, cleaning, and outlier detection, are similar to the Extract-Transform-Load (ETL) process [31, 32]. Widely used ETL and DW approaches may involve similar subsets of processes and stages, yet they differ significantly. ETL, for example, focuses on operations that combine data from multiple sources into a unified schema and store it in a consolidated data repository, such as a data warehouse, whereas DW entails tailoring data transformations to prepare a dataset for an analytical task, such as ML prediction. This variation produces different patterns for handling data difficulties faced by analysts that are not typically seen in ETL workflows; yet the similarities between the two and their required outcomes are why the adaptation of ETL-based methodologies is critical.

This section focuses on the DW tasks or functionalities identified in existing literature. Table 1 summarises these numerous DW tasks, contributing to our overall compendium of the DW tasks used in this article to generate the MA evaluation dataset. Before preparing Table 1, it is worth noting that there is no universally approved collection of data preparation stages; nonetheless, certain functionalities are widely supported and can be regarded as typical of DW [2]. Thus, while DW systems vary in the tasks they support, their functionalities significantly overlap, making it easier to distinguish different categories of systems based on the specification and composition of these tasks rather than their specific capabilities.

Table 1. Data Wrangling tasks and functionalities in existing literature contributing to the present article.

Author(s)	Focus	Data wrangling tasks and or pipeline
[33]	Document the principles of DW, practical techniques for data preparation and a data funnel.	*Data moves through stages*: Raw stage (ingest data, describe data, and assess data utility) → Refined Stage (Design and refine data, generate ad-hoc reports, prototype modelling) → Production Stage (Optimse data, regular reporting and data products, and Services) *Data value funnel:* Data sources → exploratory analysis → value (either direct or indirect)
[29]	Data lake and DW for ensuring data quality	Select data from different sources → structuring → cleaning → enrichment → validating
[34]	Indicate the iterative process of wrangling and analysis. Argue that analysts might more effectively wrangle data through new interactive systems that integrate data verification, transformation, and visualisation.	Original dataset(s) → Diagnose → Transform → First load in analysis application → Start of analysis → Repeat with new data → Document
[28]	Organise the data preparation processes into six broader categories.	Data discovery → Data validation → Data structuring → Data enrichment → Data filtering → Data cleaning
[35]	Present and describe a set of processes that collectively define the actions of DW and propose that what is really needed is a curated data lake, where the lake contents have undergone a curation process that enables its use and delivers the promise of ad-hoc data accessibility to users beyond the enterprise ICT staff.	Data procurement → Vetting Data for Licensing and Legal Use → Obtaining Data → Describing Data → Grooming data → Provisioning data → Preserving data
[3]	Investigate a number of DW pipelines in the form of workflows to find commonalities or patterns in the way DW is performed in practice, considering a multitude of data analysis tasks and data sets, devised by data analysts with varying levels of experience	Loading data → Exploring data → Cleaning → Intra-record structuring → Inter-record structuring → Integrating → Data enriching data → Transforming data → Publishing data

Table 1. *Continued.*

Author(s)	Focus	Data wrangling tasks and or pipeline
[36]	Proposed a data cleaning framework for real-world research, focusing on the three most common types of dirty data (duplicate, missing, and outlier data), and a normal workflow for data cleaning to serve as a reference for the application of such technologies in future studies.	Step 1: Back up and prepare the raw data for cleaning → Step 2: Review the data to formulate cleaning rules → Step 3: Implement the cleaning rules → Step 4: Verify and evaluate the quality of the cleaned data → Step 5: Warehouse after data cleaning
[37]	DW for big data	Data source identification → Data extraction → Data integration → Wrangled data
[30]	Data context informed DW	Identifying data sources and understanding data context → Schema matching → Schema mapping generation and validation → Value format transformation → Rule-based data repair (detecting and repairing certain classes of data errors)
[2]	Review the state-of-the-art in data preparation, by describing functionalities that are central to data preparation pipelines	Profiling → Matching → Mapping → Format transformation → Data repair

4. DATA DESCRIPTION

This article uses data from a Zimbabwean medical insurance company, Cimas Medical Aid Society, which offers several medical aid packages. The dataset includes 16- to 101-year-olds with essential diabetes (ICD-10 Code: E10-) and hypertension (ICD-10 Code: i10-). The data covered January 1–December 31, 2022. The Cimas Medical Aid Society transactional processing system (TPS) provided medical claims and patient-level data for the article. Cimas' TPS collects large medical and pharmaceutical claim data from members. This data applies to employer- or individually-sponsored private health insurance recipients. This comprises members, spouses, and dependants. We included essential diabetics and hypertensives who met the ICD-10 criteria. The provision also requires their enrollment before January 1, 2022. Monthly refills or claims of diabetes and hypertensive medications from Cimas TPS enabled medication refill adherence analytics. We used patient-level data, prescription refill thresholds, and related monthly medication refill data. The Cimas ICT personnel provided us with two de-identified raw diabetic and hypertensive datasets. The first two CSV datasets had 41 columns or features. The hypertension dataset had 218,816 records and the diabetic dataset had 52,616 rows. The dataset included variables concerning internal system operations and patient data. These variables included membership number, pharmacy practice number, medical claims invoice number, claims code, service date, gender, birthdate, current age, annual claim amount, Cimas members' annual medical aid contributions, and option name, which were medical aid package subscribed to by Cimas members. We used Python to wrangle these two datasets.

The generated dataset is available online at the Mendeley repository accessible at the following link: https://data.mendeley.com/preview/zkp7sbbx64?a=700b962c-8a07-4a53-a2fb-cf84ccf316a3. All the datasets (before and after DW) and Jupiter notebook housing the commands and functions used to generate and prepare the dataset are in the main folder named *'Datasets- Diabetes and Hypertension Data'* uploaded as a zipped folder accessible on the same link: https://data.mendeley.com/preview/zkp7sbbx64?a=700b962c-8a07-4a53-a2fb-cf84ccf316a3. This main folder contains four (4) sub-folders and a Microsoft Excel File. These are further described in Table 2 below.

Table 2. Description of the dataset folder, sub-folders, and related files.

Folder Name	Description
Datasets- Diabetes and Hypertension Data *(Main Folder)*	Inside this main folder, there are four sub-folders named, 'Datasets Before Wrangling', 'Additional Columns', 'Complications Development', 'Jupyter Notebook' and a Microsoft Excel File named, 'Final Prepared Dataset - Diabetes and Hypertension Data'. **Note:** The contents of these sub-folders and files are explained in the next rows below.
Datasets Before Wrangling *(Sub-Folder)*	Inside this sub-folder, there are 2 Microsoft Excel CSV files named 'Diabetes Adherence Data' and 'HTN Adherence Data'. These two initial raw datasets contain 41 columns/features including the main policy holder membership number, adherence threshold, gender, service date, amount claimed, dis (pharmacy practice number), units (medical supplies dispensed to a patient for various ailments), option name (medical aid package subscribed) and current age among other patient-level and internal system processing variables. **Note:** The variables deemed relevant for the dataset generation process in this article are further explained in Table 3. The 'Diabetes Adherence Data' file is one of the two initial raw datasets of individuals aged 16–101 with essential diabetes (ICD-10 Code: E10-) who were covered by Cimas medical aid. It contains 52,616 rows of data and 41 columns. The HTN Adherence Data' file is one of the two initial raw datasets of individuals aged 16–101 with essential hypertension (ICD-10 Code: i10-) who were covered by Cimas medical aid. It contains 218,816 rows of data and 41 columns.
Additional Columns *(Sub-Folder)*	Additional columns are additional features related to data points in the 'Diabetes Adherence Data' and 'HTN Adherence Data' csv files and were generated based on data and information extracted from Cimas brochures and website. Inside this sub-folder named 'Additional Columns', there are 2 Microsoft Excel CSV files named 'Additional Columns' and 'HTN_Additional Columns'. Each of the files has four columns: OPTIONNAME, SCHEMETYPE, COVERTYPE, ANNUALCONTIRUBTION. These are annual contributions towards medical aid cover by the members, hospital cover type (standard or comprehensive), the medical aid scheme types subscribed by the members and whether the scheme type is basic, medium or premium level **Note:** These variables are further explained in Table 3. The 'Additional Columns' pertains to the diabetes and contains 10 rows of data (excluding the header).

Table 2. *Continued.*

Folder Name	Description
	The 'HTN_Additional Columns' pertains to the hypertensive patients and contains 11 rows of data (excluding the header).
Complications Development *(Sub-Folder)*	This sub-folder contain four Microsoft Excel CSV files named, 'Dialysis', 'Neurology', 'Opthamol', and 'Physician'. Each file has two columns: Beneficiary number and specialist services sought as a result of a complication development related to diabetes and hypertension. The 'Dialysis' file contains all the list of all beneficiaries who visited a dialysis specialist as a result of diabetes or hypertension – related complications despite already being on medication for diabetes or hypertension. The 'Neurology' file contains all the list of all beneficiaries who visited Neurologists as a result of diabetes or hypertension – related complications despite already being on medication for diabetes or hypertension. The 'Physician' file contains all the list of all beneficiaries who visited Physicians as a result of diabetes or hypertension – related complications despite already being on medication for diabetes or hypertension. The 'Dialysis' file contains all the list of all beneficiaries who visited a dialysis specialist as a result of diabetes or hypertension – related complications despite already being on medication for diabetes or hypertension.
Final Prepared Dataset - Diabetes and Hypertension Data *(Excel File)*	This is the final cleaned, transformed, fully pre-pared and merged dataset with all related variables incorporated. It's in Microsoft Excel file format with 12 columns and 24084 rows of data. The cleaned and enriched datasets for the diabetics and hypertensive patients are merged into this final dataset. Note: The variables making up the final merged dataset are explained in Table 3.
Jupyter Notebook *(Sub-Folder)*	This is a sub-folder named, 'Jupyter Notebook' which contains two files – the same notebook in two different formats: portable document format (PDF) and Hyper Text Markup Language (HTML). The jupyter notebook contains the entire python code – commands, libraries and functions used in the generation of the dataset in this article.

Based on acquired domain knowledge, the relevant features of the dataset are described in Table 3 below.

5. MATERIALS AND METHODS

This section provides an account of how we acquired the data as well as a complete description of the materials and methods used to acquire these data. In that process, we include all codes, libraries, functions, and techniques used for data acquisition, DW, and generation. In order to address this, the DW and generation pipeline discussed in this article combines the widely recognised data preparation tasks and approaches described by existing authors. In addition to DW tasks and pipelines summarised

Table 3. Description of the eleven variables before data transformation.

Variable	Data type	Description
ADHERENCE	Numerical/ Discrete	This is the target or dependent variable. The values (adherence thresholds) ranges from 1 to 12 representing number of medicine refills or medication claims at pharmacies on a monthly basis by diabetes and hypertension patients who are covered by the Cimas Medical Aid Society. If a patient with diabetes or hypertension claims or refills their medicine at least 9 times (once a month) in a span of 12 months (equal to 75% of the expected 12 times per year), they are considered to have met the monthly refills for their condition.
GENDER	Categorical	The gender of the patients is either male or female.
AGE	Numerical/ Continuous	The age range for patients or Cimas members considered in our dataset is from 16 to 103. Prior parental consent is a customary requirement for providing medical treatment to an individual under the age of sixteen.
SCHEMETYPE	Categorical	The medical aid package subscribed by a member.
COVERTYPE	Categorical	This variable represents the various levels of medical aid cover, specifically indicating whether the hospital cover is fully comprehensive or standard.
DIAGNOSIS	Categorical	The medical condition of the patient with NCD. This article centred on individuals who were diagnosed with either diabetes or hypertension.
COMPLICATIONDEVELOPMENT	Categorical	Whether an individual has consulted medical specialists for complications related to diabetes or hypertension.
ANNUALCONTRIBUTION	Numerical/ Continuous	Employees and their dependants make payments to the medical aid company to cover their medical bills. The figures are denominated in Zimbabwe's domestic currency, specifically Zimbabwean dollars.
ANNUALCLAIMAMOUNT	Numerical/ Continuous	Amounts claimed by members to pay for hospital stays or prescription medication purchased from pharmacies. The figures are denominated in Zimbabwe's domestic currency, specifically Zimbabwean dollars.
COMORBIDITY	Categorical	Individuals who have been diagnosed with both diabetes and hypertension.
UNITS	Continuous	The individual number of medical supplies or units distributed to a patient for different medical conditions during the course of the year.

in Table 1, we also incorporate notions from [6] that classify various approaches to DW, such as data transformation, data cleaning, data reduction, and data profiling. The article also incorporates the IBM Garage Methodology, which outlines approaches for preparing data for ML, including data cleansing, data transformation, and feature engineering [38]. According to [39], major DW tasks include data source identification, data collection, and dataset imports. We added data enrichment, integration, and dimensionality reduction as crucial elements, drawing from the work of [39]. Furthermore, the hybrid DW and generation process of our article incorporates data partitioning, which is another common data preparation activity [40], in addition to the usual DW tasks such as data cleansing, reduction, scaling, and transformation. Ultimately, our hybrid DW and generation pipeline encompasses various tasks and building blocks such as data source identification and data acquisition, data importation and inspection, data pre-processing, data enrichment, data integration, profiling the data, transforming the data, dimensionality reduction, data partitioning, and conducting technical validation. Figure 1 below depicts these data-generation building blocks.

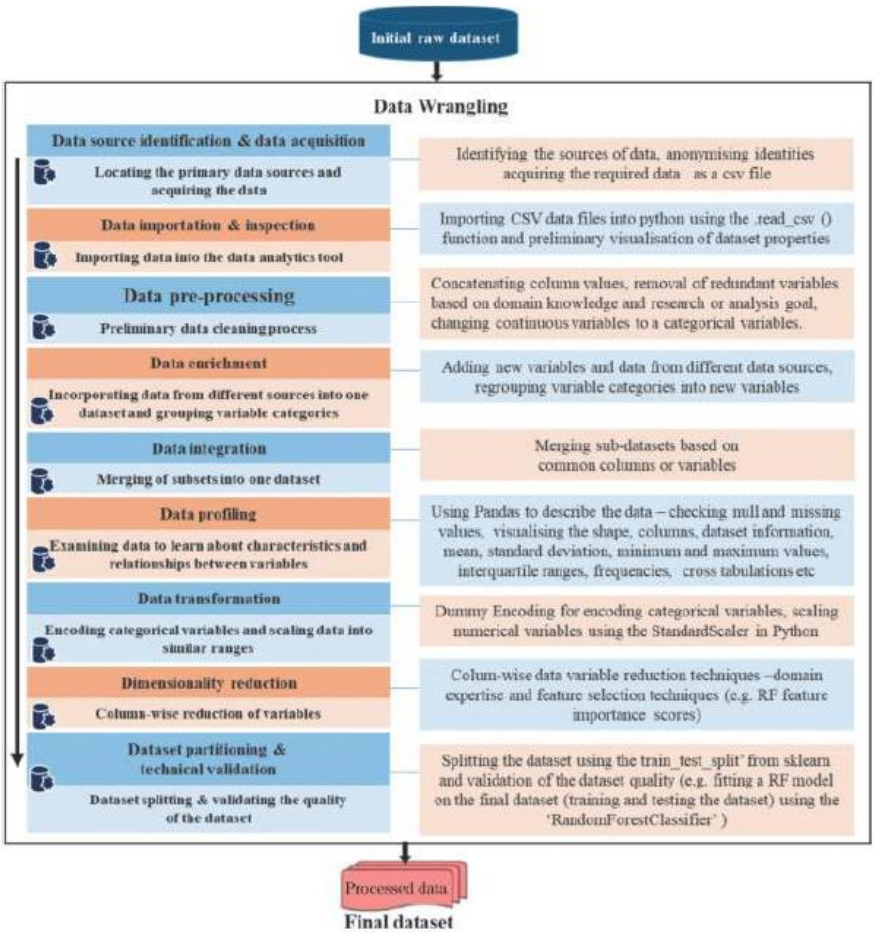


Figure 1. Building blocks of DW and generation.

We employed Python 3.11.3 and Jupyter Notebook 6.5.4 throughout the entirety of the dataset generation procedure. Primarily, we utilised the Pandas library for our DW process. This library provided us with data structures and functions that facilitate the importation, analysis, cleansing, exploration, and manipulation of data.

5.1 Data Source Identification and Data Acquisition

We used data from the Cimas Medical Aid Society, a Zimbabwean medical insurance company that offers several medical aid packages. The data source was Cimas's TPS, a system that collects, retrieves, and manages medical claim transactional data from pharmacies, hospitals, clinics, and doctors. Cimas TPS raw data include member medical claims, internal systems, and patient data. These include membership number, age, beneficiary or dependant number, claim receipt and assessment date, medical scheme type, and ICD-10 health condition among other variables. We requested Cimas ICT only to provide us with relevant data on diabetes and hypertension patients because our article focuses on these two NCDs. We divided diabetes and hypertension data using the ICD-10 code. We filtered all records for diabetes and hypertension by selecting those with 'E1' and 'i10' in the 'ICD10' column. ICD-10 codes classify diabetes as E10–E16 and hypertension as i10–i15. It is important to mention that evaluating adherence is most effective within the same measurement year. The adherence measure calculates the percentage of patients whose prescription claims are covered for the same therapeutic category during the measurement year. The data collected only included records for 2022 data from January 1 to December 31 after excluding other years such as 2020, 2021, and 2023. The final requirement was to ensure that the dataset only included medication refill entries at pharmacies. Importantly, the dataset includes all patient transactions with pharmacies, medical institutions, and professionals. Thus, the data generation process only included pharmaceutical refills and medical claims. Pharmacy-based medication refills were filtered by the discipline column, which codes whether a transaction involved a medical professional or a pharmacy. The screening process used code '60' to identify pharmacy medication refill transactions. Code '60' denoted medicine refill transactions associated with pharmacy claims. After filtering for hypertension, diabetes, 2022 records, and discipline code '60', we got a 52,616-row, 41-column CSV diabetes dataset. Hypertensive patients had 218,816 rows and 41 columns. We imported these CSV files using Python's '.read_csv' function. File formats should be inclusive. Delimited text formats like CSV are generally convertible, making them better than statistical software formats like Microsoft Power BI or SAS.

5.2 Data importation and Inspection

With Python version 3.11.3 and Jupyter Notebook version 6.5.4, we initiated the DW process by importing the Pandas library and reading the initial two raw datasets containing the 41 features. The two CSV datasets that we imported were named 'Diabetes Adherence Data' and 'HTN Adherence Data'. We imported the diabetes adherence data with its header using the '.read_csv' method and stored it into a data frame named 'diabetes'. We imported the 'HTN Adherence Data' with its header using the '.read_csv' method and stored it into a data frame named 'htn'. The first row of the data frame contains the names or labels for each column. It is essential for it to store the names or headers for every column. It essentially

assists the researcher and analysts in understanding the role of each column within the data frame. Subsequently, we inspected the two imported datasets by utilising the df.head() and df.tail() functions. By utilising the df.head() function, we successfully retrieved the initial 5 rows of data along with the names of all the columns. In contrast, the df.tail() function retrieves the latest 5 rows of the data frame, including both the rows and their corresponding column names. Concurrently, information showing 52616 rows × 41 columns for the 'diabetes' data frame and 218816 rows × 41 columns for the 'htn' data frame was also displayed, showing that all the rows in the CSV files were successfully imported into Python. The imported datasets had the columns shown in Figure 2 below.

```
Index(['MEMBER', 'INO', 'PZTIQNT NZXQ', 'ADHERENCE', 'GENDER', 'PRACTICE NO',
       'DIS', 'PROVIDER', 'AS AT NETWORKS', 'REFERRING PRACTICE', 'PAYER NAME',
       'INV REF', 'SERVICE DATE', 'ASSESS DATE', 'DATE RECEIVED', 'CLM CODE',
       'CODE DESCRIPTION', 'UNITS', 'SCRIPT CODE', 'AMOUNT CLAIMED',
       'PAID FROM RISK AMT', 'PAID FROM THRESHHOLD', 'PAID FROM SAVINGS',
       'RECOVERY AMOUNT', 'TOTAL AMOUNT PAID', 'TARIFF', 'CO-PAY', 'PAY TO',
       'REJ', 'REV', 'AUTH NO', 'DL', 'CLAIM NO', 'CLAIM LINE NO',
       'DUPLICATE CLAIM', 'DUPLICATE CLAIM LINE', 'OPTION NAME', 'PAPER/EDI',
       'ASSESSOR NAME', 'BIRTHDATE', 'CURRENT AGE'],
      dtype='object')
```

Figure 2. Columns in the imported datasets.

5.3 Data Pre-processing

During the dataset pre-processing phase, we initiated the creation of a unique identifier (column) named 'BENEFICIARY' for each imported dataset, specifically the 'diabetes' and 'htn' datasets. This unique identifier (the BENEFICIARY column) was created by concatenating the "Member" number and a suffix captioned "INO, where "Member" denoted the policyholder's unique membership number and "INO" denoted the suffix number of any dependants who would be covered by the same membership number. Therefore, it is possible to come across many dependants with the same membership number, and the only way to distinguish them from each other is by the suffix in the 'INO' column. An example of this would be the principal policyholder, 92226868, who had an "INO" or suffix of 0. Yet, there was also a dependant, who had the same membership number (92226868) as the principal policyholder, along with an "INO" or suffix of 1. Thus, to distinguish the main policyholder from their dependants under the same membership number, it was necessary to establish a unique identifier. This identifier, labelled 'BENEFICIARY', was formed by concatenating the 'Member' or membership number with the 'INO or Suffix'. By concatenating the membership number and the suffix, we managed to create a primary key that serves to uniquely identify and differentiate a dependant or beneficiary from other family members associated with the same membership number. We achieved the concatenation of the two fields by running the following command: diabetes['BENEFICIARY'] = diabetes.MEMBER.astype(str) + diabetes.INO.astype(str). We employed the same procedure to generate the 'BENEFICIARY' column within the 'htn' data frame.

Subsequently, we eliminated the majority of the attributes or columns that we considered irrelevant for evaluating MA. Variables about external medical claims and internal system processing, such as the medical claims invoice number, claims code, claim line number, claims assessor name, internal system

claim processing authorisation number, internal system claim processing authorisation type, name of the employer of the medical aid policyholders, claim assessment data, claim receipt date, and other non-relevant features or columns, were discarded. Given that the initial datasets had both the 'CURRENT AGE' and 'BIRTHDATE' variables, we chose to discard the birth date variable. This decision was based on the fact that the current ages of the members were already calculated and sufficient for our analysis concerning the age variable. To remove unwanted variables, such as 'BIRTHDATE', from the data frame, we employed the following command: diabetes.drop(['BIRTHDATE'], axis=1, inplace=True). The command specifies the column to be dropped as 'BIRTHDATE'. Here is the syntax for removing all the remaining unneeded columns: The first argument should specify the column(s) that need to be removed using the 'drop' function. Since the 'drop' method can eliminate columns or rows, we needed to indicate the axis to which the first argument pertains. In our particular scenario, specifying 'axis=1' indicates that we are eliminating the column rather than the row. It is important to mention that by default, the 'drop' method creates a new data frame instead of making changes to the current 'diabetes' data frame. By setting the 'inplace' parameter to the boolean value 'True', the original 'diabetes' data frame had the 'BIRTHDATE' column removed. To facilitate the removal of unwanted columns, we employed the DataFrame[[]].copy() method to generate duplicates of the existing data frames while exclusively specifying the columns or variables we intended to retain, as demonstrated here: diabetes = diabetes[['BENEFICIARY', 'ADHERENCE', 'GENDER', 'UNITS', 'AMOUNT CLAIMED', 'OPTION NAME', 'CURRENT AGE']].copy(). Therefore, at this point, the 'diabetes' data frame would now exclusively contain the relevant variables for our endeavour. The same procedure was executed for the 'htn' data frame. The variables 'BENEFICIARY' and 'OPTION NAME' were temporarily retained as they were necessary for the reclassification and generation of new variables in upcoming data enrichment and merging stages. Removing unneeded variables when cleaning a dataset helps prevent the profiling of unnecessary variables and data during subsequent DW tasks.

Before moving forward with more DW steps, we renamed the variable "CURRENT AGE" to the more relatable "AGE," and the variable "OPTION NAME" to "SCHEME TYPE". We used the diabetes.rename(columns={ "CURRENT AGE": "AGE", "OPTION NAME": "SCHEMETYPE") command to rename the variable. We replicated the same process with the 'htn' dataset. In additional processes, we excluded individuals below the age of 16. We made the decision to exclusively work with patient data of patients who are at least 16 years old because standard procedure requires parental agreement before treating a child less than sixteen. By using a minimum age definition, the process of eliminating inappropriate values reduces the overall size of the datasets. This allows for a more concise representation of the data with a smaller storage capacity. In order to exclude individuals under the age of 16, we first inspected all the distinct ages in the 'AGE' column of the diabetes data frame using the diabetes['AGE'].unique() function. We found that there were records with ages below 16, ranging from 1 to 15 years. We executed the command "diabetes = diabetes[diabetes['AGE'] >= 16]" to filter the records in the diabetes data frame. This command specifically filtered records with ages equal to or more than 16 years. The filtered records were then stored in the same 'diabetes' data frame. We also applied the same procedure of excluding individuals under the age of 16 to the 'htn'

dataset. The number of rows in the 'diabetes' data frame decreased from 52,616 to 52,388, whereas in the 'htn' data frame it decreased from 218,816 to 218,660.

In subsequent processes, the 'ADHERENCE' column was transformed from a continuous variable to a categorical variable using the pd.cut() function, specifying the desired bins and labels. We then assigned the resulting values to the existing adherence column. The pd.cut() function replaced the pre-existing continuous variables with the newly specified categories. The Pandas 'cut' function, often known as pd.cut(), is an effective method for converting continuous data into categorical data. In the 'diabetes' dataset, we employed the pd.cut() function to convert the continuous data of diabetes['ADHERENCE'] into categorical data. We assigned the labels ['NON-ADHERENT', 'ADHERENT'] to the data based on the specified bins [1, 8, 12]. The pd.cut() function consists of three primary components: the bins, which define the cut-off points for dividing continuous data into bins, and the labels, which are important for labelling the resulting bins. Two lists must be filled with the threshold values for the bins. It is important to note that the number of labels will consistently be one less than the number of bins. The initial value in the sequence denotes the beginning of the bin, while the subsequent value indicates the cut-off point of the bin. Within our context, the bins consist of monthly medicine refills categorised from 1 to 8, indicating a lack of adherence to medication, and from 9 to 12, indicating adherence to meds. The same procedure of converting the adherence column from a continuous variable to a categorical variable using the pd.cut(column, bins=[ ], labels=[ ]) function was also employed on the 'htn' data frame. We examined the 'ADHERENCE' column in the diabetes data frame by utilising the print(diabetes['ADHERENCE'].value_counts()) method. The value_counts() method in Pandas generates a series that includes the count of distinct values. In our case, the command returned 36195 (adherent class) and 15355 (non-adherent class) before removing duplicate records for the diabetes dataset, whereas there were 147037 adherent and 66760 non-adherent hypertensive patients before removing duplicate records as well.

5.4 Data Enrichment

Data enrichment involves augmenting the existing dataset with new attributes and extending the initial raw dataset with data obtained from external data sources, which provide data connected to the existing data points (Wang, He, Shea, and Wang, 2018). Thus, data enrichment adds value or supplementary information to existing data from separate sources. In our case, it involved sourcing additional information about the existing data points, which are data already held, and regrouping column values about already existing variables in the dataset. This process aims to create a unified and integrated single view of all the data sources. We created and coded additional variables into what we think are the most logical categories as part of the data enrichment process. As a part of data enrichment, we collected data about the types of schemes subscribed to by Cimas members, the monthly contributions made by Cimas members for each scheme type, and the hospital cover types that Cimas members are eligible for or have access to based on their monthly contributions or the medical aid packages they have subscribed to. The data about the various tiers of medical aid packages, the two categories of hospital cover (comprehensive and standard), and yearly contributions were obtained from Cimas' brochures, catalogues and website. Consequently, we

successfully generated and incorporated additional variables: COVERTYPE, ANNUAL CONTRIBUTION, and SCHEMETYPE.

The brochures indicate that the 'Standard' health cover includes the studymate private, basiccare, icare, private lite, and private hospital packages. On the other hand, the comprehensive cover includes the private hospital plus, medexec, medexec plus, and preme packages. We compiled all of this data in Microsoft Excel and saved the file as a CSV file, including columns labelled 'OPTION NAME' or 'SCHEME TYPE', 'COVERTYPE', and 'ANNUALCONTRIBUTION', along with their corresponding values. The 'OPTION NAME' included values corresponding to different medical aid packages: studymate private, basiccare, icare, private lite, private hospital, private hospital plus, medexec, medexec plus, and prime. The 'OPTION NAME' values labelled as 'Study mate private', 'basiccare', 'icare', 'private lite', 'private hospital', 'private hospital plus', 'medexec', 'medexec plus', and 'prime' (which represent various tiers of medical aid packages retrieved from Cimas Brochures, catalogues, and websites) were reclassified and recoded as 'BASIC', 'MEDIUM', and 'PREMIUM'. The classification of the 3-level schemes was determined by the monetary contributions made to each scheme. The different schemes were grouped according to the monetary value assigned to the eight distinct schemes obtained from Cimas Medical Aid Society brochures and catalogues, which depict the contribution rates for the various medical aid packages. It is worth noting that recoding is a technique that allows combining or grouping two or more categories of a variable to simplify the process of analytics. In our case, when we recoded, we took a variable such as 'OPTIONNAME' with nine valid categories and combined them to create three categories (or fewer) for easy interpretation of the analytics outputs related to the variable in question. Instead of the nine categories the "option name" variable has in the dataset, we believe that the real distinction in scheme types is whether an individual has a basic-level, medium-level, or premium-level membership type. It is important to mention that recreating the variable groups or categories in the 'SCHEMETYPE' involves reclassifying and recoding the data into the newly formed variable groups. It is crucial to keep in mind that recoding variables results in the reclassification of subjects into different groups.

Furthermore, we successfully obtained additional data from the Cimas brochures regarding the monthly payments made by the members for different medical aid society packages. As an example, Cimas members who subscribed to the 'basiccare' package made monthly contributions of $67357.00. It is worth noting that the amounts are denominated in Zimbabwe's domestic currency, specifically Zimbabwean dollars. When multiplied by 12 months, this amounts to an annual commitment of $808284.00. This means that the available dataset did not initially contain this variable. We reconstructed the data for this variable based on the monthly contributions made by members, which we classified according to their subscription scheme types. We compiled the monthly contributions to generate the annual contributions. We presumed that the primary policyholders contribute the same amount annually as the other beneficiaries under them (dependants under them). Regarding the other schemes that had varying rates of contribution based on age, we calculated the average monthly contribution for each scheme and then multiplied the result by the length of a year to obtain the average annual contribution. Therefore, in such instances, we used the mean values. Using the read_csv() function, we imported the CSV file containing the existing "OPTION NAME" column values corresponding to the newly created "SCHEMETYPE,"

"COVERTYPE," and "ANNUALCONTRIBUTION" columns and values. We integrated the computed annual contributions, cover types, and recreated scheme types into our working dataset that we were generating in Python. Subsequently, we integrated the imported CSV file (additional columns in a CSV file called Add_Columns) with the two pre-existing datasets on diabetes (diabetes) and hypertension (htn). For instance, the following: diabetes = pd.merge(diabetes, Add_Columns, on = 'OPTION NAME', how = 'left') command line merges the diabetes data frame and the newly imported Add_Columns dataset. The merge() function facilitates the merging of two datasets, namely diabetes and Add_Columns. This procedure involves combining the datasets into a single data frame or dataset and matching the rows from each dataset based on common attributes or columns, namely the 'OPTION NAME' column in our example. The merge() function is a variant of the Vlookup procedure, which merges two distinct data tables based on specified criteria, requiring at least one shared attribute (column) across the tables. Executing this operation generates a unified dataset that encompasses all the data from both tables with matching values. Our code to merge the dataset implemented the 'Left Join' operation, which retrieves all the rows from the first data frame and only the matched rows from the second data frame. We also implemented the same procedure for enriching the 'htn' dataset by incorporating additional columns ('SCHEMETYPE', 'COVERTYPE', and 'ANNUALCONTRIBUTION') along with their respective values.

We added an extra column labelled 'COMPLICATIONSDEVELOPMENT' to indicate the diabetics or hypertensive patients who sought medical care from specialists like nephrologists, dialysis specialists, cardiologists, physicians, and ophthalmologists, despite already being on medication for diabetes or hypertension. Diabetics and hypertensive patients typically seek medical care from specialists in four interconnected fields: nephrology, dialysis, internal medicine, and ophthalmology. To generate the "COMPLICATIONSDEVELOPMENT" column, we received supplementary data in the form of CSV files. These files contained data about clients who were taking medication for diabetes or hypertension in 2022. Despite their medication, these clients went on to seek medical attention from specialists such as neurologists, ophthalmologists, physicians, and dialysis specialists due to various complications related to diabetes or hypertension. From our perspective, this variable was crucial as we believed it was essential to ascertain whether NCD-related complications are linked to adherence to medication. The four distinct CSV files we received delineate the individuals who would have sought medical attention from at least one of the following healthcare professionals: neurologists, ophthalmologists, physicians, and dialysis specialists, as a result of complications arising from diabetes or hypertension. The files had two important columns: 'BENEFICIARY' number (the unique identifier) and 'SPECIALIST' visited by the patients. We used the read_csv() function in Python to import the four CSV files, specifically neurology.csv, opthamol. csv, physician.csv, and dialysis.csv. Subsequently, we consolidated them into the primary dataset under development in Python. For instance, in the diabetes data frame, we merged the neurology complications. csv file using the following command line: diabetes = pd.merge(diabetes, neurology, on = 'BENEFICIARY', how = 'left'). We merged the 'diabetes' data frame with the 'neurology' data frame on the 'BENEFICIARY' column using a left join. The code demonstrates that the two data frames that have been integrated are 'diabetes' and 'neurology'. A 'Left Join' is a data operation that merges all the rows from the first data frame with the matching rows from the second data frame. The presence of a shared column or variable

(BENEFICIARY) in both data frames facilitates the merger of the two files. NaN values are included in the outcome when there are missing values of the "on" variable (BENEFICIARY) in the right data frame (neurology). We substituted the missing values with NaN in the primary diabetes data frame. We replicated the identical process for adding the remaining complications development files (opthamol, physician, and dialysis) in both the 'diabetes' and 'htn' data frames. Subsequently, we assigned the value 'YES' to all the corresponding rows in the just-added column and replaced the NaN values with 'NO'. To do this task on the 'NEUROLOGY' column of the diabetes data frame, we employed the subsequent commands:

```
diabetes['NEUROLOGY'] = diabetes['NEUROLOGY'].replace('Neurology', 'YES')
diabetes.NEUROLOGY = diabetes.NEUROLOGY.fillna('NO')
```

We applied the same procedure to the 'htn' data frame. Finally, we intended to create a new column caption called "COMPLICATIONDEVELOPMENT" in both the "diabetes" and "htn" data frames out of these four newly added specialist visits columns. If a patient, identified by their unique 'BENEFICIARY' number, had visited any of the specialists (neurologists, ophthalmologists, physicians, and dialysis specialists) during the year of 2022 which was under consideration, the corresponding cell in the intersection of the beneficiary row and the newly created column (COMPLICATIONDEVELOPMENT) should be filled with the value of 'COMPLICATION'; otherwise, it should be filled with the value of 'NO_COMPLICATION'. In order to accomplish this task using the diabetes data frame, we employed the following command lines:

```
diabetes.loc[(diabetes['NEUROLOGY'] == 'YES') | (diabetes['OPTHAMOL'] == 'YES') |
(diabetes['PHYSICIAN'] == 'YES') | (diabetes['DIALYSIS'] == 'YES'),
'COMPLICATIONDEVELOPMENT'] = 'COMPLICATION'
```

```
diabetes.loc[(diabetes['NEUROLOGY'] ! = 'NO') & (diabetes['OPTHAMOL'] != 'NO') &
(diabetes['PHYSICIAN'] ! = 'NO') & (diabetes['DIALYSIS'] ! = 'NO'),
'COMPLICATIONDEVELOPMENT'] = 'NO_COMPLICATION'
```

We applied the same procedure to incorporate the 'COMPLICATIONDEVELOPMENT' column into the 'htn' data frame. As part of our data enrichment process, we aggregated the individual claims for medication into annual claim amounts. Similarly, we consolidated the dispensed medical supplies or units of individual transactions for diabetes and hypertensive patients into annual totals. Given that our research on MA spanned the entirety of 2022, it was sensible to use consolidated data on annual claim amounts and annual totals of medical supplies or units. In order to calculate the new annual claim amounts column, we generated a new subset of data called "diabetes_unique" and 'diabetes_unitstotal' from the 'diabetes' dataset. We calculated the aggregated 'AMOUNT CLAIMED' columns in the subset by using the unique identifier (BENEFICIARY) and the existing 'AMT CLAIMED' column, respectively. In this procedure, we employed the subsequent command:

```
diabetes_unique = diabetes.groupby('BENEFICIARY')[[ 'AMOUNT CLAIMED']].sum()
diabetes_unitstotal = diabetes.groupby('BENEFICIARY')[['UNITS']].sum()
```

The groupby() function yields a DataFrameGroupBy object that encompasses an aggregate function sum() to compute the total of the specified column ('AMOUNT CLAIMED' or 'UNITS') for each group (BENEFICIARY). The groupby() function partitions the recurring records or data into distinct groups based on a common attribute (BENEFICIARY), performs the specified operation on each group, and subsequently reassembles the data to generate the output. The Pandas 'Group by Sum' method obtains the result by grouping the data and calculating the total sum of values in either the 'AMOUNT CLAIMED' or 'UNITS' column. Subsequently, within the sub-datasets, we used the rename() method to rename the names of 'AMOUNT CLAIMED' to 'ANNUALCLAIMAMOUNT' and 'UNITS' to 'UNITSTOTAL', as shown below.

```
diabetes_unique.rename(columns = {"AMOUNT CLAIMED": "ANNUALCLAIMAMOUNT"}, inplace=True)
diabetes_unitstotal.rename(columns = {"UNITS": "UNITSTOTAL"}, inplace = True)
```

Subsequently, we proceeded to add the annual amount total (ANNUALCLAIMAMOUNT) and the total number of units (UNITSTOTAL) as additional columns to the primary dataset (diabetes) by executing the following commands:

```
diabetes = pd.merge(diabetes, diabetes_unique, on = 'BENEFICIARY', how = 'left')
diabetes = pd.merge(diabetes, diabetes_unitstotal, on = 'BENEFICIARY', how = 'left')
```

The merge() function facilitates the merging of two datasets, namely diabetes and diabetes_unique or diabetes_unitstotal. This procedure involves combining the datasets and aligning the rows from each dataset based on shared attributes or columns, specifically the 'BENEFICIARY' column in our example. The merge() function is a variant of the Vlookup procedure that combines two distinct data tables depending on specified criteria, requiring at least one shared attribute (column) across the tables. Upon executing this procedure, a merged dataset is generated, comprising all the data from both tables that have matching values. The code implemented the 'Left Join' operation, which retrieves all the rows from the first data frame and the corresponding rows from the second data frame. We replicated similar procedures to calculate the 'ANNUALCLAIMAMOUNT' and 'UNITSTOTAL' columns in the 'htn' dataset. In additional DW, it is critical to note that we only required one unique row with annual aggregated values for a patient. Therefore, we eliminated duplicate rows by using the specified unique identifier ('BENEFICIARY') through the execution of the subsequent commands:

```
diabetes.drop_duplicates(subset = ['BENEFICIARY'], inplace = True)
htn.drop_duplicates(subset = ['BENEFICIARY'], inplace = True)
```

The Pandas drop_duplicates() method facilitates the elimination of duplicate values from a Pandas data frame in Python. The subset parameter accepts either a single column label or a list of column labels. The default value is null. After passing the columns or variables, they will exclusively identify duplicate rows. We proceeded to include a new column captioned 'COMORBIDITY' to capture data on comorbidity. The objective was to ascertain information regarding people who have both diabetes and hypertension. Our article solely focused on this specific type of comorbidity. In that vein, we initiated the process by adding an entirely new column labelled 'DIAGONIS' and filling the entire column with the diagnosis value

'DIABETES'. Subsequently, we used the DataFrame.copy() function to generate a separate sub-DataFrame (diabetes_copy) containing two variables (BENEFICIARY and DIAGNOSIS) extracted from the pre-existing diabetes data frame. Subsequently, we changed the name of the 'Diagnosis' column in the diabetes_copy dataset to 'COMORBIDITY'. Next, we performed a left join operation to merge the diabetes_copy dataset with the 'htn' dataset, matching the records based on the 'BENEFICIARY' column. The utilised commands are as follows:

```
diabetes_copy = dfdiabetes[['BENEFICIARY', 'DIAGNOSIS']].copy()
diabetes_copy.rename(columns = {"DIAGNOSIS": "COMORBIDITY"}, inplace = True)
htn = pd.merge(htn, diabetes_copy, on = 'BENEFICIARY', how = 'left')
```

Therefore, hypertensive patients who were present in both the 'htn' dataset and the 'diabetes' dataset, with their beneficiary numbers showing at the same time, were classified as having comorbidity. Put simply, these were hypertensive patients with concurrent diabetes. Rows in the merged primary 'htn' data frame that did not have a match were displayed as NaN (Not A Number). To address the issue of missing values represented as NaN, we used the command line htn.COMORBIDITY = htn.COMORBIDITY.fillna('NO_COMORBIDITY') to replace them with a designated value called 'NO_COMORBIDITY'. The fillna() function substitutes the NULL values with a predefined value, 'NO_COMORBIDITY'. The fillna() method replaces the original DataFrame (htn) with a new DataFrame object unless the inplace parameter is set to True. We replicated the same procedures to calculate the 'COMORBIDITY' column in the primary 'diabetes' dataset.

5.5 Integration

Data integration entails expanding a dataset by merging it with the additional dataset(s) through the use of functions such as union, join, or concatenation methods. Following the creation of supplementary columns and the enrichment of the dataset, we proceeded to merge the primary datasets about diabetes (diabetes) and hypertension (htn). Before merging the data frames, we initiated the process by eliminating columns that were not relevant to assessing MA among the patients. We deleted the unwanted columns or variables by building new sub-datasets, only keeping portions of the required variables and data. We executed the subsequent command: diabetes = diabetes[[]].copy(). To clarify, we used the copy() function to assign the required and relevant variables and data to the diabetes data frame, again named 'diabetes'. To rearrange the columns of the Pandas DataFrame (diabetes) in Python, we employed the diabetes. reindex(columns=[]) method. We replicated the same syntax and process to generate a subset of solely relevant variables as well as rearrange the column sequence in the hypertension dataset (htn). Since we wanted to work with only one merged dataset, the two sub-datasets were merged into a new data frame named 'df' using the following command: df = pd.concat([diabetes, htn], axis = 0). The concat() function in Pandas is a simple yet potent method of merging two or more data frames. The concat() function merges the dataframes diabetes and htn, which are provided as a list. The outcome is a new data frame (df) that merges the rows of the 'diabetes' and 'htn' data frames either horizontally or vertically. The option axis = 0 is used to concatenate two or more data frames along the rows. The penultimate number of rows and columns after the two datasets had been thoroughly cleaned and concatenated was 24084 rows and

11 columns, respectively. The final list of columns includes the following: 'ADHERENCE', 'GENDER', 'SCHEMETYPE', 'DIAGNOSIS', 'COVERTYPE', 'COMORBIDITY', 'COMPLICATIONDEVELOPMENT', 'CURRENT AGE', 'ANNUALCONTRIBUTION', 'ANNUALCLAIMAMOUNT', 'UNITSTOTAL'. Following the consolidation of the cleaned dataset, we subsequently engaged in data profiling.

5.6 Data Profiling

Data profiling is a crucial step in our DW and generation procedure, which involves examining datasets to gain insights into important characteristics and the significance of relationships and correlations among their attributes [5, 41]. Data profiling encompasses several techniques, such as data preview, summarising the primary properties of the data, data description, data distribution analysis, range analysis, identifying trends, pattern analysis, detecting anomalies, and validating assumptions using summary statistics. During the process of data profiling, we carefully engaged in a preliminary investigation of data by examining the structure, content, and distribution of the data. To analyse our final merged dataset (df), which is a Pandas DataFrame, we used the df.isnull().sum() function to examine variables that have missing values. This function facilitates the verification of whether a given value is NaN (Not a Number) or not. The function is also valuable in Pandas series for identifying null values within a series. Remarkably, our dataset exhibited a complete absence of missing or null values. We successfully generated multiple data profile outcomes by employing diverse functions, as illustrated in Table 4 presented below.

Table 4. Data profiling functions and outputs before variable encoding.

Method/Function	Output
df.shape	Return a tuple representing the dimensionality of the DataFrame. The dimensions of our data frame were specified as (24084, 11) in our case.
df.columns	Displays all the 11 columns in the final dataset or index as objects: 'ADHERENCE', 'GENDER', 'SCHEMETYPE', 'DIAGNOSIS', 'COVERTYPE', 'COMORBIDITY', 'COMPLICATIONDEVELOPMENT', 'CURRENT AGE', 'ANNUALCONTRIBUTION', 'ANNUALCLAIMAMOUNT', 'UNITSTOTAL'.
display(df)	The function presents the dataset headers consisting of 11 variables and showcases the first 5 and last 5 rows of the DataFrame in a tabular format with delineated rows and columns. This facilitates the examination of our dataset. The result additionally displays the inscription "24084 rows × 11 columns."
df.info()	Displays details regarding the DataFrame. The analysis revealed that the dataset consists of 24167 entries or records, with a total of 11 columns. The variables in the dataset encompass category, integer, and float data types. Additionally, the memory use of the dataset amounts to 2.0+ MB.
df.describe()	Generate descriptive statistics for the numerical variables: count of records or entries as 24084, mean values such as 57 as the mean age, complemented by the standard deviation (14.48), minimum value (16) and maximum value (103) and the 25%, 50% and 75% percentiles for the same variable.

Table 4. *Continued.*

Method/Function	Output
crosstab()	Generate a frequency table, which is a tabular representation of a categorical variable that displays the distribution of values in a dataset by tallying the occurrences of each value. An example is the use of pd.crosstab(index=df['ADHERENCE'], columns='Frequency Count'), which revealed that 14415 (60%) of individuals were non-adherent to medication, whereas 9669 (40%) were adherent. However, out of the total number of individuals diagnosed with diabetes, only 2697 exhibited non-adherence, whereas 2108 individuals (56%) showed adherence. The remaining distributions were as follows: the number of individuals with diabetes was 4805, accounting for 20% of the total, while the number of patients with hypertension were 19279, accounting for 80%. Among the population, 11013 individuals were female, making up 46%, and 13071 were male, making up 54%.
df.mode()	Identify and choose the data column with the highest frequency in the Pandas DataFrame. The mode() function obtains the mode(s) of each element along the specified axis. Our scenario indicated a higher proportion of males compared to females, a greater number of non-adherent patients compared to adherent ones, and a larger population of hypertension patients compared to diabetics.

We observed that DW components can be repeatedly used throughout the DW process, without being limited to a single stage. For instance, after loading the datasets into the data analytics tool, we promptly carried out dataset inspection and data profiling tasks. We also continuously inspected and profiled the datasets as more DW tasks were in order to consistently monitor the modifications and improvements applied to the dataset. Again, during the DW process, we were recreating and recalibrating the columns, making the removal of unwanted columns an ongoing task.

5.7 Data Transformation

While processing the data, we were aware that the dataset included both categorical and numerical variables. Therefore, we proceeded with the transformation of both the categorical and numerical variables. We used df = pd.get_dummies(df, columns=['categorical variable name'], drop_first=True) for categorical variable encoding. The get_dummies() function generates indicator variables from categorical variables. The function creates a separate dummy variable column for each level of the categorical variable. The dummy variables take on the value "True" if the original categorical variable is at that level or "False" if not. As there are only two gender types in our dataset, it should be noted that when examining a feature like "Gender" and having a value named "male," there is no need for an additional column called "female." This is because it is already known or indicated via male=false. Therefore, if a gender category is not identified as male, it implies that it is female. Dummy features complicate the ML algorithms' fitting process or, worse, make it simpler for overfitting, which is important to consider. The pd.get_dummies() method executed on the 'DIAGNOSIS' variable generated a new variable named 'DIAGNOSIS_HYPERTENSION' with the values 'true' and 'false'. Within the article, the categories were transformed into separate columns and assigned binary values of either True or False. In terms of diagnosis, a 'True' result shows the presence of hypertension, whereas a 'False' value suggests its absence or the presence

of another disease, such as diabetes. Consequently, after the transformation, there will be an increase in the number of features. Regarding the present scenario, the total number of variables or columns has risen from 11 to 12 due to the addition of two new dummy variables ('SCHEMETYPE_MEDIUM' and 'SCHEMETYPE_PREMIUM') for the previously existing 'SCHEMETYPE' variable, notwithstanding the removal of the 'SCHEMETYPE_BASIC' column. The parameter 'drop_first=True' removes the first category (variable) when creating dummy variables. To encode dummy variables, it is necessary to exclude the first category (variable) among the newly created dummy variables for each feature. Avoiding the formation of a dummy variable trap is the reason for excluding the first category among the newly created dummy variables for each feature. The dummy variable trap arises when two or more independent dummy variables, generated through dummy encoding, display a strong correlation or are influenced by multicollinearity. We ultimately transformed the categorical variables into new labels, including GENDER_M, SCHEMETYPE_MEDIUM, SCHEMETYPE_PREMIUM, DIAGNOSIS_HYPERTENSION, COVERTYPE_STANDARD, COMORBIDITY_NO_COMORBIDITY, and COMPLICATIONDEVELOPMENT_NO_COMPLICATION.

The DataFrame method scaler.fit_transform() from the sklearn.preprocessing library was used to scale down the numerical variables: age, annual claim amount, annual contribution, and units. The StandardScaler function standardises a feature by removing the mean and then scaling it to have unit variance. The feature scaling data pre-processing approach facilitates the transformation of feature or column values in the dataset to a uniform scale. Following the application of the standard scaler, we subsequently executed the min() and max() functions to observe the minimum and maximum values of all the scaled columns. We specifically observed that the annual contribution values had been reduced to a range of -3 to 3. Consequently, the scaler transformed the data to fit within a designated range. The objective is to achieve an equal contribution from every feature to the model and prevent the overpowering influence of features with higher values.

Data Records

The data records are published in Mendeley Data: https://data.mendeley.com/datasets/zkp7sbbx64/2. All data files (including supporting files) are CSVs with a simple structure. Following variable encoding, the final dataset created consists of 24084 records and 12 columns. Table 5 describes the final collection of variables obtained after data pre-processing, dataset enrichment, and transformation.

5.8 Dimensionality Reduction

There are several main approaches for reducing the number of variables in a dataset: employing domain knowledge to directly select variables; employing statistical feature selection techniques to identify significant variables for further analytics; or employing feature extraction techniques to generate useful features for data analysis. We leveraged the inherent feature importance mechanism of the Random Forest algorithm to ascertain the significance of each feature. We derived the values presented in Figure 3.

Table 5. Description of the twelve variables after data pre-processing and transformation.

Variable	Data type	Description
ADHERENCE	Categorical	This is the target or dependent variable. After transforming the initial numerical values (adherence thresholds) into categorical data, we had two categories (either Adherent or Non-adherent). The 'Adherent' class were the patients which would have claimed or refilled their medicine at least 9 times (once a month) in a span of 12 months (equal to 75% of the expected 12 times per year). The non-adherent were those with medicine refill thresholds under 9 monthly refills.
AGE	Numerical/ Continuous	The age range for patients or Cimas members considered in our dataset is from 16 to 103. Prior parental consent is a customary requirement for providing medical treatment to an individual under the age of sixteen.
ANNUALCONTRIBUTION	Numerical/ Continuous	Employees and their dependants make payments to the medical aid company to cover their medical bills. The figures are denominated in Zimbabwe's domestic currency, specifically Zimbabwean dollars.
ANNUALCLAIMAMOUNT	Numerical/ Continuous	Amounts claimed by members to pay for hospital stays or prescription medication purchased from pharmacies. The figures are denominated in Zimbabwe's domestic currency, specifically Zimbabwean dollars.
UNITSTOTAL	Continuous	The total quantity of medical supplies or units distributed to a patient for different medical conditions throughout the year.
GENDER_M	Categorical	Newly created variable after encoding the 'GENDER' column. The 'GENDER_M' takes two values, TRUE and FALSE. The 'TRUE' value signifies that the patient is male and 'FALSE' represents the opposite – Female.
SCHEMETYPE_MEDIUM	Categorical	The medical aid package subscribed by a member. The 'SCHEMETYPE_MEDIUM' is one of the newly created dummy variables for 'SCHEMETYPE' after variable encoding. The variable takes two values - TRUE and FALSE. The 'TRUE' value signifies that the member subscribes to the medium class of the medical aid cover schemes offered by Cimas, whereas the 'FALSE' represents that a member subscribes to the lower class – the basic medical aid scheme class.

Table 5. *Continued.*

Variable	Data type	Description
SCHEMETYPE_PREMIUM	Categorical	The medical aid package subscribed by a member. The 'SCHEMETYPE_PREMIUM is one of the newly created dummy variables for 'SCHEMETYPE' after variable encoding. The variable takes two values - TRUE and FALSE. The 'TRUE' value signifies that the member subscribes to the premium class of the medical aid cover schemes offered by Cimas, otherwise a member subscribes to the lower class – the medium medical aid scheme class.
DIAGNOSIS_HYPERTENSION		The medical condition of the patient with NCD. The 'DIAGNOSIS_HYPERTENSION' variable is a newly created dummy variable after variable encoding which takes two values - TRUE and FALSE. The 'TRUE' value signifies that the patient is a hypertensive patient, otherwise the patient is a diabetic.
COVERTYPE_STANDARD	Categorical	This variable represents the various levels of medical aid cover, specifically indicating whether the hospital cover is fully comprehensive or standard. The 'COVERTYPE_STANDARD' variable is a newly created dummy variable after variable encoding which takes two values - TRUE and FALSE. The 'TRUE' value signifies that the member's medical aid subscription affords him or her a hospital cover labelled as 'STANDARD', otherwise its 'COMPREHENSIVE'.
COMORBIDITY_NO_COMORBIDITY		Individuals who have been diagnosed with both diabetes and hypertension. The 'COMORBIDITY_NO_COMORBIDITY' variable is a newly created dummy variable after variable encoding which takes two values - TRUE and FALSE. The 'TRUE' value signifies that the patient has no comorbidity, otherwise he or she is a comorbid patient (having both diabetes and hypertension).
COMPLICATIONDEVELOPMENT_NO_COMPLICATION	Categorical	Whether an individual has consulted medical specialists for complications related to diabetes or hypertension. The 'COMPLICATIONDEVELOPMENT_NO_COMPLICATION variable is a newly created dummy variable after variable encoding which takes two values - TRUE and FALSE. The 'TRUE' value signifies that the patient did not develop any diabetes or hypertension related complications, otherwise he or she develop the complications.

	Feature Score
UNITSTOTAL	0.599417
ANNUALCLAIMAMOUNT	0.306223
AGE	0.048715
COMPLICATIONDEVELOPMENT_NO_COMPLICATION	0.012243
COMORBIDITY_NO_COMORBIDITY	0.010182
DIAGNOSIS_HYPERTENSION	0.008696
ANNUALCONTRIBUTION	0.007248
GENDER_M	0.003101
COVERTYPE_STANDARD	0.001659
SCHEMETYPE_PREMIUM	0.001502
SCHEMETYPE_MEDIUM	0.001013

Figure 3. Random Forest feature importance scores.

Figure 3 illustrates the comparative significance of each feature in the context of ML-based prediction. The feature importance rankings are arranged in descending order, with the most significant feature listed first. The results suggest that the most important variable influencing MA is the 'UNITSTOTAL' (the total number of different medications and medical supplies given to a patient each year), as indicated by the highest score of feature importance, which is 0.599. The RF feature importance measure indicates that the annual claim amount has a high impact on MA, with a feature importance score of 0.306. It is the second-most influential predictor in this regard. The results suggest that the variables "units" and "annual claim amount" made the most substantial contributions of 60% and 31%, respectively, in predicting the target variable, MA. The age of diabetic or hypertensive patients contributes 5% to the model classification performance. Based on the RF importance scores, the features that have the least significance in the MA prediction model are scheme type and cover type, as their values are below 0.05. The scheme type demonstrated the least amount of importance, as indicated by the reported importance score. Nevertheless, all variables within the RF prediction model exhibited statistically significant scores above zero. As a rule of thumb, it is generally recommended to include any features with RF importance scores greater than zero. Therefore, researchers and data analysts can build ML classifiers using all 11 MA predictors that underwent RF feature importance computation in this article.

5.9 Technical Validation

The data presented in this article were carefully collected, processed, and thoroughly verified for accuracy by a team of experienced computer and data scientists who possess a deep understanding of aspects pertaining to computational intelligence. The verification process involved a two-step validation procedure. Firstly, during the data extraction phase, we implemented a double-check system. The original raw data were independently extracted by two ICT specialists who were involved in the initial data extraction process. The initial raw dataset columns were compared, and discrepancies between the two extractions were analysed. Fortunately, all the records matched. This process ensured that the data

extraction was accurate and consistent. Secondly, we performed preliminary data inspection and profiling on the imported datasets in order to describe the shape of the datasets, display the columns, check null or missing values, and generate descriptive statistics for the numerical variables. This allowed us to identify any outliers or data points that deviated significantly from the expected values based on all the extracted data points. At the end of our DW process, we validated the quality of the dataset by assessing the performance of a random forest model on our prepared dataset. Usually, the people in charge of manually gathering and providing the data may lack the commitment to guaranteeing accuracy and may resort to quick shortcuts while capturing the data. Default values provided by an information system often disproportionately represent themselves in the collected data. The automated data-gathering techniques may be flawed, leading to the generation of inaccurate or erroneous data. As such, we deployed a Random Forest model using the 'RandomForestClassifier' module imported from the 'sklearn.ensemble' library.

Before calculating the feature importance scores on our dataset, we divided the dataset into a training set (70%) and a test set (30%) using the 'train_test_split' module imported from the 'sklearn. model_selection' library. The training dataset served as the primary dataset for training the ML model to detect specific patterns or carry out the task of MA classification. The performance evaluation of the RF model on the testing set yielded an accuracy score of 80.67% and an area under the curve (AUC) value of 0.8929. Table 6 shows that the RF model achieved an accuracy score of 80.67% and an AUC value of 0.8929 when employing 11 features as predictors. The model was fitted with a max_depth of 7, used 'entropy' as the criterion, had 200 n_estimators, and had a random state of 1. The accuracy score was computed in Python using the 'accuracy_score' module imported from sklearn.metrics. Accuracy is a metric that evaluates the frequency with which an ML model accurately predicts the outcome. The parameter "n_estimators" denotes the number of trees contained in the forest. The criterion functions as a method for evaluating the quality of a split, particularly when using the "entropy" criterion from the available options. The max_depth option defines the maximum depth that the tree can reach. The random state parameter controls the degree of randomisation during the training process. Using a constant number throughout several cycles guarantees the reproducibility of the training process. Assigning a constant random state value to the model and providing it with the same parameters and training data consistently produces identical results. We conducted trials for model fitting using the default number of n_estimators, which was set to 100, along with max_depth 3, and the criterion 'entropy'. The experimental findings demonstrated a decline in performance concerning accuracy (0.7964) and AUC (0.8816). The experimental findings demonstrated a decline in performance concerning accuracy (0.7964) and AUC (0.8816). Thus, modifying the settings, notably increasing n_estimators to 200 and changing max_depth to 7, led to an improvement in the model's performance. As part of an ablation analysis, we examined the performance of the RF model by eliminating certain features in order to better understand the contribution of certain features to the overall model. As part of the ablation analysis and guided by the RF variable importance ratings in Figure 3, we eliminated five variables with the lowest feature importance scores. Table 6 and Figure 4 below show that after eliminating the five features, the RF model accuracy increased from 0.8067 to 0.8074, as did the AUC score, from 0.8929 to 0.8935. When another feature was removed, the RF model's accuracy and AUC increased to 0.8182 and 0.8977,

Table 6. Feature ablation and ablation analysis results.

RF Classifier Parameters	Number of Predictors	Number of tupples	Predictors included	Accuracy	AUC
max_depth = 7, n_estimators = 200, criterion = "entropy"	10	4805	'GENDER', 'AGE', 'SCHEMETYPE', 'COVERTYPE', 'ANNUALCONTRIBUTION', 'COMPLICATIONDEVELOPMENT', ANNUALCLAIMAMOUNT', 'UNITSTOTAL', 'DIAGNOSIS', 'COMORBIDITY'	0.7864	0.8704
max_depth = 3, n_estimators = 100, criterion = "entropy"	11	24084	AGE', 'ANNUALCONTRIBUTION', 'ANNUALCLAIMAMOUNT', 'UNITSTOTAL', 'GENDER_M', 'SCHEMETYPE_MEDIUM', 'SCHEMETYPE_PREMIUM', 'DIAGNOSIS_HYPERTENSION', 'COVERTYPE_STANDARD', COMORBIDITY_NO_ COMORBIDITY', COMPLICATIONDEVELOPMENT_ NO_COMPLICATION'	0.7964	0.8816
max_depth = 7, n_estimators = 200, criterion = "entropy"	11	24084	'AGE', 'ANNUALCONTRIBUTION', 'ANNUALCLAIMAMOUNT', 'UNITSTOTAL', 'GENDER_M', 'SCHEMETYPE_MEDIUM', 'SCHEMETYPE_PREMIUM', 'DIAGNOSIS_HYPERTENSION', 'COVERTYPE_STANDARD', COMORBIDITY_NO_ COMORBIDITY', 'COMPLICATIONDEVELOPMENT_ NO_COMPLICATION'	0.8067	0.8929
max_depth = 7, n_estimators = 200, criterion = "entropy"	7	24084	'AGE', 'ANNUALCONTRIBUTION', 'ANNUALCLAIMAMOUNT', 'UNITSTOTAL', 'DIAGNOSIS_ HYPERTENSION', 'COMORBIDITY_ NO_COMORBIDITY', 'COMPLICATIONDEVELOPMENT_ NO_COMPLICATION'	0.8074	0.8935
max_depth = 7, n_estimators = 200, criterion = "entropy"	6	24084	'AGE', 'ANNUALCLAIMAMOUNT', 'UNITSTOTAL', 'DIAGNOSIS_ HYPERTENSION', 'COMORBIDITY_ NO_COMORBIDITY', 'COMPLICATIONDEVELOPMENT_ NO_COMPLICATION'	0.8182	0.8977

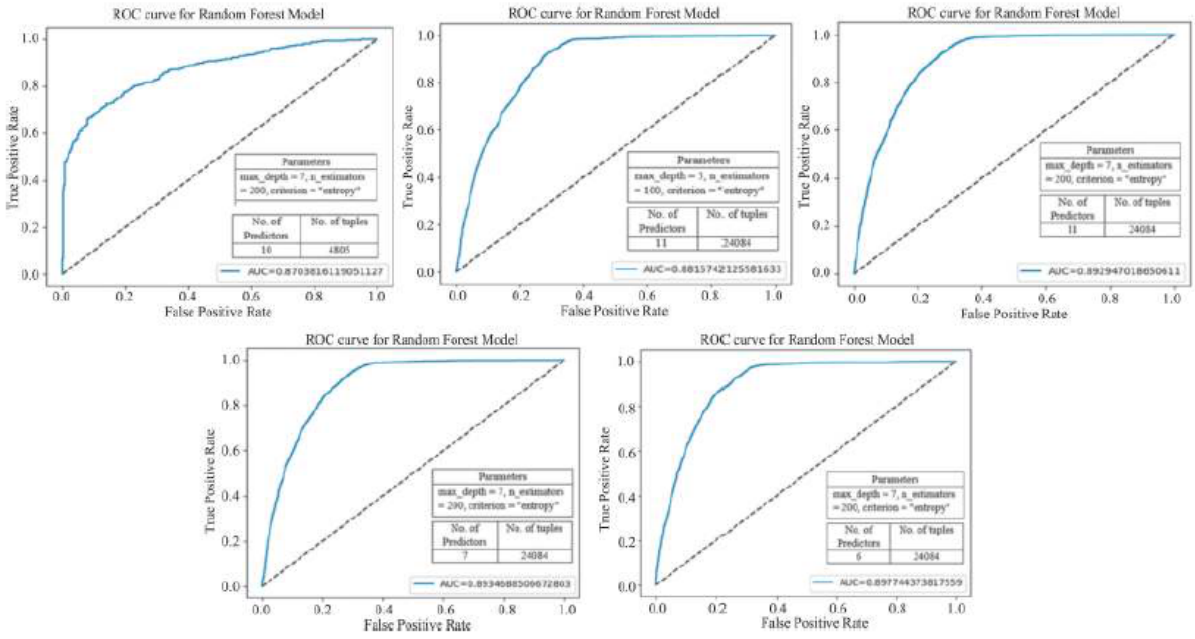


Figure 4. AUC scores for the ablation analysis.

respectively. Future researchers and analysts can use the full feature set or a subset of the available dataset for MA analytics, guided by the results of their ablation analysis techniques. Again, as part of the ablation analysis, we fitted an RF model to a smaller dataset of 4805 records relating to diabetes patients only. The RF model computed on the smaller diabetes dataset exhibited lower performance (accuracy 0.7864 and AUC 0.8704). Indeed, the model performance was lower when compared to the RF model, which was fitted on the final large dataset of 24084 tuples, containing not only the diabetes records but also the 19279 hypertensive records. Table 6 clearly shows that all of the RF models fitted to the 24084-records dataset with the same or different RF classifier parameters produced model accuracies and AUC scores of more than 0.7864 and 0.8704, respectively. The dataset size is regarded as an important factor in influencing the performance of an ML model [42]. Large datasets typically result in improved classification performance, but small datasets can lead to overfitting [43, 44].

The arduous nature of the data generation process leads to numerous technical challenges along the way. Table 7 outlines the key technical challenges in creating the new dataset and articulates their resolution.

6. USAGE NOTES

Future research can directly benefit from the key features offered in the availed dataset. First, researchers can continue to investigate the association between the available independent variables (NCD patient-level features and other factors associated with MA) and the outcome variable (MA). Thus, the

Table 7. The main technical challenges in generating the new dataset.

Technical Challenge	Description of the challenge(s)	How to resolve the challenge
Data acquirement stage challenges	Understanding the data and variables: Before obtaining the initial raw dataset, the available dataset included more than 45 variables. Therefore, we invested considerable effort to familiarise ourselves with the data and ensure its suitability for MA analysis.	We held three meetings with the data provider's (Cimas) ICT department and then worked together to grasp all of the variables individually as well as the data file contents in general. Because the focus of our study was on patients with diabetes and hypertension, we needed to understand the codes that might be used to identify and sort through information related to the NCDs in question. We obtained sub-dataset extracts for diabetes and hypertension using the ICD10 code for disease classification.
	Data reduction and filtering challenge: during preliminary data gathering meetings with the Cimas ICT department, the raw dataset in CSV format exceeded 500000kb. MS Excel 2013 could not fully load this large file due to its limited capacity of 1048576 rows, despite the raw dataset containing over 1700000 tuples.	After citing the MS Excel issue of failing to load all of the original raw dataset tuples, Python 3.11.3 was used to read the initial huge raw dataset. The adherence measure calculates the percentage of patients whose prescription claims are covered for the same therapeutic category during the same measurement year; hence, only records from 2022 were included. Another criterion for lowering the dataset size was to include records of prescription refills and medical claims data. We used the code '60' to identify pharmacy medicine refill transactions. After filtering for hypertension, diabetes, 2022 records, and discipline code '60', we obtained a CSV diabetes dataset with 52,616 rows and 41 columns. The hypertensive patients had 218,816 rows and 41 columns.
	Data confidentiality issue- Medical records are sensitive and confidential.	Before getting the required initial dataset from Cimas ICT department, the data records were anonymised to protect the patients' privacy.

Table 7. *Continued.*

Technical Challenge	Description of the challenge(s)	How to resolve the challenge
Data pre-processing challenges	Record unique identifier: Some dependants had the same membership number, and the only way to distinguish them was to use the suffix in the 'INO' field. For example, the primary policyholder, 92226868, had a "INO" or suffix of 0. However, there was a dependant who shared the same membership number (92226868) as the primary policyholder, along with an "INO" or suffix of 1. Looking at the 'INO' column, the suffixes were not unique across all the rows.	Concatenate the 'Member' or membership number with the 'INO or Suffix' in Python. By concatenating the membership number and the suffix, we managed to create a primary key that serves to uniquely identify and differentiate a dependant or beneficiary from other family members associated with the same membership or policyholder number.
	Irrelevant columns or features	Remove irrelevant variables such as internal system processing such as claim line number and claims assessor name. To facilitate the removal of unwanted columns, we employed either the drop function or the DataFrame[]].copy() method in Python to generate duplicates of the existing data frames while exclusively specifying the columns or variables we intended to retain.
Data enrichment challenges	Incorporating supplementary information from other external data sources that were not originally available in the original raw dataset.	Add columns or features (medical aid scheme types, hospital cover types, and annual contribution towards medical aid) extracted from Cimas brochures. We added an extra column labelled (Complications Development) to indicate the diabetics or hypertensive patients who sought medical care from specialists during the course of medication for their NCD conditions. The additional columns were sub-datasets, which were then merged to the original dataset using the merge function (a variant of the VLOOKUP procedure) in Python.
	Duplications: We only required one unique row with annual aggregated values for a patient. Therefore, there was a need to eliminate duplicate rows by using the specified unique identifier ('BENEFICIARY') or primary key through the execution of the subsequent commands.	Drop the duplicate aggregated records using the drop_duplicate function in Python.

Table 7. *Continued.*

Technical Challenge	Description of the challenge(s)	How to resolve the challenge
Integration challenge	Sub-dataset integration (hypertension datasets and diabetes datasets).	Merged the datasets using the concat() function, which merges the dataframes 'diabetes' and 'htn'.
Transformation challenges	High-value outliers and varying value ranges, which cause bias impact.	Used the StandardScaler function to facilitate the transformation of the numerical values in the dataset to a uniform scale. The scaler transforms the data to fit within a designated range (-3 to 3).
	Non-encoded categorical variables—nominal data—do not have any order and can introduce unintended relationships that mislead the ML algorithm.	We used the get_dummies function to convert the categorical data into numerical format so that it could be used as input for ML algorithms to process and so that the data could be provided to different models.
Data reduction challenge	Elimination of the features could potentially degrade the performance of an ML model. While feature reduction can be highly beneficial, in our case, it presents the challenge of information loss. Reducing the number of features leads to the loss of important information, potentially degrading the total value of the generated dataset.	Employed the RF variable importance mechanism, which demonstrates that all the features had importance scores greater than 0. Thus, all 11 predictors in the final generated dataset can be included for MA analysis. The full feature set or a subset of the availed entire dataset can be used in MA analytics as guided by the output of their ablation analysis procedures. Table 6's demonstration of our ablation analysis can serve as a guide.

created dataset can be used in several future studies to evaluate MA among NCD patients by comparing the performance of various ML algorithms and statistical methods on this availed dataset. Similarly, the dataset can be divided into training and test beds and then used to create virtually new ML algorithms and models for evaluating MA. Consequently, the data serves as a valuable tool for training newly developed ML and deep learning MA assessment models - which then can be tested later on newly provided MA-related datasets.

After developing and pre-training ML models on our available dataset and optimising them to predict MA in NCD patients, the ML models are made available to developers as pre-trained models or application programming interfaces (APIs) that can be used to build intelligent MA tracking or analysis applications quickly and easily. Pre-built models save developers time and resources by reducing the need to create models from scratch [45]. This means that developers may easily plug in these pre-built models and start solving complex issues without worrying about the underlying methods or model training specifics.

Furthermore, data analysts working with policymakers will benefit from insights gained through studies that will use this present dataset, particularly in identifying which variables contribute the most

to medication non-adherence. Thus, the Ministry of Health within countries might develop strategies to improve MA among chronic disease patients by focusing on the specific variables that strongly predict MA among patients. Policymakers can also directly benefit from the ML models built using our dataset. ML can accurately predict MA, thereby enabling the development of interventions aimed at decreasing or preventing nonadherence. According to [46], predictive models can analyse enormous amounts of data, such as demographics, health records, and environmental factors, allowing for the identification of high-risk populations and the development of targeted interventions.

Our data descriptor article provides a full compilation of the systematically arrayed techniques used in DW and generation. Though there is no universally approved compendium of DW stages, this data article presents a more comprehensive approach, entailing nine all-inclusive DW building blocks for generating a dataset. Thus, researchers and data analysts can replicate these DW techniques and activities in their future data-generation research tasks.

The dataset's deposition also allows researchers to reproduce past discoveries. Due to the current crisis surrounding reproducibility [47], replication assumes a critical role in generating reliable and accurate evidence. Reproducibility is an important aspect of science since the replication of discoveries is the process by which they become knowledge [47]. Our dataset has been published, making it publicly available and freely accessible to researchers. Making this dataset openly available will allow academics from all across the world to continue using it for further research. This promotes scientific transparency, which is critical in today's research context. Making the data set available increases transparency and encourages open evaluation and conversation among scholars. This promotes transparency as well as the quality and trustworthiness of published results. Thus, past findings will be subjected to a post-publication review. Hopefully, this technique will advance the cause of open access while preventing irreproducibility. There is a need for transparency and openness in all areas of scholarly publishing by making the data on which published research is built available, discoverable, and re-usable so that the research can be verified to ensure true scientific rigour; in this way, scholarship is built on scholarship and new knowledge is created [48]. Furthermore, because the cost of science is a major concern in the global scientific community, our data set is an invaluable resource for in-depth research into MA evaluation and modelling. In a broader sense, making the dataset available will cut the cost of science and considerably benefit scientific communities worldwide.

7. LIMITATIONS

The dataset does not consider the medically uninsured citizens, who are typically unemployed in formal jobs and represent the most impoverished population in developing nations. It is necessary to create and compile more datasets based on electronic records of patients who can only afford to access public health facilities and do not have health insurance in order to use them in MA analytics in the future. The available raw datasets utilised in the development of this article had limited relevant features; consequently, in future dataset generation articles should make use of raw datasets with large sets of relevant features in the context of a developing country are indispensable.

8. CONCLUSION

In this data article, we generated and published a state-of-the-art dataset that is valuable for examining MA among patients with diabetes and hypertension in the setting of a developing nation, Zimbabwe. We meticulously documented a comprehensive compendium of systematically arrayed DW tasks and techniques that we followed to prepare the dataset. Nevertheless, we recognise that there is no universally agreed-upon set and flow of steps for DW. DW often incorporates specific fundamental building blocks and functionalities that substantially resemble and represent the essence of DW. In light of this, we take a pragmatic approach to the generation of the dataset, incorporating nine essential DW building blocks and tasks, which we arrange in an iterative and incremental manner, akin to an agile approach to dataset generation. Thus, we could use some of the DW tasks iteratively and incrementally throughout the DW process without restricting them to a single stage. For instance, one can promptly carry out dataset inspection and data profiling tasks right after loading the dataset into the data analytics tool. Following each round of DW, one can repeat these actions to consistently monitor the modifications and improvements applied to the dataset. During the DW process, new columns may be recreated or recalibrated, making the removal of unwanted columns an ongoing task. After meticulously preparing the dataset, we used the generated dataset to train an ML model, which obtained a commendable level of performance with an accuracy of 81%. Hence, DW is essential in any ML workflow and has a substantial influence on the prediction capabilities of an ML model. In the future, it will be necessary to create and compile more datasets based on electronic records of patients who can only afford to access public health facilities and do not have health insurance in order to use them in MA analytics. Researchers should prioritise the development of automated methods for performing operational DW tasks to improve the efficiency of data analytics.

ETHICS STATEMENT

The article adhered to ethical guidelines by getting ethical approval on June 8, 2023, after submitting a written request to gather data to the management of the Cimas Medical Ad Society. The authors implemented procedures to safeguard the confidentiality of the Cimas Organisation and its clients. To resolve this matter, we reached an agreement with Cimas' ICT department, which was responsible for extracting the initial raw dataset from Cimas' TPS. The agreement was to adhere to our suggested minimal standard of de-identifying datasets before using them in our research work and publishing them in peer-reviewed journals. In order to ensure the protection of sensitive patient information and adhere to the confidentiality standards outlined in the Health Insurance Portability and Accountability Act, personal data underwent a de-identification procedure. We excluded participants below the age of 16 from the dataset. Though it may vary depending on medical situations, youth under 16 cannot make healthcare decisions without their parents' consent, hence their exclusion from our article. The authors have read and followed the ethical requirements for publication in Data in Brief and confirming that the current work does not directly involve human subjects, animal experiments, or any data collected from social media platforms.

ACKNOWLEDGEMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS

Wellington Kanyongo: Conceptualization, Methodology, Coding, Data curation, Original draft preparation and Validation;

Absalom El-Shamir Ezugwu: Conceptualization, Original draft preparation, Review, Supervision, Validation and Editing;

Tsitsi Moyo: Data curation;

Jean Vincent Fonou Dombeu: Review and Editing.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AND CODE AVAILABILITY

All the data are available at the Mendeley data repository titled. "Dataset for analysing MA among diabetes and hypertension patients: Patient-level and medication refill data", with the following data identification number (DOI number): 10.17632/zkp7sbbx64.2 or URL link: https://data.mendeley.com/datasets/zkp7sbbx64/2, licensed under a Public Domain Dedication licence (CC0 1.0). Data processing, validation and plotting were performed using Jupyter notebooks in a Python 3 environment. Code for data cleaning and analysis is provided as part of the replication package. It is also included in the same repository.

REFERENCES

[1] Joshi, A.P., Patel, B.V.: Data preprocessing: the techniques for preparing clean and quality data for data analytics process. Oriental Journal of Computer Science and Technology 13, 78–81 (2020). doi: DOI: http://dx.doi.org/10.13005/ojcst13.0203.03

[2] Fernandes, A. A. A., Koehler, M., Konstantinou, N., Pankin, P., Paton, N. W., Sakellariou, R: Data Preparation: A Technological Perspective and Review. SN Computer Science 4(4), p. 425 (2023). doi: https://doi.org/10.1007/s42979-023-01828-8

[3] Almasaud, A., Sampaio, S., Sampaio, P.: Mining Data Wrangling Workflows for Patterns, Reuse and Optimisation Opportunities. In: EDBT/ICDT Workshops (2023)

[4] Kandel, S., Paepcke, A., Hellerstein, J., Heer, J.: Wrangler: Interactive visual specification of data transformation scripts. In: Proceedings of the Sigchi Conference on Human Factors in Computing Systems, pp. 3363–3372 (2011)
[5] Abedjan, Z., Golab, L., Naumann, F.: Profiling relational data: a survey. The VLDB Journal 24, pp. 557–581 (2015). doi: https://doi.org/10.1007/s00778-015-0389-y
[6] Barapatre, D., Vijayalakshmi, A.: Data preparation on large datasets for data science. Asian Journal of Pharmaceutical and Clinical Research 10(13), 485–488 (2017). doi: http://dx.doi.org/10.22159/ajpcr.2017.v10s1.20526
[7] Abdallah, Z. S., Du, L., Webb, G. I.: Data Preparation, pp. 318–327. Humana Press, Boston (2017)
[8] Hrynaszkiewicz, I., Norton, M. L., Vickers, A. J., Altman, D. G.: Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. The BMJ 340 (2010). doi: DOI:10.1186/1745-6215-11-9
[9] Yang, B., Nazari, R., Elmo, D., Stead, D., Eberhardt, E: Data preparation for machine learning in rock engineering. In: IOP Conference Series: Earth and Environmental Science, vol. 1124, 12072 (2023). doi: DOI:10.1088/1755-1315/1124/1/012072
[10] Walport, M., Brest, P.: Sharing research data to improve public health. The Lancet 377, 537–539 (2011)
[11] Brand, D., Singh, J. A., McKay, A. G. N., Cengiz, N., Moodley, K.: Data sharing governance in sub-Saharan Africa during public health emergencies: Gaps and guidance. South African Journal of Science 118(11–12), 1–6 (2022)
[12] Julius, M., Alo, R. U., Onu, F. U., Iganga, O., Eze, C. N., Abanobi, G., Okpan, S. O.: Survey Data on the knowledge, belief, behavioral pattern, and perception towards medication non-adherence among outpatients with chronic illness in AE-FUTHA, Nigeria (2022). doi: 10.17632/9gbpwk97zf.1
[13] Julius, M., Onu, F., Okorie, K., Alo, U., Eze, C., Abanobi, G.: Multi-Dimensional patients' dataset for: Development of online assessment and machine learning prediction of medication non-adherence with intervention system (2023). doi: 10.17632/3t39fctrzv.1
[14] Prell, T., Grosskreutz, J., Witte, O., Kunze, A., Franke, G. H., Mendorf, S.: Data for: Clusters of non-adherence to medication in neurological patients (2019). doi: DOI: 10.17632/ny2krr3vgg.1
[15] Dhakal, B.: Working data for adherence to anti-hypertensive therapy (2022). doi: 10.17632/k7xchzr33y.1
[16] Sanders-Pinheiro, H., Carminatti, M., Ferreira, R., Paula, R. de, Pinhati, R., Colugnati, F. A.: Data for: Adherence to antihypertensive medication in severely hypertensive patients after referral to secondary health care: a prospective cohort study (2020). doi: doi: 10.17632/8tw3c4jw5n.1
[17] Awwad, O., AlMuhaissen, S., Al-Nashwan, A., AbuRuz, S.: Translation and validation of the Arabic version of the Morisky, Green and Levine (MGL) adherence scale. PLoS One 17(10), e0275778 (2022)
[18] Tegegn, H. G., Spark, M., Wark, S., Tursan D'espaignet, G. D. E.: Medication adherence among people with cardiovascular disease: A multivariable predictive model development and validation. Phd dissertation. University of New England (2023). Available at: https://hdl.handle.net/1959.11/55620. Accessed 3 January 2024
[19] Koesmahargyo, V., Abbas, A., Zhang, L., Guan, L., Feng, S., Yadav, V., Galatzer-Levy, I.R.: Accuracy of machine learning-based prediction of medication adherence in clinical research. Psychiatry Research. 294, 113558 (2020)
[20] Dima, A. L., Dediu, D.: Computation of adherence to medication and visualization of medication histories in R with AdhereR: towards transparent and reproducible use of electronic healthcare data. PLoS One 12(4), e0174426 (2017)
[21] Gu, Y., Zalkikar, A., Liu, M., Kelly, L., Hall, A., Daly, K., Ward, T.: Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data. Scientific Reports 11(1), 18961 (2021)

[22] Lauffenburger, J.C., Yom-Tov, E., Keller, P.A., McDonnell, M.E., Bessette, L.G., Fontanet, C.P., Sears, E.S., Kim, E., Hanken, K., Buckley, J.J. Barlev, R.A.: REinforcement learning to improve non-adherence for diabetes treatments by Optimising Response and Customising Engagement (REINFORCE): study protocol of a pragmatic randomised trial. BMJ Open 11(12), e052091 (2021)

[23] Park, H. R., Kang, H. S., Kim, S. H., Singh-Carlson, S.: Effect of a smart pill bottle reminder intervention on medication adherence, self-efficacy, and depression in breast cancer survivors. Cancer Nursing. 45(6), 874–882 (2022)

[24] Ellsworth, G.B., Burke, L.A., Wells, M.T., Mishra, S., Caffrey, M., Liddle, D., Madhava, M., O'Neal, C., Anderson, P.L., Bushman, L. Ellison, L.: Randomized pilot study of an advanced smart-pill bottle as an adherence intervention in patients with HIV on antiretroviral treatment. JAIDS Journal of Acquired Immune Deficiency Syndromes 86(1), 73–80 (2021)

[25] Lo-Ciganic, W.H., Donohue, J.M., Thorpe, J.M., Perera, S., Thorpe, C.T., Marcum, Z.A., Gellad, W.F.: Using machine learning to examine medication adherence thresholds and risk of hospitalization. Medical Care 53(8), 720-728 (2015)

[26] Gao, W., Liu, H., Ge, C., Liu, X., Jia, H., Wu, H., Peng, X.: A clinical prediction model of medication adherence in hypertensive patients in a Chinese community hospital in Beijing. American Journal of Hypertension 33(11), 1038–1046 (2020)

[27] Lucas, J. E., Bazemore, T. C., Alo, C., Monahan, P. B., Voora, D.: An electronic health record based model predicts statin adherence, LDL cholesterol, and cardiovascular disease in the United States Military Health System. PLoS One 12(11), e0187809 (2017)

[28] Hameed, M., Naumann, F.: Data preparation: A survey of commercial tools. ACM SIGMOD Record 49(3), 18–29 (2020)

[29] Azeroual, O., Schöpfel, J., Ivanovic, D., Nikiforova, A.: Combining data lake and data wrangling for ensuring data quality in CRIS. Procedia Computer Science 211, 3–16 (2022)

[30] Koehler, M., Bogatu, A., Civili, C., Konstantinou, N., Abel, E., Fernandes, A.A., Keane, J., Libkin, L., Paton, N.W.: Data context informed data wrangling. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 956–963 (2017)

[31] Ali, S. M. F., Wrembel, R.: From conceptual design to performance optimization of ETL workflows: current state of research and open problems. The VLDB Journal 26(6), 777–801 (2017)

[32] M Dahiya, M., Malik, N., Rana, S.: Essentials of Data Wrangling. Applications and Tools, pp. 71–90 (2023)

[33] Rattenbury, T., Hellerstein, J. M., Heer, J., Kandel, S., Carreras, C.: Principles of data wrangling: Practical techniques for data preparation. O'Reilly Media, Inc. (2017)

[34] Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Van Ham, F., Riche, N.H., Weaver, C., Lee, B., Brodbeck, D., Buono, P: Research directions in data wrangling: Visualizations and transformations for usable and credible data. Information Visualization 10(4), 271–288 (2011)

[35] Terrizzano, I. G., Schwarz, P. M., Roth, M., Colino, J. E.: Data Wrangling: The Challenging Yourney from the Wild to the Lake. In: Conference on Innovative Data Systems Research, Asilomar (2015)

[36] Guo, M., Wang, Y., Yang, Q., Li, R., Zhao, Y., Li, C., Zhu, M., Cui, Y., Jiang, X., Sheng, S., Li, Q.: Normal Workflow and Key Strategies for Data Cleaning Toward Real-World Data. Interact. Journal of International Medical Research 12(1), e44310 (2023)

[37] Furche, T., Gottlob, G., Libkin, L., Orsi, G., Paton, N. W.: Data wrangling for big data: Challenges and opportunities. In: 19th International Conference on Extending Database Technology, pp. 473–478 (2016)

[38] IBM: Prepare data for machine learning:Improve data quality with data cleansing, data transformation, and feature engineering (n.d.). Available at: https://www.ibm.com/garage/method/practices/reason/prepare-data-for-machine-learning/. Accessed 15 January 2024

[39] Abdallah, Z. S., Du, L., Webb, G. I.: Data Preparation in Encyclopedia of Machine Learning and Data Mining. Humana Press, New Jersey (2016)
[40] Xiao, F., Fan, C.: Data mining in building automation system for improving building operational performance. Energy and Buildings 75, 109–118 (2014). doi: http://dx.doi.org/10.1016/j.enbuild.2014.02.005.
[41] Bosagh Zadeh, R., Meng, X., Ulanov, A., Yavuz, B., Pu, L., Venkataraman, S., Sparks, E., Staple, A., Zaharia, M.: Matrix computations and optimization in apache spark. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 31–38 (2016)
[42] Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A.B., Alzakari, N., Abou Elwafa, A., Kurdi, H.: Impact of dataset size on classification performance: an empirical evaluation in the medical domain. Applied Sciences 11(2), 796 (2021)
[43] Prusa, J., Khoshgoftaar, T. M., Seliya, N.: The effect of dataset size on training tweet sentiment classifiers. In: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), pp. 96–102 (2015)
[44] Rahman, M. S., Sultana, M.: Performance of Firth-and logF-type penalized methods in risk prediction for small or sparse binary data. BMC Medical Research Methodology 17, 1–15 (2017)
[45] Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in NLP. arXiv Preprint arXiv1906.02243 (2019)
[46] Olawade, D. B., Wada, O. J., Ling, J.: Using artificial intelligence to improve public health: a narrative review. Frontiers in Public Health 11, 1196397 (2023)
[47] Colliot, O., Thibeau-Sutre, E., Burgos, N.: Reproducibility in machine learning for medical imaging. In: Colliot, O. (eds.) Machine Learning for Brain Disorders, pp. 631–653 Humana Press, New York (2023)
[48] Mahony, S.: Toward openness and transparency to better facilitate knowledge creation. Journal of the Association for Information Science and Techonlogy 73(10), 1474–1488 (2022)

Data Wrangling and Generation for Machine Learning Models in Medication Adherence Analytics: A practical Standpoint using Patient-Level and Medical Claims Data

AUTHOR BIOGRAPHY

WELLINGTON KANYONGO earned an MSc in Information Systems and an MSc in Strategic Management from the National University of Science and Technology and Chinhoyi University of Technology, respectively. He also holds a BSc in information systems from Midlands State University. He is currently pursuing and finalising his PhD in computer science at the University of Kwazulu-Natal. His research interests are in applied data science, particularly the application of machine learning algorithms for big healthcare and biomedical data analytics, as well as problem-solving in health informatics and bioinformatics.
Email:

4.2 Chapter Summary

The conclusion of this chapter is a distinctive dataset generated by utilising EHRs for patients with NCDs in Zimbabwe, a developing country. This dataset is the first of its kind in Zimbabwe and is designed to assess the level of medical adherence in patients with diabetes and hypertension using modern technology and methods. The chapter provides usage notes to guide the use of the generated dataset in ML modeling, including optimal methods for data manipulation and any challenges to consider. By and large, this chapter provides a detailed explanation of each process involved in DW, which is essential for creating robust and reliable machine-learning models. This chapter ensures a robust foundation for the subsequent development of ML models in the next chapter (Chapter 5), leading to more accurate and reliable predictions and classifications of MA analytics through meticulous data preparation.

CHAPTER 5

Development and validation of ML models for medication adherence analytics (Paper 4)

5.1 Introduction

Chapter five (5) entails the actual development, validation, discussion of interpretability, and explainability of ML models specifically built to classify the MA of patients with NCDs to their prescribed medications. This chapter commences with delineating the process of choosing appropriate ML algorithms, such as LR, DT, RF, support vector machines, naïve bayes, k-NN, and DNN. Each method is selected based on its appropriateness for addressing the classification problem at hand around MA analytics, the intricacy and size of the dataset, and the capability to represent the non-linear connections inherent in the healthcare data. The chapter proceeds with the process of training the model, which begins by partitioning the dataset into training and testing sets. This enables the evaluation of the ML models with previously unseen data. Methods such as cross-validation were used to improve model parameters and avoid overfitting, ensuring that the models can effectively generalise to new unknown data. The RF feature importance mechanism was used to assess the relevance of features, while the SHAP beeswarm plots were used to get insights into the contribution of each feature to the model's predictions. Not only does this improve the ability to understand and explain the model, but it also helps in explaining the key variables that impact MA.

The chapter dedicated significant attention to evaluating the performance of the developed models. The models' effectiveness was assessed using various ML model evaluation metrics, such as accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic (ROC) curve. Given the imbalanced nature of the dataset, with a higher frequency of non-adherence cases compared to adherence cases, it was important to utilise weighted averages for precision, recall, and F1-score. This ensures a fair evaluation across different classes. The chapter addresses the trade-offs between these measures and emphasises how crucial it is to choose the appropriate evaluation criteria in light of the particular circumstances surrounding the MA data. In addition, the confusion matrix was

employed to present detailed analytics of true positives, true negatives, false positives, and false negatives, thus, providing a more profound understanding of the performance of the developed model performance. ROC curves were computed to visually evaluate the balance between sensitivity and specificity at various threshold values. This helps in choosing the most suitable model for practical implementation.

The models were subjected to a vigorous validation process to guarantee model reliability and robustness. In order to improve the robustness of the model, a technique known as 5-fold cross-validation was utilised. This technique entails dividing the dataset into five separate subsets and then repeatedly using four subsets for training the model and one subset for testing its performance. The process was iterated five times, wherein each subset was used as the test set once. This ensures a full evaluation of the model and prevents its performance from being excessively reliant on any particular group of data. The thesis concludes this chapter with presentation and comparison of the various developed ML models which provide useful and precise classifications of MA. Healthcare practitioners can use these models to identify patients who are likely to not follow medical advice, enabling specific treatments that can improve health outcomes and reduce the economic impact of medication non-adherence among NCD patients. The rigorous methodology employed for developing and validating models guarantees that the results are reliable, robust, and suitable for practical use in the healthcare field.

Paper status: Under Review –Second Revision

Journal Name: Scientific Reports

Journal impact factor: 5.516 (Q1)

Student's contribution: conception, planning, and design of the study, coding, data analysis, data curation, original drafting of the manuscript, and review.

Other authors' contributions: Data acquisition, supervision, editing, and revision

Percentage of student's contribution: 90%

Machine Learning Classification of Medication Adherence in NCD Patients: A Retrospective Study

Wellington Kanyongo[1], Absalom E. Ezugwu[2], Tsitsi Moyo[3], Jean Vincent Fonou Dombeu[1]

[1] *School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg Campus, Pietermaritzburg, 3201, South Africa.*
[2] *Unit for Data Science and Computing, North-West University, 11 Hoffman Street, Potchefstroom, 2520, South Africa.*
[3] *Cimas Medical Aid Society, Borrowdale Office Park, Borrowdale Road, P. O Box 1243, Harare, Zimbabwe*

Corresponding author's email addresses: absalom.ezugwu@nwu.ac.za

Abstract. Non-adherence to medication among individuals with non-communicable diseases (NCDs) leads to increased morbidity, mortality, and healthcare costs. The integration of electronic drug prescription and dispensation systems enables comprehensive analysis of medication adherence (MA). Patient-level and medical claims data for 8141 diabetic and hypertensive patients in Harare, Zimbabwe, were analyzed. Non-adherence was defined as medication refills falling below 75% of the intended 12 monthly claims, while adherence required at least 75% of the refills. Classification employed multiple machine learning algorithms, including SVM, KNN, DT, Naïve Bayes, DNN, LR, and RF in Python 3.11.3. Significant variables for MA were identified through the Random Forest (RF) feature importance mechanism. These included the annual quantity of medical supplies, annual claim amount, patient age, wellness program subscription, medical aid cover, contribution towards medical aid cover, comorbidity, diagnosis, hospital cover type, complications development, gender, and medical aid scheme. The total units of medical supplies dispensed annually emerged as the most significant predictor of MA. Classification accuracy of the ML classifiers ranged from 84.7% to 87.6%, with AUC values from 0.857 to 0.934. RF, an ensemble learning technique, was the most robust classifier, achieving 87.6% accuracy, an AUC of 0.9351, and superior precision, recall, and F1-score. This model shows promise as a prognostic tool for enhancing MA, aiding in identifying adherence levels among patients. These findings contribute to addressing disparities in medication refilling and adherence rates among patients with NCDs. The ML model holds potential for the development of intelligent MA and intervention applications to improve patient adherence to medication in the chronic disease domain.

Keywords: Machine learning, medication adherence, non-communicable diseases, chronic disease, diabetes, hypertension, deep neural network, SVM, logistic regression, naïve bayes, decision trees, random forest, logistic regression, k-NN.

Nomenclature Descriptions

AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Programming Interface
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BART	Bidirectional and Auto-Regressive Transformer
CART	Classification and Regression Trees
CGM	Continuous Glucose Monitoring
CHAID	Chi-square Automatic Interaction Detector
CNN	Convolutional Neural Networks
CoLT	Computational Learning Theory
DNN	Deep Neural Network
DT	Decision Tree
EDA	Exploratory Data Analysis
EHR	Electronic Health Records
EMR	Electronic Medical Records
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
GaussianNB	Gaussian Naive Bayes
GB	Gradient Boosting

GDM	Gestational Diabetes Mellitus
HbA1c	Hemoglobin A1c / Glycated Haemoglobin
ICD	International Classification of Diseases
ICD-10	International Classification of Diseases 10th Revision
IDE	Integrated Development Environment
IDF	International Diabetes Federation
IoT	Internet of Things
k-NN	K-Nearest Neighbour
LR	Logistic Regression
LSTM	Long Short-Term Memory Networks
LSVM	Linear Support Vector Machine
MA	Medication Adherence
ML	Machine Learning
MLP	Multi-Layer Perceptron
MPR	Medication Possession Ratio
MSE	Mean Squared Error
NCD	Non-Communicable Diseases
OD	Odds Ratio
QUEST	Quick Unbiased Efficient Statistical Tree
PDC	Proportion of Days Covered
PCA	Principal component analysis
RBF	Radial Basis Function
ReLU	Rectified Linear Unit
RF	Random Forest
ROC	Receiver Operating Characteristic Curve
RMSE	Root Mean Square Error
SAS	Statistical Analysis System
SD	Standard Deviation
SSB	Smart Sharps Bin
SHAP	SHapley Additive exPlanations
SOM	Self-Organizing Map
ST	Survival Trees
SVM	Support Vector Machines
SVR	Support Vector Regression
T1D	Type 1 Diabetes
T1DM	Type 1 Diabetes Mellitus
T2D	Type 2 Diabetes
TN	True Negatives
TP	True Positives
TPS	Transactional Processing System
TPR	True Positive Rate
XGBoost	Extreme Gradient Boosting
WHO	World Health Organisation

1. Introduction

Non-communicable diseases (NCDs) are responsible for the mortality of 41 million individuals per year, accounting for approximately 74% of total global deaths [1]. These diseases arise from a complex interplay of genetic, physiological, environmental, and behavioural variables [2]. Thus, the focus of this study is on a population with NCDs, not only because they account for the majority of total global deaths but also because they place a significant economic burden on family and national budgets as NCD patients become less productive, work for fewer years and die prematurely [3]–[5]. It is vital to focus on hypertensives and diabetics, as the number of people with hypertension has risen to 1.28 billion in the last two decades [6]. Globally, hypertension is the primary cause of cardiovascular illness and early mortality [7]. On the other hand, a minimum of 422 million individuals globally suffer from diabetes, with the majority residing in low- and middle-income nations. The disease is directly responsible for 1.5 million fatalities annually [1]. Amidst the alarmingly unfavourable statistics surrounding non-communicable diseases (NCDs), non-adherence to the medication for NCDs such as hypertension and diabetes has been identified as a significant factor contributing to the resulting repercussions of these diseases [8], [9]. A WHO study on drug adherence for chronic illnesses indicates that patients in developed countries only take 50% of their prescribed medications; in developing countries, this number is even lower [10].

Accurately identifying non-adherence early on is essential for allocating resources to those who require more care and for personalising the intervention to address the unique causes of non-adherence to medication.

With the availability of vast EHR datasets and a shortage of medical professionals in the healthcare sector, the process of examining a patient's medical history and assessing outcomes would be time-consuming [11]. However, ML in the field of medicine and public health offers the potential to address various healthcare obstacles by effectively managing vast and varied data, thereby offering real-time and cost-efficient evidence [12]. While trying to extract actionable insights and predictions from these kinds of data sets, scientists and engineers are shifting towards leveraging advanced analytics and ML-related suite of techniques to make accurate and informative decisions [13]. [14] highlight that contemporary clinical decision support tools that incorporate ML algorithms have demonstrated significant promise in various areas, such as disease classification or profiling, risk prediction, dose recommendation, pattern recognition, precision treatment, and health monitoring. In contrast to more widely used statistical methodologies, big data spotlights challenges and datasets for which ML algorithms perform exceptionally well as compared to traditional statistical approaches. Various ML modelling techniques have been employed to assess MA in individuals with NCDs such as diabetes and hypertension. [15] undertook a comparative examination of ML-based classification models in order to assess the occurrence of treatment default in the context of various diseases. The anticipated accuracy ratings of four supervised ML methods, specifically GB, LR, RF, and SVM, were 0.87, 0.90, 0.81, and 0.77, respectively.

In the further empirical literature, three studies developed MA models using decision trees (DTs) alone [16]–[18]. [16] used CART to build DTs for an ML model to predict poor MA in elderly hypertensives. In that study, the validated CART model accuracy showed 26% misclassification. [17] used hypertension data to construct a clinical prediction model for MA using DTs. Unlike [16] who used CART, [18] used the CHAID algorithm to identify demographic factors connected to antihypertensive non-adherence. In contrast to [16], [18] found household income and neighbourhood poverty as risk variables for MA levels in elderly hypertensives. Similar to [18], [17] shows that concomitant diabetes increases MA adherence risk. Though [19] applied more than one ML algorithm to assess MA, their study focused on hypertensive people as well. In their, study, the Wilcoxon signed ranked test showed no statistically significant difference between actual scores and ML model predictions. Another study involving hypertensives by [20] examined hypertensive prescription refill adherence using electronic medical records and dispensing data, focusing on predictive and pattern analysis. With baseline predictors and past prescriptions, the RF and GB algorithms outperformed LR and k-NN on both validation and test sets. The best model is RF, with AUC values of 0.90–0.91 [20].

Five past studies used real-time MA measurement data with ML [21]–[25]. [21] used CNN and CGM data to identify once-daily basal insulin injection non-adherence. Their findings suggest that CGM data can be used to predict adherence, especially with additional data on the classification day. Similar to Mohebbi et al. (2017), deep learning algorithms were used to detect T2D adherence using CGM signals. They found that their best CNN model had 77.5% accuracy. In contrast, [23] used a smart sharps bin (SSB) and ML algorithms (RF, extreme GB (XGBoost), GB, and MLP) to predict injectable MA. HealthBeacon Ltd.'s "SSB," an IoT device, tracked injection disposal at patients' homes. The analysis predicted 81.3% of patients not taking their prescriptions on time using HealthBeacon SSB data. [24] created a user-friendly medication intake monitoring system using an Android wearable app and a cloud-based data pipeline to improve MA. The study achieved a high F1 score of 0.977 using the Android smartwatch app, which collected activity sensor data using an accelerometer and gyroscope. In another work, [25] created an interactive electronic system to track diabetes patients using ML. The software used Bluetooth movement data from wearable insole devices to track MA, carbohydrate intake, blood glucose levels, and physical activity. For these tasks, SVM and DT models were 86% accurate.

Existing literature also houses ML-based predictive models for diabetic MA or non-adherence [26]–[30]. Three studies—[26], [29], [30] focused on type 2 diabetes (T2D) patients at Sichuan Provincial People's Hospital in China. The recent study by [26] employed ANN, rapid unbiased efficient statistical tree (QUEST), Bayes network, discriminant analysis, CART, and ensemble models to predict risks of diabetic complications and poor glycaemic control in non-adherent T2D, while [29] used XGBoost, RF, LR, SVM, GB, k-NN, DT, Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, and ensemble learning to predict patients' non-adherence risks, whereas [30] compared Bayesian networks, neural networks, SVM, LR, k-NN, least squares support vector machines (LSVM), RF, DT, and ensemble models; all came to the same conclusion that the ensemble learning models yield superior AUC scores. [27] applied ML to predict diabetes treatment dropout. Integrating ML with

the EHR created a TD risk score for diabetics. The score outperformed the Cox regression model in prediction. [28] used various supervised ML algorithms to predict type 1 diabetes self-care and glycaemic management in a proof-of-concept study. All ML models had similar discriminatory strength; however, the RF model had better calibration.

A discrepancy remains that necessitates additional investigation, as [31] asserts that a clear distinction between ML models and conventional statistical models does not exist. Again, the presence of conflicting findings within the current body of literature regarding the efficacy of different ML models in predicting MA highlights the need for further research in this area. Research by [15] compares the performance of different models and reports that LR achieved the highest predicted score of 0.90, followed by GB (0.87) and LR (0.87), while RF had the lowest score of 0.77. In contrast, the study conducted by [32] revealed that SVM exhibited significantly higher classification accuracy compared to LR and back-propagation neural networks (BPNN). In contrast, Lee et al. (2013) found that SVM exhibited higher performance with a classification accuracy of 97.3% in comparison to LR, which achieved a classification accuracy of 71.1%. LR was surpassed by RF in research aimed at predicting the likelihood of individuals discontinuing a therapy programme [15], [34].

An alternative viewpoint with conflicting research outcomes found that race, ethnicity, and gender do not significantly predict treatment discontinuation [35], yet, [18] discovered that race or ethnicity is a significant factor in predicting non-adherence to antihypertensive drugs. Additionally, [36] and [37] found that gender can be used to predict medical adherence in patients with high fever and opioid use disorder, respectively. There were divergent results about how gender can predict MA in patients with the same chronic illness. This shows that differences in the study setting, time frame, and participant demographics affect the results of the research. While [35] found no significant impact of age on model prediction accuracy, other literature demonstrated that younger individuals are more prone to non-adherence to prescribed treatment plans [38]. It is imperative to acknowledge that the factors influencing MA vary across different research environments. The variation in the susceptibility to different types of NCD under investigation can be attributed to several factors, including the techniques used for feature selection, the specific study setting, cultural and regional differences, the level of economic development, education level, the presence of social desirability bias, differences in living habits, residing in urban or rural areas, variations in social and healthcare systems, and other relevant factors [39]–[41]. These challenges pose difficulties in the development of accurate adherence prediction models and patient treatment programmes. Consequently, it is imperative that MA be analysed in this study using a number of ML algorithms, additional unique features, and a unique Zimbabwean dataset to classify MA among patients with diabetes and hypertension.

i. To establish features that predict MA in diabetic and hypertensive patients.
ii. To develop ML models for predicting MA in diabetic and hypertensive patients using K-NN, SVM, RF, DNN, DT, LR, and Naïve Bayes.
iii. To compare and assess the accuracy of K-NN, SVM, RF, DNN, DT, LR, and Nave Bayes models for classifying MA in patients with diabetes and hypertension.

By addressing the aforementioned research objectives, the following technical contributions are made:

- The study involves the creation of seven ML models to predict MA in patients with diabetes and hypertension. This process included identifying the most effective ML algorithms based on the specific dataset and variables that were chosen. The models that have been developed play a crucial role in facilitating clinical decision-making by accurately identifying instances of non-adherence. This enables healthcare professionals to prioritise patients who require more attention and tailor interventions to address the specific circumstances that contribute to medication non-adherence.
- Feature engineering which is the process of selecting relevant predictors for MA enables the identification of important features that patients, carers, and healthcare professionals should prioritise in order to enhance patients' health outcomes.

2. Machine Learning Theory

The foundation of this research is rooted in the theory of ML, also known as computational learning theory (CoLT). ML is a field that exhibits a strong interdisciplinary nature, drawing upon concepts and principles from several domains such as cognitive science, computer science, statistics, and optimisation, among others [42]. The objectives of this theory encompass two main aspects: enhancing the development of effective automated learning systems and gaining comprehensive insights into the fundamental intricacies of the learning process. The fundamental difference between traditional statistical techniques and ML lies in the fact that ML, as a discipline within computer science, seeks to facilitate the learning of models or computers from data, as opposed to being explicitly coded with rules [43]. An ML algorithm typically represents its data in terms of features. For instance,

our EHR dataset is represented by patient-level features, disease-related features, and medication information (such as MA thresholds). After processing this data, the ML learning system predicts certain outcomes with performance measures, such as which patient is taking their prescription as prescribed. Therefore, ML techniques can be employed in the development of MA prediction models that offer assurances of their efficacy. The selection of an algorithm is contingent upon the nature of the problem to be addressed, the number of variables involved, and the most suitable model for the problem at hand, among other factors. ML can be categorised based on whether the computer's learning or model-fitting process is supervised or unsupervised. The diagram presented in Figure 1 illustrates the comprehensive categorisation of ML and the specific tasks that are encompassed under each category.

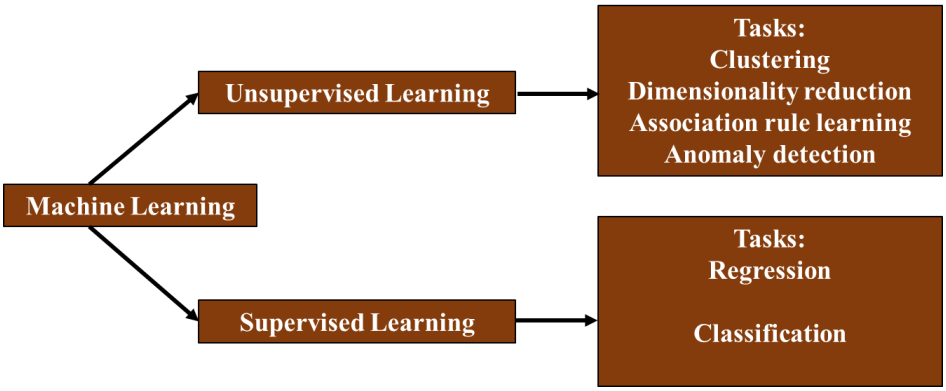


Figure 1: Broad ML taxonomies and tasks

2.1 Unsupervised Learning

There has been a resurgence of interest in unsupervised (or self-supervised) learning algorithms due to the aim to lessen reliance on curated, labelled datasets and exploit the extensive amounts of unlabelled data [44]. In a general sense, unsupervised learning encompasses the analysis of unlabelled data while operating on assumptions about the inherent structural features of the data [45]. The objective of unsupervised learning is to ascertain the hidden patterns inherent in a given dataset, cluster the data points based on their common characteristics, and then visualise the dataset in an organised way. According to [46], the essential requirements for conducting unsupervised ML involve several key steps including the careful selection of samples or objects that will be subjected to clustering, the identification and selection of specific features that will be utilised in the clustering process, the determination of an appropriate similarity metric for comparing and clustering the selected samples, and the choice of an algorithm to be employed in the analytics. Clustering, dimensionality reduction, association rule learning, and anomaly detection are among the most widely used unsupervised learning tasks. Dimensionality reduction is a process that strives to reduce the number of input features in a given dataset, whereas clustering is a data mining methodology that involves the segmentation of unlabelled data points according to their shared similarities or dissimilarities.

2.2 Supervised Machine Learning

Supervised ML involves the training of machines using accurately labelled training data, which serves as the basis for the machines' output predictions. Labelled data refers to input data that has been annotated or labelled with the correct outcome or response. The dependent variable, commonly referred to as the label in supervised ML, is predetermined for each observation, thereby indicating the value of the outcome. In the context of a supervised ML, the task focused on predicting drug adherence, the potential outcomes can be categorised as either adherence (designated as 1) or non-adherence (designated as 0). The model undergoes training until it acquires the ability to discern the fundamental adherence patterns and connections between the input data and the output labels. This enables the model to produce precise labelling outcomes when presented with previously unseen data containing input and output data. The inputs in this context are commonly referred to as features or 'X variables', whereas the output is generally denoted as the target or 'y variable'. The objective of a supervised learning algorithm is to identify a mapping function that establishes a relationship between the input variable (x) and the output variable (y). The training dataset consists of a set of (x, y) pairs, where x represents the input query and y represents the corresponding output. The objective is to provide a prediction y* for a given query x*. Supervised learning systems often generate predictions by utilising a learned mapping function f(x), which yields an output y for each input x [45]. The diagram in Figure 2 illustrates the workflow of supervised ML algorithms.

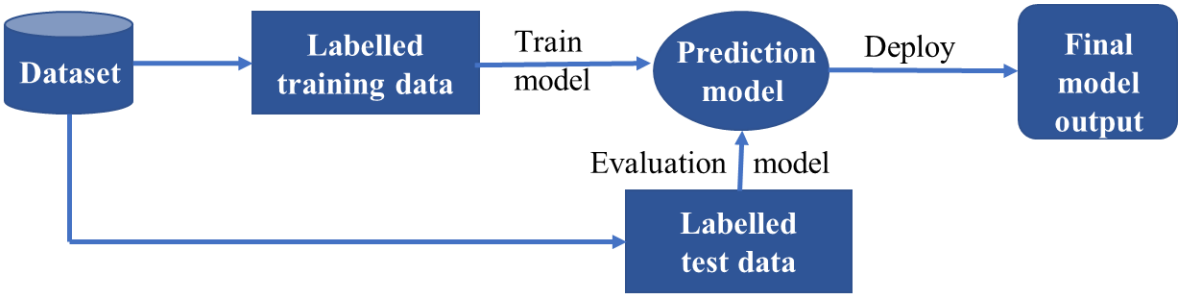


Figure 2: Workflow of supervised ML

The process of supervised ML involves many key elements. Firstly, it is crucial for the data analyst to carefully select a dataset that is most relevant to the problem at hand. Additionally, the data must be appropriately tagged or labelled. The data analyst needs to verify that the datasets possess input features that can be effectively mapped by the ML model to known labelled output values. The subsequent procedure involves dividing the data into two distinct sets: a training set and a test set. The test set is used to evaluate the model's performance rather than train it. The subsequent stage involves the identification of an appropriate ML algorithm for the model. An ML model is trained by feeding training data to the learning algorithm. During the training phase, the prepared dataset is fed into the ML model to identify underlying patterns and generate predictions. The process of training involves multiple iterations, with each iteration being followed by an evaluation to assess the performance of the model. After the completion of the training process, it becomes possible to assess the performance of the model in predicting data that has not been seen before by measuring the model's generalisation error. Real-world applications of supervised learning include the prediction of medication non-adherence risk, image classification, regression, and many other classification tasks. Classification, a supervised learning technique in ML, involves the analysis of a provided dataset to create a model that effectively classifies the data into an identifiable and distinctive set of classes. Classification algorithms are employed in situations where the output variable is categorical. This is particularly relevant in the current study in which there are two distinct class values: "adherent" (which could be labelled as 1) and "non-adherent" (which could be labelled as 0). Common examples of supervised ML algorithms are ANN, DT, LR, SVM, RF, Naive Bayes, and k-NN.

3. Materials and Methods

3.1 The Dataset

The data used covered the period spanning from January 1 to December 31, 2022. Patients in the study are those who were part of the CIMAS transactional processing system (TPS), the system used to collect and manage member medical claims as well as patient-level data. The CIMAS TPS encompasses a comprehensive collection of data pertaining to medical and pharmacy claims. This data pertains to individuals who are beneficiaries covered by either employer-sponsored or individually-sponsored private health insurance, including the members themselves, their spouses, and dependents. The inclusion criteria for this study encompassed individuals who met the diagnostic criteria for essential diabetes and hypertension as defined by the ICD-10. Additionally, these patients were required to have been recruited for the study prior to January 1, 2022. Participants below the age of 16 and individuals with chronic conditions other than diabetes and hypertension were excluded from the study. The availability of data related to monthly diabetes and hypertension medication refills or claims data in CIMAS's TPS makes it possible to evaluate medication refill adherence. This is achieved by utilising medication refill thresholds, monthly medication collection data, and patient-related features. The original dataset, formatted in CSV, consisted of 45 columns (features) and slightly over 1.7 million rows of data. A quick exploration of the dataset led to the removal of unnecessary columns before proceeding to the full-fledged data pre-processing step. The 12 relevant features included in the extracted and designated dataset are diagnosis, age, hospital cover type, scheme type, annual contribution, subscription in a wellness programme, gender, annual claim amount, comorbidity, development of complications, units (number of all different medical supplies or units given to a patient annually), and MA thresholds. The evaluation of adherence in our study encompasses a 12-month timeframe, with particular emphasis on the number of monthly claims made by the patients of interest. This evaluation is based on their medical claims data or dispensation data, which were accessible for the period spanning from January 1, 2022, to December 31, 2022. A description of the chosen relevant variables may be found in Table 1. Prior to utilising the dataset for modelling purposes, various pre-processing steps were performed.

3.2 Data Pre-processing

Variables related to internal system processing, as well as other columns deemed unnecessary for our analytics, including the anonymised membership number, practice number of pharmacies that supplied medication to CIMAS members, medical claims invoice number, claims code, and other irrelevant columns, were discarded. The initial processing of the dataset also encompassed the reclassification of the data into more refined categorical variables within Excel prior to its importation into Python. The pre-processed dataset housing the available, usable, and relevant features was imported into Python. Patients who had annual refill thresholds less than 9 were categorised as non-adherent (coded as 0), whereas those with a threshold of 10 or more medical claims or refills associated with diabetes and hypertension medication were classified as adherent (coded as 1). The typical duration of medication regimens spans one month within a 12-year timeframe, although this timeframe may be subject to variation due to premature discontinuations or extensions. It is imperative to acknowledge the importance of utilising annual data, as it provides an accurate representation of adherence or non-adherence across a span of 12 months. The dataset consists of 8141 unique rows of data for the year 2022. Table 1 below describes the 12 variables in the initial dataset.

Regarding the coding of gender data as part of the pre-processing phase, males were assigned a code of 1 while females were assigned a code of 2. Other variables include comorbidity (coded as 1 for yes and 2 for no), complication development (coded as 1 for yes and 2 for no), scheme type (coded as 1 for basic level, 2 for medium level, and 3 for premium level), hospital cover type (ranging from basic to grade B and private grade A), wellness programme (coded as 1 for the subscribed and 2 for the non-subscribed), and MA (coded as 1 for adherent and 0 for non-adherent). To ensure the integrity of the data, all instances of duplicate records pertaining to individual patients were eliminated. This was done following the requirement that each patient have only a single distinct record with annual totals. The imported data had no missing values. Since there were no missing values for the chosen variables in the dataset, imputation of the data was not necessary. Outliers in the numerical data were standardised to get them closer to zero and within the same range. This makes the data less sensitive to extreme values and the optimisation problem more numerically stable. As a component of the data preparation procedure, exploratory data analysis (EDA) was conducted on the initially pre-processed dataset prior to data encoding and feature scaling. With the use of descriptive statistics like frequency count, mean scores, median, interquartile ranges, modal scores, and SD, as well as early investigations to identify patterns and anomalies, EDA assisted us in making sense of the data we had on hand. It also made it possible for us to use a correlation matrix to visually represent the relationships between the features in the dataset. Once again, as a component of the data preparation procedure, categorical data were transformed and numerical variables were scaled.

Table 1: Description of the selected variables and pre-processing procedures

Variable	Description	Pre-processing	Data type
Diagnosis	The medical condition afflicting the patient. The medications the patient is claiming on a monthly basis are used to detect the diagnosis. The focus of this study was on individuals diagnosed with diabetes and hypertension.	Given the diverse range of diseases present in the dataset, a process was undertaken to refine the data, specifically focusing on two specific diagnoses, namely diabetes and hypertension. To specifically identify cases of diabetes and hypertension, the monthly prescription claims were used as a filtering mechanism to exclude any rows that did not correspond to diagnoses of diabetes and hypertension. From an alternative standpoint, the medical records of individuals diagnosed with diabetes and hypertension were examined by utilising the ICD10 codes (E10- and i10-).	Categorical
Age	Age of the patient/member	Filtered out ages under 16 years. It is standard procedure to need parental approval before treating a youngster younger than sixteen.	Numerical
Scheme type	The actual cover or grade one gets for medical costs.	The dataset initially consisted of nine distinct scheme types or packages, namely secure primary, secure premium, secure prime, secure essential, essential, classic, deluxe, healthsave, and secure private. The data was categorised into three distinct groups, namely basic level, medium level, and premium level, based on the actual coverage or grade received for medical expenses and the annual contributions made towards each programme.	Categorical
Hospital cover type	Whether one gets a healthcare package or access covering private hospital access up to Grade A or Grade B.	This variable was not initially present in the dataset that was accessible. The variable was recreated and patients or members were grouped into two categories: 'basic to grade B' and 'private to grade A', based on the scheme types.	Categorical
Annual contribution	Payments made to the medical aid society to cover the medical expenses of employees, their close relatives and dependents.	This variable was not originally present in the dataset that was accessible. The data for this variable was generated by recreating it using the monthly contributions made by members, which were categorised according to their scheme types of subscription. The monthly contributions were aggregated to form annual contributions. It was posited that the yearly contribution made by the primary account holder is equivalent to that made by the beneficiaries under their supervision.	Numerical
Wellness programme	Whether one has subscribed or not to a wellness programme over and above the accessed primary service (medical aid cover).	Individuals who enrolled in a wellness plan are categorised as "subscribed," whereas those who were not enrolled are categorised as "not subscribed."	Categorical
Gender	Gender of the patients.	Gender of the patients were labelled into 1 (Males) and 2 (Females)	Categorical
Annual claim amount	Amount claimed by members to cover hospital costs or costs for medical drugs bought through pharmacies.	The monthly claim amounts were aggregated to derive the annual claim amounts. The total claim amount refers to the aggregate sum of claims made by the members for their diverse medical conditions throughout the course of the year.	Numerical
Complication development	Whether one had visited medical specialists with diabetes or hypertension – related complication	Only records were selected that indicated members had sought medical care from specialists for complications connected to diabetes or hypertension at least once throughout the span of the year. The identification of these records was conducted by examining the documentation that establishes a correlation between a patient diagnosed with diabetes or hypertension and their subsequent visits to a medical professional for the purpose of receiving treatments connected to diabetes or hypertension.	Categorical
Comorbidity	Coexistence of diabetes or hypertension	Individuals who are diagnosed with both diabetes and hypertension are commonly referred to as comorbid patients, while those who do not have these two specified disorders are categorised as non-comorbid patients.	Categorical
Units	Number of medical supplies or units dispensed to a patient for various ailments over a year.	The discrete data for the selected patients taken as is. Number of medical supplies or units dispensed to a patient for various ailments over a year.	Numerical
Medication adherence	Whether a diabetes or hypertensive patient monthly claimed or refilled medication for his or her diabetes or hypertension condition at 75% of the expected 12 times in a year [equivalent to at least 9 monthly claims or refills throughout a span of 12 months] (monthly over the 12-month period).	The monthly records within a given year were aggregated in order to calculate the total occurrences of prescription claims or refills for patients with diabetes or hypertension. MA was defined as meeting or above the 75% medication refill threshold, while any refill threshold below 75% was considered indicative of non-adherence. Therefore, the data was categorised into two distinct groups, namely adherent and non-adherent.	Categorical

3.3 Data Encoding and Feature Scaling

The encoding of categorical makes it possible to convert the predictor variables into numerical representations that the ML algorithms can examine efficiently, categorical data for the variables were encoded. Data encoding was conducted on the variables diagnosis, gender, comorbidity, complications development, and wellness programme. Within the framework of our research, the categorical data encoding technique involves the conversion of the categorical variable into a set of dummy binary variables, represented by the values 0 and 1. The categories were thus converted into distinct columns and assigned binary values of either 0 or 1. A value of 1 indicates the presence of the respective category, while a value of 0 signifies its absence. The steps for variable encoding led to the creation of a new dataset with differently captioned variables. The procedure for encoding dummy variables involves excluding the first variable among the newly generated dummy variables for each feature to prevent the occurrence of a dummy variable trap. For example, the variable 'gender' initially consisted of two categories, namely 1 for males and 2 for females. It would have been appropriate to transform this variable into two separate variables, namely *gender_1* for males and *gender_2* for females. However, in this case, only the *gender_2* (female) column is retained, while the *gender_1* (male) column is dropped to prevent the occurrence of the dummy variable trap. The dummy variable trap refers to a situation where two or more independent dummy variables, created through dummy encoding, exhibit high correlation or are affected by multicollinearity. This implies that the presence of multicollinearity poses challenges in interpreting the coefficients of predictor variables in regression models, namely LR as employed in this particular study. It is important to acknowledge that the inclusion of the gender_2 dummy variable is sufficient for our purposes. If an individual is not classified as female (1) in the gender_2 column, they are classified as male (0). Conversely, if an individual is classified as female, they are assigned a value of 1. Therefore, it is unnecessary to use both variables in our ML models. As an additional component of data pre-processing, we implemented feature scaling, a procedure we applied to the independent variables of the dataset to normalise the data within a specific range. The numerical variables, namely age, annual claim amount, annual contribution, and units were scaled down using a data scaling method to standardise the numerical input variables for classification. After data scaling, the numerical data were transformed into a standardised range of values spanning from -5 to 5.

3.4 Feature Selection

This study employed an embedded feature selection method which also incorporates the benefits of the filter method. The built-in feature importance mechanism in the RF method determined the significance of the encoded category and scaled numerical independent variables used to create the various ML-based classifiers. The most significant feature appears first in the form of a bar graph when the feature rankings are computed and sorted using the argsort method in descending order. All variables in the RF prediction model that possess importance scores above zero were actively included in the development of the ML classifiers utilised for the prediction of MA. The RF classifier, along with its corresponding Gini feature importance metric, facilitates the explicit removal of features. The Gini significance or mean decrease in impurity (MDI) metric computes the importance of each feature by summing the number of splits that involve the feature over all trees, weighted by the number of samples it splits [47]. Within the binary trees of the RF, the Gini impurity is employed at each node to determine the ideal split. Since the Gini impurity is a fast-to-calculate approximation of entropy, it is used to figure out how well a potential split separates samples from the two classes in a given node. The Gini importance measures the overall discriminative importance of a feature in the investigated classification problem and quantifies its frequency of selection for a split. Furthermore, for each ML method used in this study, SHAP bar plots and beeswarm plots were created to show the global relevance of each feature, which is expressed as the mean absolute value for that feature across all of the samples.

3.5 Model Development

The adherence prediction problem was conceptualised as a binary classification problem. Seven ML models—SVM, RF, DNN, DT, RF, Naïve Bayes, and K-NN were implemented to analyse MA in diabetes and hypertension patients. Prior to model development, the relevance of features in model development was measured using an embedded feature selection method known as the RF feature importance mechanism. Hence, this feature selection method embedded in RF was used to train only one ML model (RF) to select the significant features. The features found to be significant were then included in implementing all the ML models developed in this study. We utilised thirteen variables, consisting of dummy and scaled variables (twelve independent variables) and one dependent variable (MA), for developing the MA analytics models. The seven ML algorithms were implemented using the scikit-learn Python toolbox. The dataset was divided into two sets using random assignment: the training set consisting of 5698 patients (70%) and the test set consisting of 2443 patients (30%). Following the completion of training the ML models, we proceeded to fit the models to the test data. To mitigate the risk of overfitting and provide a more credible evaluation of the model's anticipated performance on unseen data, we employed K-fold cross-validation. Additionally, we used the GridSearchCV tool to tune hyperparameters and optimise model

performance. To improve the interpretability and explainability of the developed ML algorithms, we used the SHAP bar and beeswarm plots to explain the ML model outputs and interpret the features' contributions to the prediction or target outcome.

3.5.1 Algorithms

K-NN is a majority voting classification system. Similarity measures like distance functions determine the proximity between new and existing data points to classify them. This approach selects and compares K data points to determine their proximity to a specific class. A criterion of more than 50% of nearby data points determines the class. The k-NN algorithm finds the nearest data points to a query point to classify it. The k-NN technique locally approximates the target function f(x) = y during each prediction rather than globally. This study used the Euclidean distance measure to calculate the distances between data points and their neighbours to identify them as adherent or non-adherent. In Python, we employed a method wherein the distances are saved and then arranged in ascending order. We created a list of the top k items to assign a class to a new sample based on the majority of classes in nearby data points. The k-NN classifier was created using the KNeighborsClassifier module from sklearn.neighbor. In the k-NN model, the number of neighbours (n_neighbors) was set at 9. The k-NN classifier's neighbour count was manually varied. The values we evaluated were 3, 5, 6, and 7. Our results show that the classifier performed best with 9 neighbours.

In this study, the SVM algorithm locates the optimal hyperplane and positions it at the midpoint of the support vectors. This placement ensures that the hyperplane is maximally distant from the nearest points in each class, thereby minimising test errors. Our study used these SVC class parameters: We applied the kernel trick using the 'RBF' method, set the 'gamma' to 0.01, and used a random state of 1. The "C" parameter balances accurate training sample classification with maximisation of the margin of the decision function. The gamma parameter represents the kernel coefficient for the RBF kernels. We interpret the gamma parameters as the reciprocal of the radius of influence of the samples selected by the model as support vectors. In a high-dimensional space, the RBF kernel represents data points as vectors and measures their similarity by radial distance. SVMs use RBF kernels because they can capture complex input-output relationships. We employed extrapolation to assign lower values to the 'C' parameter and higher values to gamma, and vice versa. Ultimately, a relatively high ‘C’ of 1000 traded off with a relatively smaller ‘gamma’. The random state '1' denotes the seed of the pseudo-random number generator that is utilised during the process of shuffling the data.

Logistic regression estimates the probability of an event occurring, such as ‘adherent’ or ‘non-adherent’, based on a given data set of independent variables. This study initially examines the scenario in which the response variable (MA) is binary, with only two values coded as one (adherent) or zero (non-adherent) for ease of analysis. The LogisticRegression() module from the sklearn.linear_model library was utilised in building an LR model using Python. In addition, we employed the Statsmodels library to perform statistical computations for the LR analysis, including the estimation of coefficients, determination of p-values, calculation of odds ratios (ORs), and generation of other relevant outputs.

The study used Sequential() from Keras to build the neural network. The "sequential" function in Keras implies sequential model development, where each layer's outcome feeds the next layer. The first step is to create an input layer with 12 nodes, each representing one of our dataset's 12 input features. Each node is representative of an independent variable. The hidden layer connects all input nodes. Two hidden layers were added to the model. In this study, the hidden layers used the ReLU activation function. The activation function ReLU is known for its ability to minimise disappearing gradients during training in deep neural networks. The output layer had one node because there were only two target classes: adherent and non-adherent. The sigmoid activation was used at the output layer. It provided us with the probability of an MA. In the model's construction, compilation, and training, the "sequential" parameter specifies that we built the DNN model sequentially, using each layer's output as input for the next. The element "dense" is employed to designate the fully connected layer. For dense arguments, the input dimension is 12. In the last layer, specifically the output layer, there is a single node due to the implementation of a binary cross-entropy loss function and the utilisation of the sigmoid activation function. The single node represents adherent (1) and non-adherent (0) in the output layer. Selecting and computing the sigmoid function yields an output range of 0 to 1. The DNN model was optimised with the Adam optimizer due to its computational efficiency and performance across numerous deep-learning tasks. We used the binary cross-entropy loss for the classification task. The model was trained with 150 epochs and 10 batches. After conceptualising the Keras neural network architecture, we used TensorFlow to generate additional DNN outputs, which were visualised on TensorBoard. TensorBoard lets researchers and analysts monitor and visualise loss, accuracy, weight, and bias shifts using histograms. Histograms help visualise weights and biases and validate anticipated changes. The histograms facilitate the assessment of the model's learning progress.

The RF models built in this study use ensemble learning to aggregate the results of numerous DTs to reduce overfitting. Multiple DT classifiers were trained on distinct dataset subsets using sklearn.ensemble RandomForestClassifier. The parameters were: n_estimators = 200, max_depth = 7, criterion = "entropy," and random state = 1. The term "n_estimators" refers to the number of trees present within the forest. The "entropy" criterion is used to evaluate split quality. The max_depth parameter represents the upper limit on the depth of the tree. The random state parameter is used to control training randomization. The model will always produce the same results with a fixed random state value, identical parameters, and training data. We ran experiments with the default n_estimators value of 100 and additional model fitting parameters. The results of these experiments indicated a decrease in performance in terms of model accuracy. Increasing n_estimators to 200 and max_depth improved model performance. Each of the 200 DTs has rules that predict "adherent" or "non-adherent." RF interpretation would be difficult due to the maximum depth of 7 and the significant number of DTs in the visualisation; thus, we exported RF trees as textual representations to simplify analysis and interpretation. To export the RF trees, we used the tree.export_text method, which generates a textual report of the rules for creating the whole ensemble of 200 DTs.

Decision trees (DTs) are a type of non-parametric supervised learning technique that is commonly employed for classification. The objective is to develop a predictive model that estimates the value of the target variable by acquiring basic decision rules derived from the properties of the data. The *DecisionTreeClassifie*r module from the *sklearn.tree* library was utilised to construct the DT, which is a classification model organised in a tree form. The maximum depth was set to a widely used value of 5. Regarding the settings for splitting, we employed the default supported techniques, namely "best" for selecting the optimal split and "random" for selecting a random but optimal split. The criterion employed in our study is "gini," a recognised measure of tree quality.

The Naive Bayes algorithm is a probabilistic classification method that relies on Bayes' theorem, also known as the Nayes rule [48]. It is characterised by the "naive" assumption of independence among the predictive variables. The Gaussian Naive Bayes technique for classification was implemented using the *GaussianNB* module. The *GaussianNB* algorithm from the *sklearn.naive_bayes* library was utilised in developing the model.

In summary, we manually set various hyperparameters before training the ML models. It is important to acknowledge that during the training of ML models, each model requires a distinct set of hyperparameters, which are a type of variable. The hyperparameters used for building the ML models are displayed in Table 2.

Table 2: Hyperparameters set for the ML models

Model	Hyperparameters
ANN	activation: relu, loss: binary_crossentropy, batch_size:10, optimizer: adam, epochs: 150, number of hidden layers: 2.
RF	n_estimators: 200, bootstrap: True, criterion: entropy, max_depth: 7, max_features: sqrt, min_impurity_decrease: 0.0, min_sample leaf: 1, min_samples_split: 2, min_weight_fraction_leaf: 0.0, n_estimators: 200, random_state: 1, verbose: 0.
k-NN	algorithm: auto, cv: 5, leaf_size: 30, n_neighbors: 5, weights: uniform, metric: minkowski, euclidean distance p: 2.
DT	criterion: gini, max_depth: 5, cv: 5, min_impurity_decrease: 0.0, min_samples_leaf: 1, min_samples_split: 2, min_weight_fraction_leaf: 0.0, splitter: best.
LR	C: 1.0, fit_intercept: True, cv: 5, intercept_scaling: 1, max_iter: 100, multi_class: auto, penalty: l2, solver: lbfgs, tol: 0.0001, verbose: 0.
SVM	C: 1.0, cache_size: 200, cv: 5, decision_function_shape: ovr, degree: 3, gamma: 0.01, kernel: rbf, max_iter : -1, random_state: 1, shrinking: True, tol: 0.001 'verbose: False
Naïve Bayes	var_smoothing: 1e-09, cv: 5, alpha: 1.0, priors: None, fit_prior: True, force_alpha: warn, norm: False

3.5.2 Tools and Materials

The study utilised Anaconda Python 3.11 to create various ML models, including k-NN, SVM, RF, Naïve Bayes, and DT. Additionally, Python 3.11.3 was employed specifically for the development of the DNN model. These Python systems were used in conjunction with Jupyter Notebook 6.5.4, a widely used and esteemed integrated development environment (IDE) for Python. Several libraries were installed and imported into the Python IDE to enhance the ease and convenience of Python programming. Pandas, NumPy, Scikit-learn, TensorFlow, Keras, matplotlib, and Seaborn are the libraries used in the study.

3.6 Model Evaluation

The prediction models were evaluated using several metrics to facilitate a balanced comparison. Evaluation metrics used for evaluating the performance of the created ML models include the area under the receiver operating characteristics (ROC) curve (AUC) and the confusion matrix, also linked with accuracy, precision, recall, and the F1 score. The weighted average (weighted average) statistic was selected for representing precision, recall, and F1 score in our study due to the imbalanced nature of our dataset, with 19% representing the adherent class and 81% representing the non-adherent class. The utilisation of a weighted average is advantageous in situations where the dataset exhibits an imbalance, as it allows for a higher emphasis to be placed on a class with a larger number of cases (namely, the non-adherent class in our case) throughout the computation process. The weighted average method incorporates the consideration of class imbalance by multiplying the metrics for each class by the proportion of that class. ROC curves were generated to visually represent the performance of the binary classifiers across all classification thresholds. These curves plot the true positive rate (TPR) against the false positive rate (FPR) at different threshold settings. By doing so, they offer a comprehensive evaluation of performance across all potential classification thresholds. The AUC metric represents the mean true positive rates across all potential false positive values. The metric quantifies the model's ability to effectively discriminate across different classes.

Table 3: Model evaluation metrics

Metric	Description	Formula
Precision	Precision can be defined as the proportion of accurately classified positive samples (true positives) to the total number of positive samples that have been classified (whether rightly or wrongly). Intuitively, precision refers to the classifier's capacity to avoid labelling a negative sample as positive.	$\frac{TP}{TP + FP}$
Recall	Recall, alternatively referred to as the true positive rate (TPR), denotes the proportion of data instances that a ML model accurately recognises as belonging to a certain class of interest, commonly referred to as the "positive class," out of the overall number of instances inside that class. As the recall increases, there is a corresponding increase in the detection of positive samples.	$\frac{TP}{TP + FN}$
F1-score	The F1 score is an ML evaluation metric that assesses the accuracy of a model. The F1 score can be mathematically expressed as the harmonic mean of the precision and recall metrics of a classification model. The F1 score is a ML assessment metric that provides an alternative approach to assessing the predicted accuracy of a model. Unlike accuracy, which measures overall performance, the F1 score focuses on the class-wise performance of the model.	$\frac{2 * (Precision * Recall)}{Precision + Recall}$
Accuracy	The accuracy measures the number of true predictions generated by a model over the entirety of the dataset. Accuracy refers to the proportion of accurately predicted data points in relation to the total number of data points.	$\frac{TP + TN}{TP + TN + FP + FN}$

Table 3 demonstrates the calculation of precision, recall, F1-score, and accuracy. These are all computed based on the four metrics associated with the confusion matrix as described in Table 4.

Table 4: Metrics' combinations in the confusion matrix

Metric	Description
True Positives (TP)	The accuracy of a model in properly identifying a positive sample as positive.
False Negatives (FN)	The frequency at which a model erroneously assigns a positive sample as a negative classification.
False Positives (FP)	The frequency at which a model erroneously assigns a negative sample as positive.
True Negatives (TN)	The accuracy of a model in properly identifying a negative sample as negative.

By and large, we evaluated the ML classification models by employing various metrics to establish a fair comparison. Figure 3 below illustrates the sequential process of data extraction, data preparation (including variable encoding and scaling), feature selection, ML modelling, and model performance evaluation.

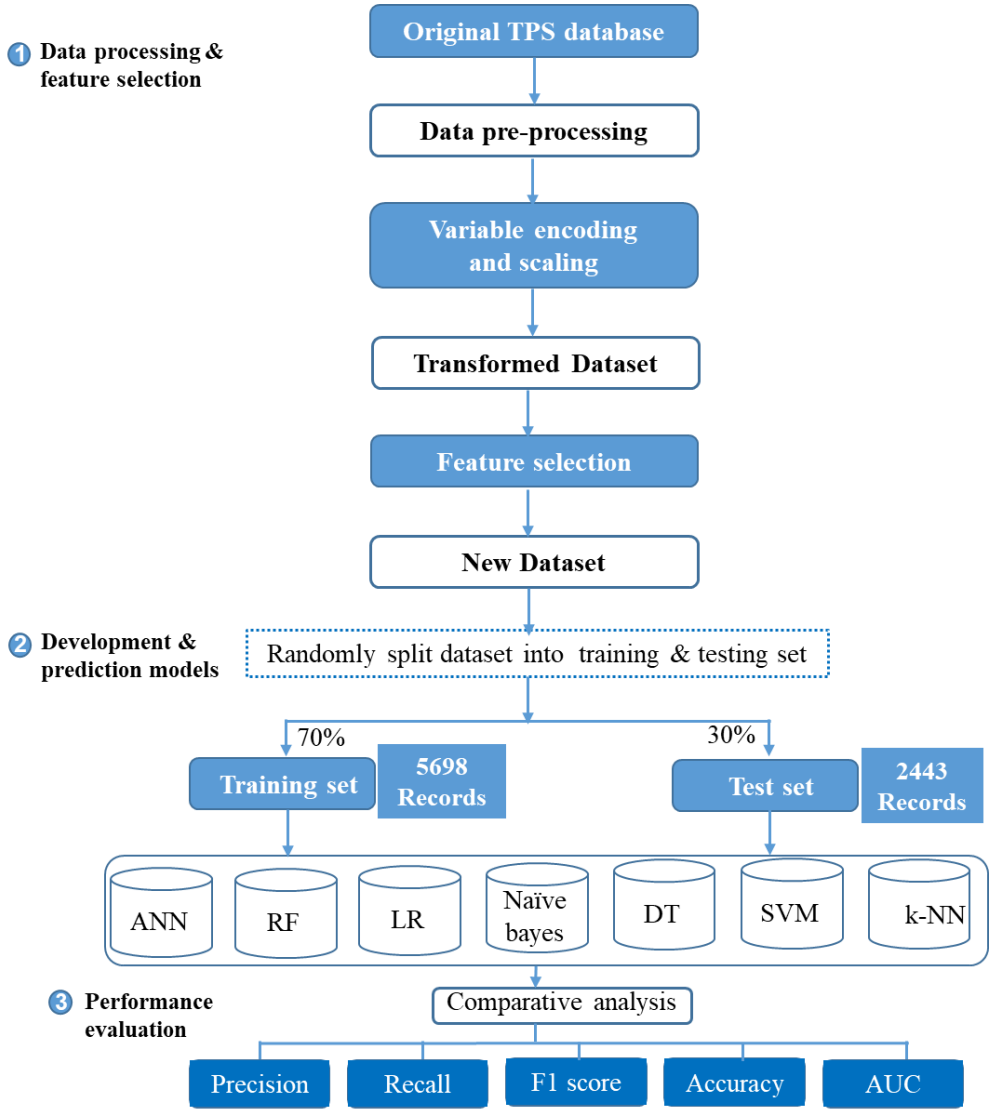


Figure 3: ML models' development flow chart

4. Results

This section reports the findings related to the research objectives. The ML problem at hand involves classifying individuals with diabetes and hypertension based on their adherence to medication refills. The objective of the study was to apply various ML algorithms to classify patients based on their adherence to the recommended minimum medication refill threshold of 75% (corresponding to at least 9 months out of 12). We classified patients as adherent if they met or exceeded this threshold, and as non-adherent if they fell below 75%. In this case, formulating the problem in this way yields a binary classification task. Before presenting the model results, we first present the results of the exploratory data analysis.

4.1 Exploratory Data Analysis Results

The data analysis technique comprises a summary of the statistics for each study variable. Table 5 reveals that there are 6461 tuples of data for hypertensives and 1680 for diabetics. Of the 8141 patients, 4768 (59%) were females and 3373 (41%) were males. At least 53% of Medical Aid Society members had basic-to-grade B hospital health insurance coverage, while 47% had private-to-grade A coverage. Comorbidity: the coexistence of diabetes and hypertension in 30% (2402 out of 8141), whereas 70% (5739%) did not. The remaining 70% had either diabetes or hypertension, with 30% having both. In the period under investigation, 3417 people (42% of the total) sought specialist medical attention for diabetes or hypertension-related complications. However, 4724 patients (58% of the sample) experienced no diabetic or hypertension problems. Statistics reveal that 19% of participants followed their scheduled prescriptions. Adherence entails refilling diabetic or hypertensive medication every month for at least nine months in 2022, or 75% of all scheduled refills. In the data sample, 81% had not refilled

the medication more than 75% of the recommended 12 times. Tables 5 and 6 provide summary statistics for the study's categorical and continuous variables.

Table 5: The summary statistics of the categorical features prior to data encoding (n = 8141)

Feature	Attributes	Value	Percentage (%)
Diagnosis	Diabetes	1680	21
	Hypertension	6461	79
	Mode	Hypertension	-
Gender	Female	4768	59
	Male	3373	41
	Mode	Female	-
Scheme type	Basic Level	1453	18
	Medium Level	5180	64
	Premium Level	1508	18
	Mode	Medium Level	-
Hospital cover type	Basic to Grade B	4281	53
	Private to Grade A	3860	47
	Mode	Basic to Grade B	-
Wellness program	Subscribed	3160	39
	Not subscribed	4981	61
	Mode	Not Subscribed	-
Complication Development	Yes	3417	42
	No	4724	58
	Mode	No	-
Cormobidity	Yes	2402	30
	No	5739	70
	Mode	No	-
Adherence	Adherent	1541	19
	Non-Adherent	6600	81
	Mode	Non-Adherent	-
Age	Mean + SD	53 + 14	-
	Age range	16-101	-
	Median	52	-
Annual contribution (US$)	Mean + SD	3206 + 7307	-
	Age range	84-25452	-
	Median	768	-
Annual claim amount	Mean + SD	222 + 1128	-
	Age range	0.10-40325	-
	Median	69	-
Units	Mean + SD	363 + 461	-
	Age range	1-13140	-
	Median	216	-

Table 6 shows the mean scores and SDs of the numerical variables in the dataset prior to encoding and scaling of the data, along with other descriptive statistics. Regarding the variable of age, it is noteworthy that the minimum and maximum values observed among the individuals were 16 years (this age group was filtered out due to ethics considerations) and 101 years, respectively. The average age of the sampled subjects was 53, as reflected by the mean of 53 (SD = 14). The average annual contribution per member of the Medical Aid Society was $3206.00, whereas the average annual claim amount was $222.00.

Table 6: The summary statistics of the numerical variables prior to variable scaling (n = 8141)

Feature	Attributes	Value
Age	Mean + SD	53 + 14
	Age range	16-101
	Median	52
Annual contribution (US$)	Mean + SD	3206 + 7307
	Age range	84-25452
	Median	768
Annual claim amount	Mean + SD	222 + 1128
	Age range	0.10-40325
	Median	69
Units	Mean + SD	363 + 461
	Age range	1-13140
	Median	216

The correlation heatmap in Figure 4 shows correlations between the transformed variables in the final dataset. The Pearson correlation coefficient of 0.71 in Figure 4 shows a positive correlation between annual contribution and scheme type 3 (premium membership level). This implies that enrollment in a premium medical aid plan or scheme is closely associated with a greater annual contribution in comparison to lower-tier medical aid schemes (basic and medium schemes). In essence, transitioning from a lower-tier medical aid membership plan to a higher-tier or premium medical aid coverage results in a corresponding rise in the annual contribution. Another description is that patients who use more medical supplies or units were more likely to adhere to diabetic and hypertension medicine, as demonstrated by the positive Pearson correlation value of 0.47. Thus, there is a positive correlation between the number of medical supplies (units) a patient takes and MA. The reported correlation values were less than 0.8, indicating that the variables used in the current study were not highly correlated. The absence of multicollinearity in Pearson correlation coefficients below 0.8 has minimal or no impact on prediction model accuracy. Collinearity increases the variances of parameter estimates, resulting in inaccurate inferences about the relationships between explanatory and response variables [49].

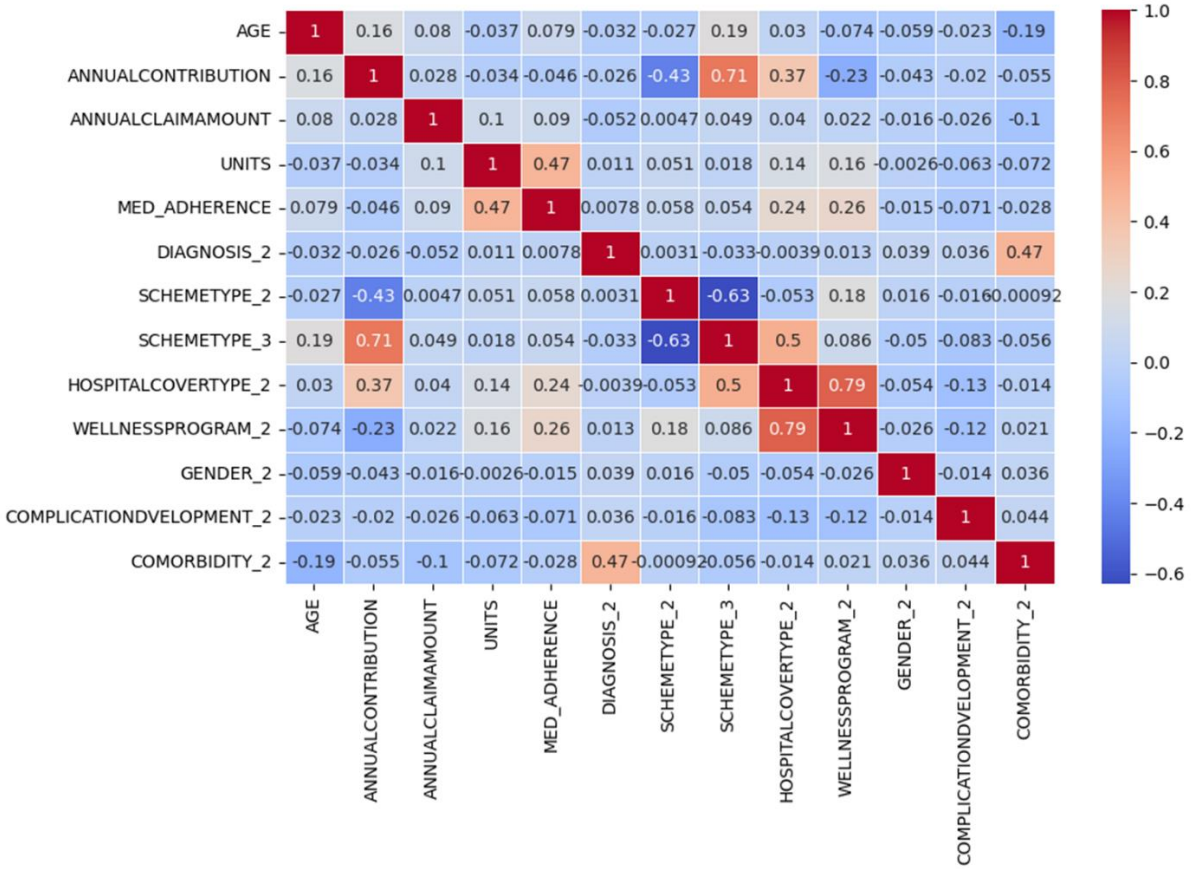


***Figure 4: Correlation heatmap visualising* relationships between the transformed variables**

4.2 Feature importance analysis results

Figure 5 shows the ranked list of features determined through the RF feature importance analysis. The argsort method ranks features in decreasing order, with the most important one first. The bar chart displayed the features in descending order of importance. The highest feature importance score (0.500) indicated that the cumulative number of medical supplies or units supplied to a patient annually was the most important factor in predicting MA. At 0.259, the RF feature importance metric ranks the annual claim amount as the second-most significant predictor of MA. The findings indicate that the two variables, units and annual claim amount, had significant contributions of 50% and 20%, respectively, in predicting the target variable. Moreover, the SHAP plots in Figure 6 confirm that the total number of medical supplies or units provided to a patient each year is the most important predictor of MA. Complications development, gender, and scheme type have RF importance scores below 0.01, making them the least significant features in the MA prediction model. The RF importance scores showed that

the scheme type was the least important. It is noteworthy that all the variables inside the RF prediction model had significant scores over zero. Therefore, the study used all the features to build seven ML classifiers.

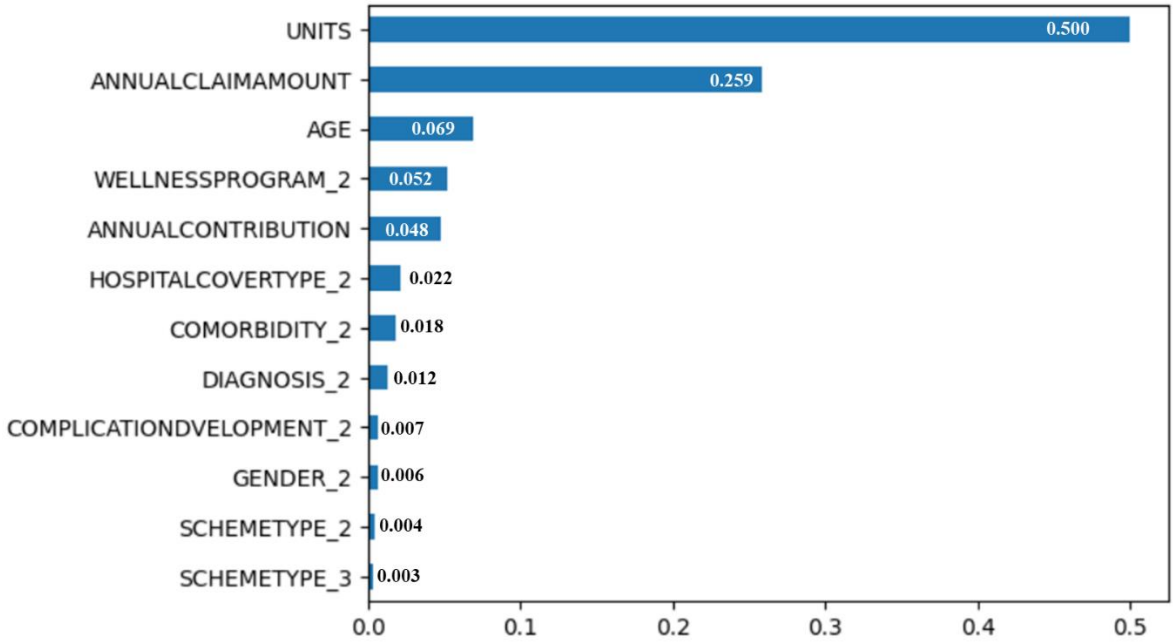


Figure 5: Visualisation plot for feature importance of the encoded and scaled variables using RF Classifier

4.3 Model Results

The RF and DNN models emerged as the top performers in MA classification among the seven ML models created. On the other hand, the LR model didn't perform as well, although it still showed decent results. The RF classifier demonstrates exceptional performance, as shown in Table 7. We obtained an array of accuracy scores when we used the 5-fold validation technique to train and validate the RF model. These scores, displayed in Table 8 are 0.867, 0.860, 0.860, 0.859, and 0.866. On the whole, the average cross-validation score is 0.862 ± 0.004. The RF model's accuracy is consistent, as evidenced by the 5-fold accuracy scores and the very low standard deviation (0.004). Therefore, the RF model demonstrates a strong capability to make accurate predictions on new MA-related data not included in its training. Through the utilisation of grid search, the RF model's accuracy is enhanced from 0.862 to 0.876. Utilising feature importance is a popular and valuable approach for interpreting and conveying additional results of RFs. According to the RF model presented in Table 7, the variable of greatest importance is the 'units' feature, followed by the annual claim amount and the age of the patient. In summary, the 'units' feature refers to the variable that exhibited the lowest impurity and was utilised to perform the split in our RF model. This finding indicates that the number of medical supplies or units distributed to a patient for various ailments was the primary factor that significantly reduced the weighted impurity in the RF classifier. The SHAP beeswarm plot in Figure 6 confirms the RF importance scores, indicating that the three most important features in the model are units, age, and annual claim amount. To aid the interpretation of the RF model, the SHAP beeswarm plot clearly validates the three most important features in the model's predictions: units, age, and annual claim amount.

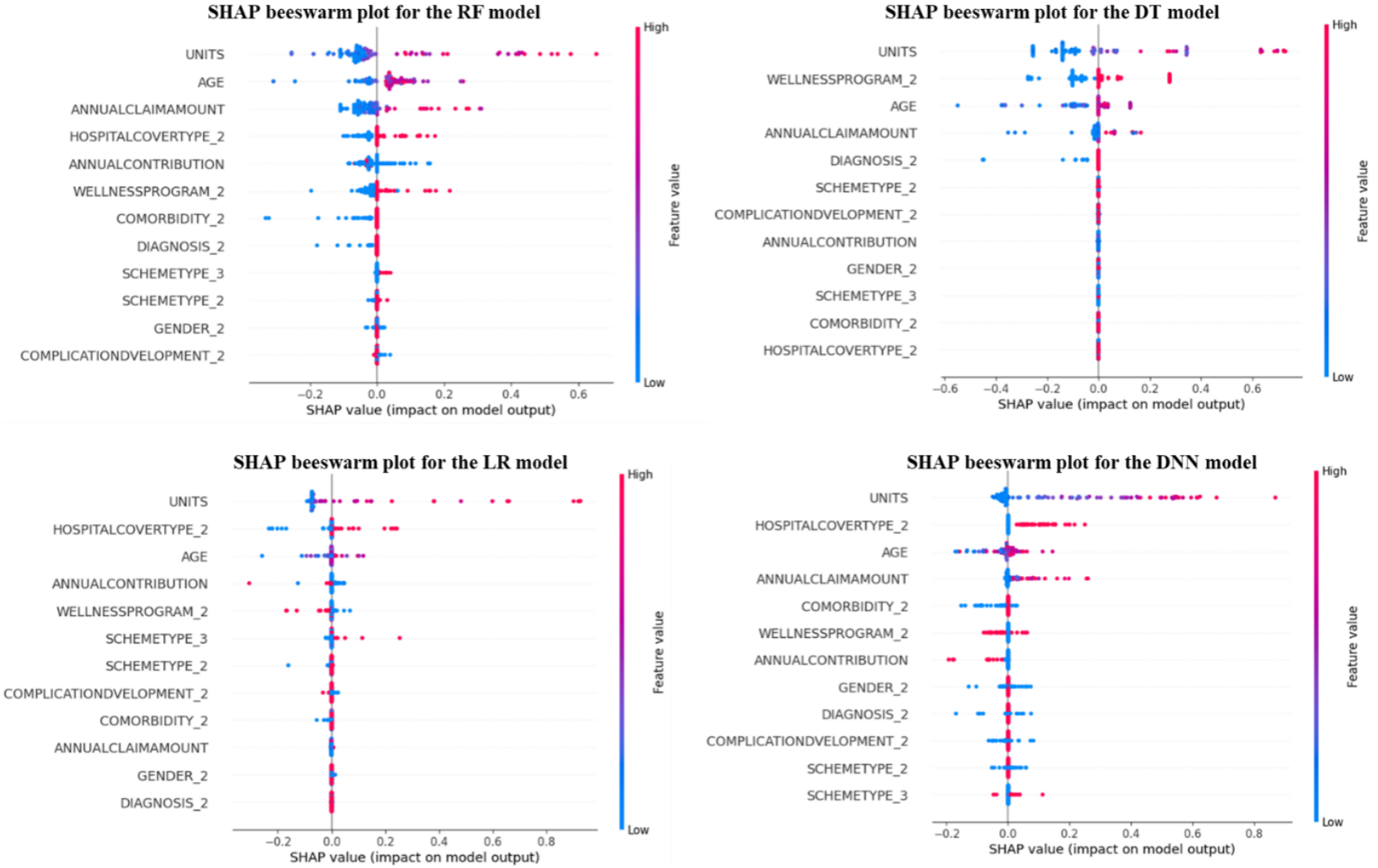


Figure 6: SHAP beeswarm plots. The beeswarm plots display an information-dense summary of how the top features in a dataset are impacting the models' output. The SHAP plots show the features with the biggest impact on the models' predictions.

Table 7: RF model development feature rankings

	Feature Score
UNITS	0.500
Annualclaimamount	0.259
Age	0.069
Wellnessprogram_2	0.052
Annualcontribution	0.048
Hospitalcovertype_2	0.022
Comorbidity_2	0.018
Diagnosis_2	0.012
Complicationdvelopment_2	0.007
Gender_2	0.006
Schemetype_2	0.004
Schemetype_3	0.003

The DNN model structure and how different layers connect through several neurons is illustrated in Figure 7, as shown below. The input layer consists of 12 nodes, each corresponding to one of the 12 input or independent features available in our dataset. The hidden layers were additionally receiving input from a set of 12 nodes, with an output size that was likewise equivalent to 12. The output layer consists of a single node that represents either an adherent (1) or non-adherent (0) outcome. The DNN achieved an accuracy level above 87% after undergoing 150 epochs and being built with two hidden layers.

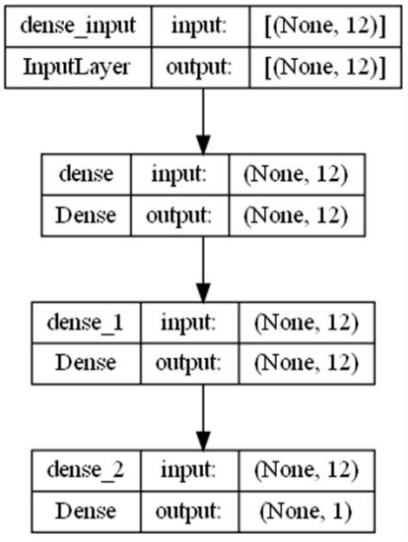


Figure 7: DNN model structure and how different layers connect to each other through a number of neurons

The fluctuations in accuracy and loss across each epoch during model training can be observed in Figure 8. As the number of epochs increases, accuracy is shown to increase before reaching a saturation point, and vice versa for loss. In general, the model demonstrates a peak accuracy of no less than 87% while utilising two hidden layers, a batch size of 10, and training for 150 epochs. The larger the number of epochs provided, the more the model learns about the data. The findings show that the model was learning quickly because the accuracy in the first epoch was above 80%.

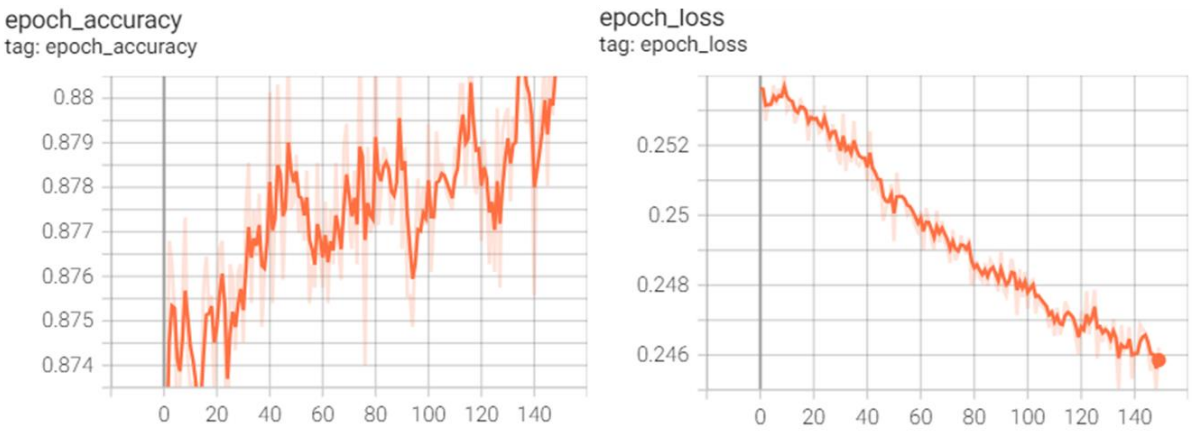


Figure 8: Accuracy or cross entropy loss versus epoch

This 3D histogram in Figure 9 below encompasses three distinct axes, namely the values of the kernels and biases, the epoch number, and the count of weights. The histograms facilitate the assessment of the model's learning progress. The histogram displays the values corresponding to each epoch within the range of 0 to 149, which corresponds to a total of 150 epochs. The histograms indicate that the values within the bias vector of the dense_1 layer fall between the range of -0.619 and 1.03. In essence, bias denotes the extent to which predictions deviate from their intended value. A bias value of -0.619 suggests that the neural network is inclined to make more assumptions regarding the form of the output. Conversely, a bias value of 1.03 shows that the network is inclined to make fewer assumptions about the form of the output. The inclusion of the bias unit ensures that the neuron will exhibit activation even in scenarios where all input values are 0. In the context of the study, the weights within the DNN serve an essential role in the transformation of input variables to impact the resulting outputs. It is worth noting that a lower weight value will have a negligible effect on the input, whereas a higher weight value will exert a more pronounced contribution to the output. It is noteworthy in our study that the histograms exhibit variability across epochs, indicating that the model underwent effective training with little or negligible errors.

Similar to the RF model, the DNN model's SHAP plot shows that the top four features that have the biggest influence on the model's predictions are the hospital cover type, age of the patient, annual claim amount, and the "units" feature, which refers to the number of medical supplies or units given to a patient for various ailments.

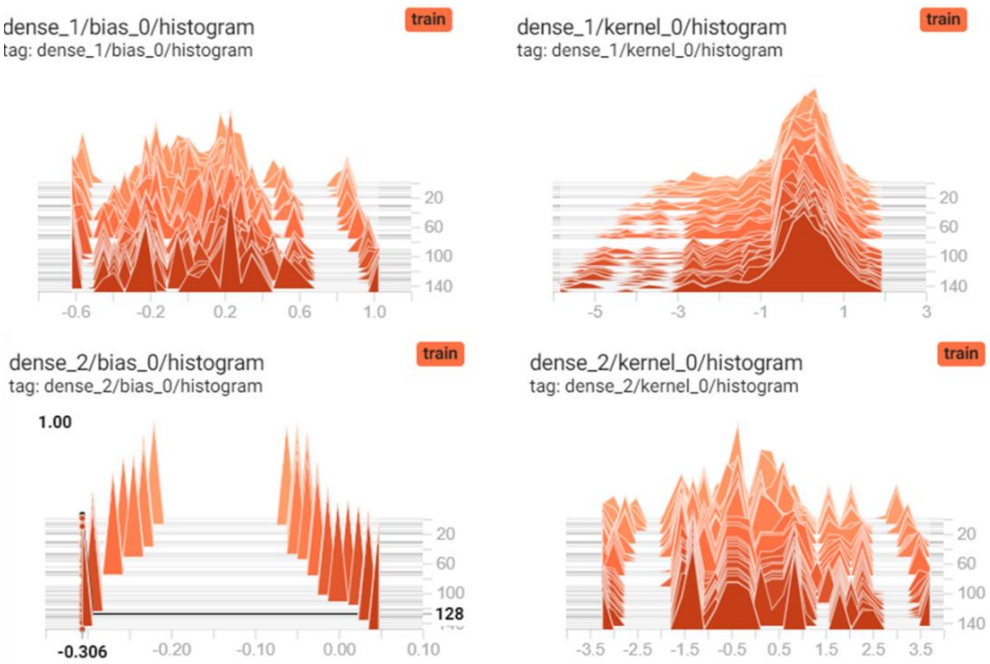


Figure 9: TensorBoard histograms showing bias and weights (kernels)

The equation representing the model, LR is provided below.

$$ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = 0.05 + 1.46X1 + 0.28X2 + 1.09X3 + 4.21X4 + 0.97X5 + 1.11X6 + 1.53X7 + 52.45X8 + 0.06X9 + 0.96X10 + 0.88X11 + 1.29X12$$

Where X1 represents the variable *AGE*, X2 represents the variable *ANNUALCONTRIBUTION*, X3 represents the variable *ANNUALCLAIMAMOUNT*, and X4 represents the variable *UNITS*. X5 - *DIAGNOSIS_2*, X6 - *SCHEMETYPE_2*, X7 - *SCHEMETYPE_3*, X8 - *HOSPITALCOVER_2*, X9 -*WELLNESSPROGRAM_2*, X10 - *GENDER_2*, X11 - *COMPLICATIONDEVELOPMENT_2*, X12 - *COMMORBIDITY_2*, and const. is the constant shown in Figure 10. The regression coefficients, p-values, relevant statistical measures, and odds ratios (ORs) are presented in Figure 10. An OR of more than 1 signifies an increased likelihood of an event (in our case, MA), whereas an OR below 1 signifies a reduced likelihood of an event (in our case, MA). The statistics presented in Figure 10 provide ORs for various features, including age, units, and annual claim amount. These ORs indicate that an increase in a patient's age, units (representing the total number of different medications and medical supplies provided to a patient), and annual claim amount are associated with a higher likelihood of experiencing the outcome (MA). For each additional year in the age of a patient, the odds that the patient will adhere to medication increase by 1.46 times. Consequently, older age groups exhibit a higher predisposition for drug adherence. The largest OR is approximately 52.45, which indicates that patients with higher hospital cover access (private to Grade A hospital access) are 52.45 times more likely to adhere to medication than those with lower hospital cover grade or access. Once again, individuals classified under the higher medical aid scheme type category (*SCHEMETYPE_2* or medium) had a 1.11-fold higher likelihood of MA in comparison to those belonging to the lower scheme type (basic level). The succeeding section presents the results pertaining to the performance of all the seven ML models (RF, DNN, LR, k-NN, SVM, DT, and Naïve Bayes) built in this study.

Logit Regression Results

Dep. Variable:	MED_ADHERENCE	No. Observations:	5698
Model:	Logit	Df Residuals:	5685
Method:	MLE	Df Model:	12
Date:	Sun, 17 Sep 2023	Pseudo R-squ.:	0.2876
Time:	15:52:44	Log-Likelihood:	-1941.6
converged:	True	LL-Null:	-2725.6
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.9056	0.306	-9.496	0.000	-3.505	-2.306
AGE	0.3780	0.045	8.442	0.000	0.290	0.466
ANNUALCONTRIBUTION	-1.2569	0.661	-1.900	0.057	-2.553	0.039
ANNUALCLAIMAMOUNT	0.0895	0.034	2.623	0.009	0.023	0.156
UNITS	1.4384	0.053	27.086	0.000	1.334	1.542
DIAGNOSIS_2	-0.0259	0.118	-0.220	0.826	-0.257	0.205
SCHEMETYPE_2	0.1062	0.156	0.679	0.497	-0.200	0.413
SCHEMETYPE_3	0.4263	0.230	1.856	0.063	-0.024	0.876
HOSPITALCOVERTYPE_2	3.9598	2.159	1.834	0.067	-0.272	8.192
WELLNESSPROGRAM_2	-2.8579	2.144	-1.333	0.183	-7.061	1.345
GENDER_2	-0.0437	0.084	-0.522	0.602	-0.208	0.120
COMPLICATIONDVELOPMENT_2	-0.1230	0.084	-1.460	0.144	-0.288	0.042
COMORBIDITY_2	0.2561	0.106	2.419	0.016	0.049	0.464

	OR	Lower CI	Upper CI
const	0.054714	0.030037	0.099665
AGE	1.459407	1.336776	1.593286
ANNUALCONTRIBUTION	0.284526	0.077824	1.040225
ANNUALCLAIMAMOUNT	1.093627	1.022872	1.169276
UNITS	4.213755	3.797239	4.675959
DIAGNOSIS_2	0.974434	0.773419	1.227693
SCHEMETYPE_2	1.112084	0.818356	1.511237
SCHEMETYPE_3	1.531636	0.976461	2.402461
HOSPITALCOVERTYPE_2	52.447023	0.761585	3611.799093
WELLNESSPROGRAM_2	0.057392	0.000858	3.839174
GENDER_2	0.957226	0.812380	1.127898
COMPLICATIONDVELOPMENT_2	0.884250	0.749659	1.043004
COMORBIDITY_2	1.291929	1.049816	1.589880

Figure 10: Logistic Regression Analysis Results. The outputs are for a medication adherence model generated with logistic regression (LR). The output consists of the dependent and independent variables, the number of observations in the training set, variable coefficients, p-values, and the model's odds ratio.

4.4 Performance of the ML models for the MA classification task

As depicted in Table 8 and Figure 11, all of the models have a classification accuracy of no less than 84% and an AUC value over 0.85, indicating satisfactory overall performance across all seven ML models employed in this classification problem. In relation to the accuracy of the models, the RF method demonstrated superior performance compared to the other algorithms, attaining the highest classification accuracy of 87.6%. This indicates that the RF algorithm had superior performance in accurately predicting the classification of patients as either adherent or non-adherent. Although not extremely low, the LR model exhibited the lowest accuracy rate, with a statistic of 84.7%. Interestingly, all the algorithms achieved a good level of accuracy since an accuracy score of at least 0.7 is deemed decent or a good level. The AUC values for RF, DNN, DT, SVM, k-NN, Naïve Bayes, and LR models are 0.934, 0.916, 0.914, 0.907, 0.905, and 0.857 respectively. As illustrated by the ROC curves presented in Figure 12 and Figure 13, all of the curves deviate from the AUC 0.5 diagonal line (which representsnts an ideal bad or worthless test), and instead approach the upper corner, which indicates excellent models since those with AUC scores around 1 demonstrate a high degree of separability. A test's overall performance is considered better as the ROC curve approaches the upper left corner. Therefore, achieving an AUC greater than 0.8 for all classifiers (LR, SVM, DNN, DT, k-NN, Naïve Bayes, and RF) indicates that the models exhibited exceptional performance in

accurately classifying MA as adherence or non-adherence. The findings presented in Table 8 indicate that the RF model, followed by the DNN model, exhibited superior performance compared to the other models in accurately diagnosing adherence versus non-adherence among patients diagnosed with diabetes and hypertension. This is evident from the achieved AUC scores of 0.934 and 0.916, respectively.

Naive Bayes Classification report:

	precision	recall	f1-score	support
Non Adherent	0.8718	0.9539	0.9110	1954
Adherent	0.7049	0.4397	0.5416	489
accuracy			0.8510	2443
macro avg	0.7884	0.6968	0.7263	2443
weighted avg	0.8384	0.8510	0.8371	2443

SVM Classification report:

	precision	recall	f1-score	support
Non Adherent	0.8817	0.9575	0.9181	1954
Adherent	0.7414	0.4867	0.5877	489
accuracy			0.8633	2443
macro avg	0.8116	0.7221	0.7529	2443
weighted avg	0.8536	0.8633	0.8519	2443

DT Classification report:

	precision	recall	f1-score	support
Non Adherent	0.9198	0.9090	0.9144	1956
Adherent	0.6510	0.6817	0.6660	487
accuracy			0.8637	2443
macro avg	0.7854	0.7954	0.7902	2443
weighted avg	0.8662	0.8637	0.8649	2443

LR Classification report:

	precision	recall	f1-score	support
Non Adherent	0.8628	0.9621	0.9098	1954
Adherent	0.7197	0.3885	0.5046	489
accuracy			0.8473	2443
macro avg	0.7912	0.6753	0.7072	2443
weighted avg	0.8341	0.8473	0.8287	2443

RF Classification report:

	precision	recall	f1-score	support
Non Adherent	0.908	0.941	0.924	1954
Adherent	0.724	0.618	0.667	489
accuracy			0.876	2443
macro avg	0.816	0.779	0.795	2443
weighted avg	0.871	0.876	0.873	2443

k-NN Classification report:

	precision	recall	f1-score	support
Non Adherent	0.8800	0.9498	0.9136	1954
Adherent	0.7066	0.4826	0.5735	489
accuracy			0.8563	2443
macro avg	0.7933	0.7162	0.7436	2443
weighted avg	0.8453	0.8563	0.8455	2443

DNN Classification report:

	precision	recall	f1-score	support
Non Adherent	0.9010	0.9360	0.9182	1954
Adherent	0.6973	0.5890	0.6386	489
accuracy			0.8666	2443
macro avg	0.7992	0.7625	0.7784	2443
weighted avg	0.8602	0.8666	0.8622	2443

Figure 11: ML classification reports. The classification reports were generated as part of the medication adherence models built using seven machine learning models: naïve bayes, support vector machines (SVM), decision trees (DT), logistic regression (LR), random forest (RF), k-nearest neighbor (k-NN), and deep neural networks (DNN).

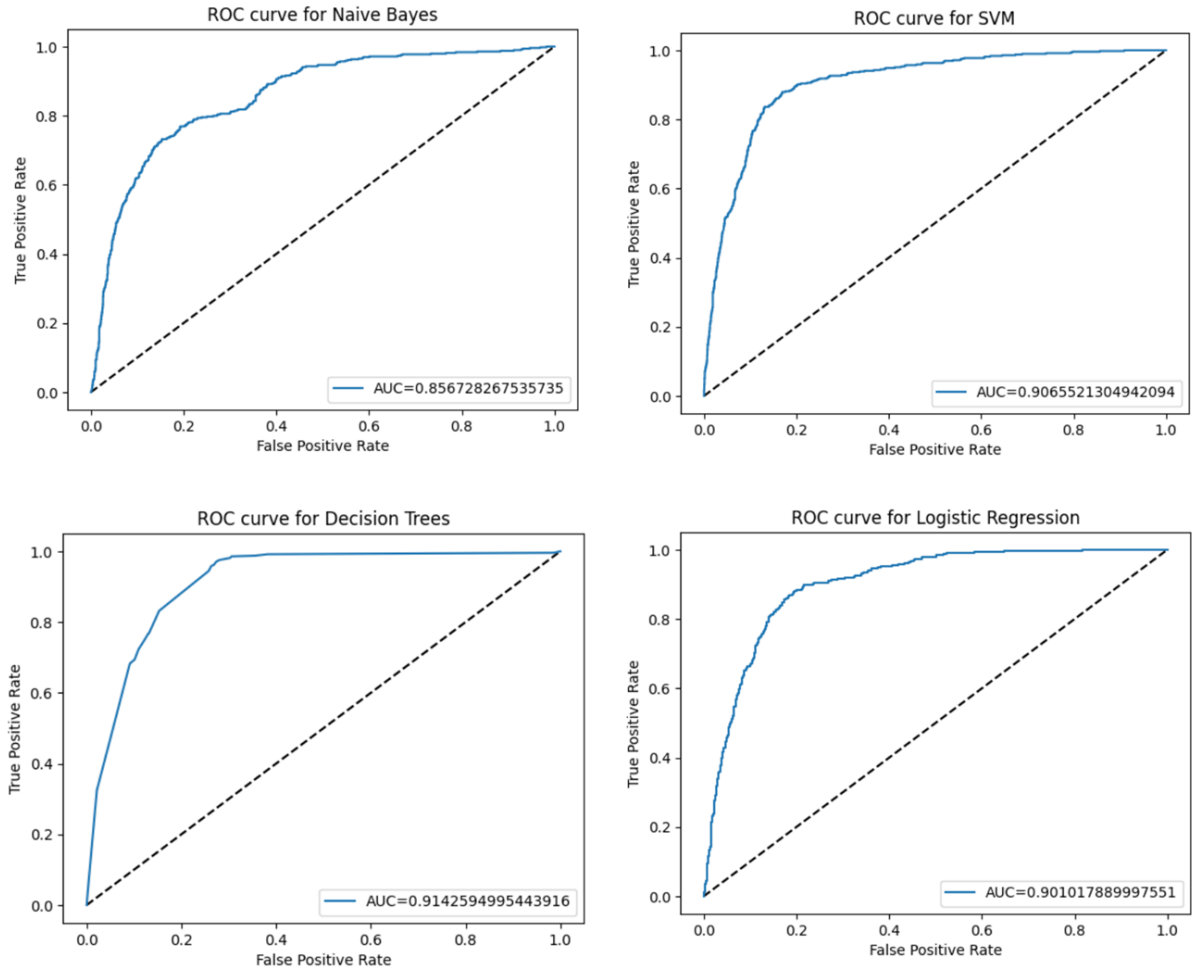


Figure 12: ROC curves for the naïve bayes, SVM, DT and LR. These are ROC curves produced for the medication adherence models built using four different machine learning models (Naïve bayes, SVM, DT and LR). These curves demonstrate the classification models' performance at all classification thresholds. The curves' plots have two parameters: True Positive Rate. False Positive Rate.

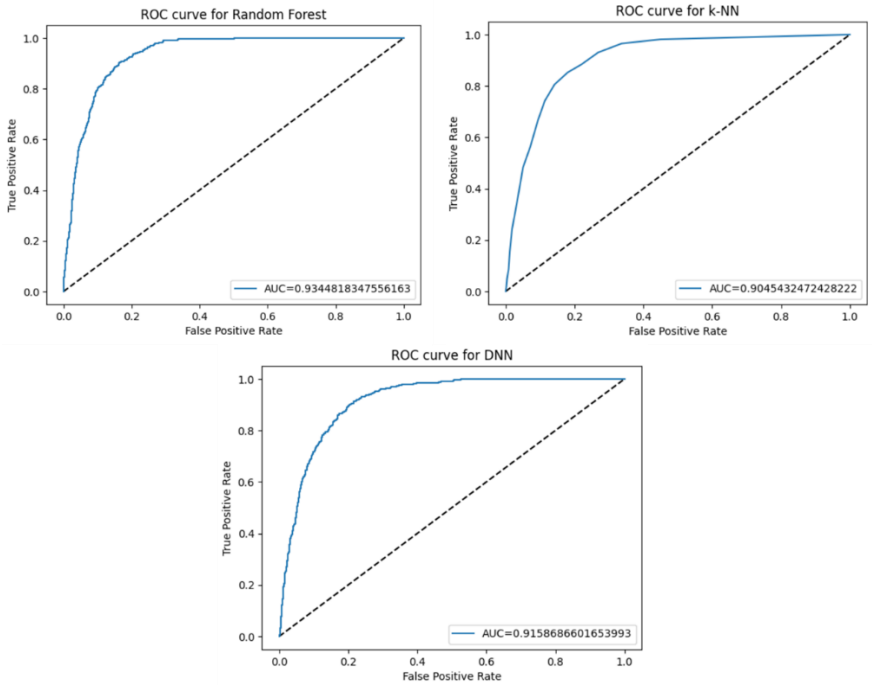


Figure 13: ROC curves for the RF, k-NN and DNN. These are the ROC curves for medication adherence models developed using three different machine learning models (RF, k-NN, and DNN). These visualisations represent the classification model's performance across all classification thresholds. The curve plots feature two parameters: True Positive Rate. False positive rate.

In additional results, Table 8 shows the weighted average precision scores for the seven models varying between 0.834 (LR) and 0.871 (RF). This indicates that the RF model correctly predicts MA in around 87% of cases for the RF model. The weighted average recall scores for the ML models generated ranged from 0.847 (LR) to 0.876 (RF), as determined by the recall metric. This metric measures the accuracy of the models in correctly predicting all potential positive observations. Therefore, in the case of the high-performing RF model, the recall metric indicates a value of 88%, indicating the model's ability to correctly identify and retrieve 88% of the objects belonging to the target class. A range of the weighted average F1 scores between 0.8 and 0.9 is considered indicative of a high level of accuracy. The study produced seven ML models, each of which achieved weighted average F1 scores ranging from 0.8281 to 0.8724. In the context of binary classification models, a high F1 score approaching 1 signifies exceptional precision and recall, whereas a low value suggests subpar model performance. These results indicate that the models performed well in terms of class-wise evaluation. Furthermore, the accuracy scores obtained on the entire dataset ranged from 0.847 to 0.876, which is consistent with the performance of the developed models. By and large

Table 8 Results of the k-fold cross validation and ML performance

Fold Number	RF Accuracy	DNN Accuracy	DT Accuracy	SVM Accuracy	k-NN Accuracy	Naïve bayes Accuracy	LR Accuracy
1	0.867	0.863	0.859	0.863	0.866	0.850	0.850
2	0.860	0.852	0.850	0.840	0.840	0.846	0.837
3	0.860	0.860	0.864	0.855	0.844	0.841	0.845
4	0.859	0.865	0.862	0.849	0.849	0.849	0.848
5	0.866	0.853	0.860	0.863	0.855	0.859	0.857
Mean accuracy	0.862	0.858	0.859	0.853	0.851	0.849	0.847
Std. Deviation	0.004	0.005	0.005	0.009	0.010	0.006	0.007
ML model performance after hyperparameter tuning using GridSearchCV							

Metric / Algorithm	RF	DNN	DT	SVM	k-NN	Naïve bayes	LR
Macro average precision	0.816	0.799	0.785	0.812	0.793	0.788	0.791
Macro average recall	0.779	0.763	0.795	0.722	0.716	0.697	0.675
Macro average f1-score	0.795	0.778	0.790	0.753	0.744	0.726	0.707
Weighted average precision	0.871	0.860	0.866	0.854	0.845	0.838	0.834
Weighted average recall	0.876	0.867	0.864	0.863	0.856	0.851	0.847
Weighted average f1-score	0.873	0.862	0.865	0.852	0.846	0.837	0.829
Accuracy	0.876	0.867	0.864	0.863	0.856	0.851	0.847
AUC	0.934	0.916	0.914	0.907	0.905	0.857	0.901

5. Discussion

In this study, MA was assessed by utilising drug claims data acquired throughout 12 months from CIMAS, a medical aid insurance firm. The utilisation of this data facilitated the attainment of a more impartial and all-encompassing depiction of MA or non-adherence. In ML models, it is customary to assign rankings to features based on their individual "contributions" or levels of importance. The results of the feature indicate that the total number of medications and medical supplies dispensed to patients for a range of ailments is the most influential factor in predicting MA, as shown by the computed RF importance scores mechanism and beeswarm plots for the various ML models built in this study. This phenomenon may be attributed to the observation that patients who are prescribed diverse drugs are more likely to develop a habitual practice of taking drugs regularly. The individual's historical pattern of increased medication utilisation and acquired understanding of MA may potentially have a favourable impact on their consistent engagement in medication refilling, administering, or taking their prescribed prescription drugs. There is evidence that people who regularly take their hypertension medication score higher on MA [50]. This could imply that the patients would have developed a tradition or practice of refilling or taking medication as prescribed or scheduled.

Age is one of the primary factors that had a considerable predictive value for MA. In our study, the average age of the participants was found to be 53. Consequently, it can be inferred that this particular age group predominantly accounts for the observed variations in the level of MA within our study. This is consistent with the findings of [51] who used real-world data and ML techniques to examine adherence among commercially insured NCD patients (metastatic breast cancer patients) in the United States. They discovered that one of the top two predictors of MA was age 55. In another study conducted in the United States by [18], it was found that age had a crucial role in the lack of adherence to hypertension treatment. The researchers employed a population segmentation method based on ML techniques. Once more, age was taken into account as a risk factor by [52] then they classified MA in patients with hypertension using LR, DT, neural networks, and RF. Therefore, age is identified as an essential factor that is both incorporated and observed to exert a substantial influence within the context of this research. Similar to the results of a prior study by [37] that identified insurance type as a significant predictor in identifying patients who were likely to discontinue medication within the twelve months, hospital cover type has also been reported as a significant predictor of MA in this study. The inclusion of the annual claims amount as one of the primary predictors of MA suggests that the utilisation of health insurance plans for regular medication claims to manage or treat chronic conditions may significantly impact the level of adherence among patients with NCDs. This study also revealed that higher prescription claims have a significant impact on medication refills and adherence.

This study presents evidence of the efficacy of ML approaches in accurately classifying patients with NCDs who are prone to non-adherence within a designated timeframe. All seven models built in the study demonstrate reliable and accurate predictions, achieving a minimum weighted average accuracy rate of 84.7% and an AUC value of 0.857. All seven models generated in our study outperform [19] ML models (RF, DNN, and support vector regression). Their ML models, which were developed to assess hypertension patients' adherence factors, produced only accuracies of less than 81%. Again, the RF model built by [53] to predict patient risk of non-adherence in T2D management only achieved a model accuracy of 0.73, whereas [21] deep learning model (CNN) for detecting adherence in T2D patients achieved an average accuracy of less than 79%. Son et al. (2010) used SVM for MA prediction and attained an accuracy of 77.6%. The observed satisfactory performance in our study can be attributed to the careful selection of a feature set comprising features that have a significant impact on predicting MA. It is noteworthy that in our study, the RF algorithm exhibited consistent superiority across all evaluated metrics. The findings indicate that the RF

classifier is a suitable choice for effectively analysing intricate clinical data in the context of predicting MA in individuals with NCDs. RF, being an ensemble learning algorithm that utilises the bagging technique, is indeed a robust and valuable ML approach for addressing the MA classification problem. The robustness of ensemble learning stems from the amalgamation of numerous diverse classifiers, namely independent DTs, which together enhance each other's performance as a cohesive unit [23]. The results above are similar to those from [26], who found that ensemble diabetes models are better at predicting nephropathy and diabetic angiopathy problems than other predictive models.

6. Strengths and Limitations

A notable aspect of the study pertains to the development of a dataset that will be readily accessible to the public. The study presents a newly created dataset within the setting of a developing country, which can be utilised by future researchers in the analysis of MA among diabetic and hypertensive patients. The utilisation of seven different ML algorithms facilitated the opportunity to evaluate and contrast the efficacy of the different ML models. This analysis afterward enables the identification of the most powerful algorithm for classifying and predicting MA or non-adherence, based on the provided dataset and its associated features. However, there are certain limitations inherent in our study. Due to restricted access to public health-related datasets in Zimbabwe and the neighbouring region, data analysis focused solely on individuals who had enrolled in commercial medical insurance. At this juncture, it is premature to assert the generalisability of the findings due to their reliance on a restricted set of clinical data. One additional constraint is the utilisation of an imbalanced dataset in relation to the classes of the target variable (MA). The non-adherent group had a higher number of instances in comparison to the adherent group. It is important to acknowledge that in practical applications where one works with an actual real-world dataset, imbalanced datasets are by far the most common scenario. Indeed, many problems that are worth solving are inherently imbalanced. Again, we used patient pharmacy dispensing data as a surrogate for the MA measure. This phenomenon has the potential to result in an overestimation of MA, as there is a possibility that patients may not self-administer the medications following their collection.

7. Conclusion

This study undertook the development of three ML models to classify MA among individuals diagnosed with diabetes, hypertension, or both. Throughout the model-building process, we confirmed the predictive relationship between several variables and MA levels. Notable variables included the number of medical supplies dispensed to a patient annually, annual claims, patient age, wellness program subscription, annual contribution towards medical aid, and hospital coverage type. Furthermore, our study showcased the feasibility of constructing ML classification models using medical claims data. The RF and DNN classifiers outperformed SVM, DT, LR, Naïve Bayes, and k-NN models, with the RF model particularly excelling in minimizing classification errors between non-adherent and adherent patients. Comprehensive evaluations of metrics such as AUC, F1 score, accuracy, precision, and recall support this conclusion. The RF model holds promise for effectively detecting medication non-adherence and offering targeted follow-up interventions, ultimately enhancing the effectiveness of clinical procedures. It has significant potential for guiding the development of intelligent MA and intervention applications in the context of chronic diseases, with applicability in clinical practice. Our findings contribute to understanding and addressing disparities in prescription refill rates, adherence, and attrition among individuals with different NCDs. Future research should prioritise the development of more extensive ML systems that consolidate data from diverse patient populations, with a focus on national health institutions and major referral hospitals. This would further advance the field and improve MA management in chronic disease contexts.

Ethical Statements

The article adhered to all ethical guidelines, after submitting a written request to the management of the Cimas Medical Aid Society to gather data. The authors implemented procedures to safeguard the confidentiality of the Cimas Organisation and its clients. They reached an agreement with Cimas' ICT department to de-identify the initial raw dataset before using it in their research and publishing. Personal data underwent a de-identification procedure to ensure the protection of sensitive patient information and adhere to confidentiality standards. Participants below the age of 16 were excluded from the dataset, as youth under 16 cannot make healthcare decisions without their parents' consent. The authors have read and followed the ethical requirements for publication, confirming that the current work does not directly involve human subjects, animal experiments, or any data collected from social media platforms.

Acknowledgements
The authors express their gratitude to the management of the CIMAS Medical Aid Society in Zimbabwe for providing the dataset utilised in this research work.

Author Contributions
W.K. contributed to the conception, planning and design of the study, coding, data analysis, data curation, original drafting of the manuscript and review. A.E.E.: drafting of the manuscript, supervision, review, validation and editing. Tsitsi Moyo: data acquisition and preparation. J.V.F.D.: editing. All authors approved the manuscript.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interest
The authors assert that they do not possess any identifiable conflicting financial interests or personal relationships that may have potentially influenced the findings presented in this research article.

Data and Code Availability
Datasets and codes are available in the online repository:
https://www.kaggle.com/datasets/dd353efc2f392e03a781933a42c0e326ceda8639cfa21874990e16b7aa6f65d1.

References

[1] World Health Organisation, "Noncommunicable diseases." [Online]. Available: https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/ncd-mortality#:~:text=Noncommunicable diseases (NCDs) kill 41,- and middle-income countries.

[2] S. Swed *et al.*, "Assessment the awareness toward hypertension and diabetes mellitus: Syrian cross sectional study," *BMC Public Health*, vol. 23, no. 1, p. 789, 2023.

[3] A. Odunyemi, T. Rahman, and K. Alam, "Economic burden of non-communicable diseases on households in Nigeria: evidence from the Nigeria living standard survey 2018-19," *BMC Public Health*, vol. 23, no. 1, p. 1563, 2023.

[4] P. T. Thanh and P. B. Duong, "The economic burden of non-communicable diseases on households and their coping mechanisms: Evidence from rural Vietnam," *World Dev.*, vol. 151, p. 105758, 2022.

[5] X. Yang, J. Yan, S. Lai, C. Shen, and N. Duan, "What affects the direct economic burden of non-communicable diseases on middle-aged and older adult people in Shaanxi Province?," *Front. Public Heal.*, vol. 11, p. 1219199, 2023.

[6] World Health Organisation, "Hypertension." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hypertension

[7] K. T. Mills, A. Stefanescu, and J. He, "The global epidemiology of hypertension," *Nat. Rev. Nephrol.*, vol. 16, no. 4, pp. 223–237, 2020.

[8] F. M. Algabbani and A. M. Algabbani, "Treatment adherence among patients with hypertension: findings from a cross-sectional study," *Clin. Hypertens.*, vol. 26, no. 1, pp. 1–9, 2020.

[9] S. M. Hamrahian, "Medication non-adherence: a major cause of resistant hypertension," *Curr. Cardiol. Rep.*, vol. 22, pp. 1–7, 2020.

[10] I. Ahmed *et al.*, "Medication adherence apps: review and content analysis," *JMIR mHealth uHealth*, vol. 6, no. 3, p. e6432, 2018.

[11] D. M. Berwick and M. E. Gaines, “How HIPAA harms care, and how to stop it,” *Jama*, vol. 320, no. 3, pp. 229–230, 2018.

[12] A. A. Abujaber, A. J. Nashwan, and A. Fadlalla, “Enabling the adoption of machine learning in clinical decision support: A total interpretive structural modeling approach,” *Informatics Med. Unlocked*, vol. 33, p. 101090, 2022.

[13] S. Sajid, “A methodology to build interpretable machine learning models in organizations.” University of Twente, 2023.

[14] J. Goecks, V. Jalili, L. M. Heiser, and J. W. Gray, “How machine learning will transform biomedicine,” *Cell*, vol. 181, no. 1, pp. 92–101, 2020.

[15] M. Owusu-Adjei, J. Ben Hayfron-Acquah, F. Twum, and G. Abdul-Salaam, “Machine Learning Modeling of Disease Treatment Default: A Comparative Analysis of Classification Models,” *Adv. Public Heal.*, vol. 2023, 2023.

[16] H.-Y. Chu *et al.*, “A predictive model for identifying low medication adherence among older adults with hypertension: A classification and regression tree model,” *Geriatr. Nurs. (Minneap).*, vol. 42, no. 6, pp. 1309–1315, 2021.

[17] W. Gao *et al.*, “A clinical prediction model of medication adherence in hypertensive patients in a Chinese community hospital in Beijing,” *Am. J. Hypertens.*, vol. 33, no. 11, pp. 1038–1046, 2020.

[18] Y. Li *et al.*, “Decoding nonadherence to hypertensive medication in New York City: a population segmentation approach,” *J. Prim. Care Community Health*, vol. 10, p. 2150132719829311, 2019.

[19] F. Aziz, S. Malek, A. M. Ali, M. S. Wong, M. Mosleh, and P. Milow, “Determining hypertensive patients’ beliefs towards medication and associations with medication adherence using machine learning methods,” *PeerJ*, vol. 8, p. e8286, 2020.

[20] A. Galozy and S. Nowaczyk, “Prediction and pattern analysis of medication refill adherence through electronic health records and dispensation data,” *J. Biomed. Inform.*, vol. 112, p. 100075, 2020.

[21] D. N. Thyde, A. Mohebbi, H. Bengtsson, M. L. Jensen, and M. Mørup, “Machine learning-based adherence detection of type 2 diabetes patients on once-daily basal insulin injections,” *J. Diabetes Sci. Technol.*, vol. 15, no. 1, pp. 98–108, 2021.

[22] A. Mohebbi, T. B. Aradóttir, A. R. Johansen, H. Bengtsson, M. Fraccaro, and M. Mørup, “A deep learning approach to adherence detection for type 2 diabetics,” in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2017, pp. 2896–2899.

[23] Y. Gu *et al.*, “Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data,” *Sci. Rep.*, vol. 11, no. 1, p. 18961, 2021.

[24] D. Fozoonmayeh *et al.*, “A scalable smartwatch-based medication intake detection system using distributed machine learning,” *J. Med. Syst.*, vol. 44, pp. 1–14, 2020.

[25] M. B. Alazzam, H. Mansour, F. Alassery, and A. Almulihi, “Machine learning implementation of a diabetic patient monitoring system using interactive E-app,” *Comput. Intell. Neurosci.*, vol. 2021, 2021.

[26] Y. Fan, E. Long, L. Cai, Q. Cao, X. Wu, and R. Tong, “Machine learning approaches to predict risks of diabetic complications and poor glycemic control in nonadherent type 2 diabetes,” *Front. Pharmacol.*, p. 1485, 2021.

[27] H. Kurasawa *et al.*, “Treatment Discontinuation Prediction in Patients With Diabetes Using a Ranking Model: Machine Learning Model Development,” *JMIR Bioinforma. Biotechnol.*, vol. 3, no. 1, p. e37951, 2022.

[28] S. Kurdi *et al.*, “Proof-of-Concept Study of Using Supervised Machine Learning Algorithms to Predict Self-Care and Glycemic Control in Type 1 Diabetes Patients on Insulin Pump Therapy,” *Endocr. Pract.*, vol. 29, no. 6, pp. 448–455, 2023.

[29] M. Li *et al.*, “Development and assessment of novel machine learning models to predict medication non-adherence risks in type 2 diabetics,” *Front. Public Heal.*, vol. 10, p. 1000622, 2022.

[30] X.-W. Wu, H.-B. Yang, R. Yuan, E.-W. Long, and R.-S. Tong, “Predictive models of medication non-adherence risks of patients with T2D based on multiple machine learning algorithms,” *BMJ Open Diabetes Res. Care*, vol. 8, no. 1, p. e001055, 2020.

[31] A. L. Beam and I. S. Kohane, “Big data and machine learning in health care,” *Jama*, vol. 319, no. 13, pp. 1317–1318, 2018.

[32] L. Wang *et al.*, “Applying machine learning models to predict medication nonadherence in Crohn’s disease maintenance therapy,” *Patient Prefer. Adherence*, pp. 917–926, 2020.

[33] S. K. Lee, B.-Y. Kang, H.-G. Kim, and Y.-J. Son, “Predictors of medication adherence in elderly patients with chronic diseases using support vector machine models,” *Healthc. Inform. Res.*, vol. 19, no. 1, pp. 33–41, 2013.

[34] A. Gottlieb, A. Yatsco, C. Bakos-Block, J. R. Langabeer, and T. Champagne-Langabeer, “Machine Learning for Predicting Risk of Early Dropout in a Recovery Program for Opioid Use Disorder,” in *Healthcare*, MDPI, 2022, p. 223.

[35] I. S. Stafford, M. M. Gosink, E. Mossotto, S. Ennis, and M. Hauben, “A systematic review of artificial intelligence and machine learning applications to inflammatory bowel disease, with practical guidelines for interpretation,” *Inflamm. Bowel Dis.*, vol. 28, no. 10, pp. 1573–1583, 2022.

[36] Y.-J. Son, H.-G. Kim, E.-H. Kim, S. Choi, and S.-K. Lee, “Application of support vector machine for prediction of medication adherence in heart failure patients,” *Healthc. Inform. Res.*, vol. 16, no. 4, pp. 253–259, 2010.

[37] M. M. Hasan, G. J. Young, J. Shi, P. Mohite, L. D. Young, and S. G. Weiner, “A machine learning based two-stage clinical decision support system for predicting patients’ discontinuation from opioid use disorder treatment: retrospective observational study,” *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, pp. 1–21, 2021.

[38] S. N. Lappan, A. W. Brown, and P. S. Hendricks, “Dropout rates of in-person psychosocial substance use disorder treatments: a systematic review and meta-analysis,” *Addiction*, vol. 115, no. 2, pp. 201–217, 2020.

[39] S. Comer *et al.*, “National practice guideline for the use of medications in the treatment of addiction involving opioid use,” *Am Soc Addicit Med*, vol. 66, pp. 39–42, 2015.

[40] W. Kanyongo and A. E. Ezugwu, “Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives,” *Informatics Med. Unlocked*, p. 101232, 2023, doi: https://doi.org/10.1016/j.imu.2023.101232.

[41] T. Kassahun, H. Gesesew, L. Mwanri, and T. Eshetie, “Diabetes related knowledge, self-care behaviours and adherence to medications among diabetic patients in Southwest Ethiopia: a cross-sectional survey,” *BMC Endocr. Disord.*, vol. 16, pp. 1–11, 2016.

[42] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, “A comprehensive survey on support vector machine classification: Applications, challenges and trends,” *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[43] A. L. Samuel, “Some studies in machine learning using the game of checkers,” *IBM J. Res. Dev.*, vol. 3, no. 3, pp. 210–229, 1959.

[44] L. Zaadnoordijk, T. R. Besold, and R. Cusack, “Lessons from infant learning for unsupervised

machine learning," *Nat. Mach. Intell.*, vol. 4, no. 6, pp. 510–520, 2022.

[45] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science (80-.).*, vol. 349, no. 6245, pp. 255–260, 2015.

[46] F. Hahne, W. Huber, R. Gentleman, S. Falcon, R. Gentleman, and V. J. Carey, "Unsupervised machine learning," *Bioconductor case Stud.*, pp. 137–157, 2008.

[47] B. H. Menze *et al.*, "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC Bioinformatics*, vol. 10, pp. 1–16, 2009.

[48] N. Salmi and Z. Rustam, "Naïve Bayes classifier models for predicting the colon cancer," in *IOP conference series: materials science and engineering*, IOP Publishing, 2019, p. 52068.

[49] H. Midi, S. K. Sarkar, and S. Rana, "Collinearity diagnostics of binary logistic regression model," *J. Interdiscip. Math.*, vol. 13, no. 3, pp. 253–267, 2010.

[50] K. A. S. A. S. Biria, "M Malepour A Heydari H Determinants of patient's adherence to hypertension medications: application of health belief model among rural patients," *Ann Med Heal. Sci Res*, vol. 46, pp. 922–927, 2014.

[51] G. Yerrapragada, A. Siadimas, A. Babaeian, V. Sharma, and T. J. O'Neill, "Machine learning to predict tamoxifen nonadherence among US commercially insured patients with metastatic breast cancer," *JCO Clin. Cancer Informatics*, vol. 5, pp. 814–825, 2021.

[52] E. J. Kleczyk and D. Evans, "Leveraging Predictive Analytics to Derive Patient Adherence Drivers," *Pharm. Manag. Sci. Assoc.*, 2021.

[53] X. I. N. CHEN, G. FERNANDES, J. I. E. CHEN, Z. LIU, and R. BAUMGARTNER, "1311-P: Machine Learning (ML) Application to Predict Patient Risk of Nonadherence in Type 2 Diabetes Management Using US Claims Databases," *Diabetes*, vol. 68, no. Supplement_1, 2019.

Machine Learning Classification of Medication Adherence in NCD Patients: A Retrospective Study

Wellington Kanyongo[a,b], Absalom E. Ezugwu[c], Tsitsi Moyo[d], Jean Vincent Fonou Dombeu[e]

[a] *School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg Campus, Pietermaritzburg, 3201, South Africa.*
[b] *Department of Computer Science, Faculty of Science Engineering, Bindura University of Science Education, Bindura, Zimbabwe.*
[c] *Unit for Data Science and Computing, North-West University, 11 Hoffman Street, Potchefstroom, 2520, South Africa.*
[d] *Cimas Medical Aid Society, Borrowdale Office Park, Borrowdale Road, P. O Box 1243, Harare, Zimbabwe*
[e] *School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg Campus, Pietermaritzburg, 3201, South Africa.*

Email Addresses: *wkanyongo@gmail.com (W. Kanyongo), Absalom.Ezugwu@nwu.ac.za (A.E. Ezugwu), tmoyo@cimas.co.zw, tsitsimoyo13@gmail.com, (Tsitsi Moyo), and fonoudombeuj@ukzn.ac.za (Jean Vincent Fonou Dombeu).*

General Information

This supplementary document contains visualisations and results for the medication adherence classification models created as part of a research project. The project involved building seven machine learning models, including naïve Bayes, support vector machines (SVM), decision trees (DT), logistic regression (LR), random forest (RF), k-nearest neighbour (k-NN), and deep neural networks (DNN). We used these models to classify medication adherence among diabetes and hypertension patients. These supplemental materials are intended to clarify some of the findings from the scientific paper "Machine Learning Classification of Medication Adherence in NCD Patients: A Retrospective Study."

Supplementary Figure S1: ML classification reports. The classification reports were generated as part of the medication adherence models built using seven machine learning models: naïve bayes, support vector machines (SVM), decision trees (DT), logistic regression (LR), random forest (RF), k-nearest neighbor (k-NN), and deep neural networks (DNN).

Naive Bayes Classification report:

	precision	recall	f1-score	support
Non Adherent	0.8718	0.9539	0.9110	1954
Adherent	0.7049	0.4397	0.5416	489
accuracy			0.8510	2443
macro avg	0.7884	0.6968	0.7263	2443
weighted avg	0.8384	0.8510	0.8371	2443

SVM Classification report:

	precision	recall	f1-score	support
Non Adherent	0.8817	0.9575	0.9181	1954
Adherent	0.7414	0.4867	0.5877	489
accuracy			0.8633	2443
macro avg	0.8116	0.7221	0.7529	2443
weighted avg	0.8536	0.8633	0.8519	2443

DT Classification report:

	precision	recall	f1-score	support
Non Adherent	0.9198	0.9090	0.9144	1956
Adherent	0.6510	0.6817	0.6660	487
accuracy			0.8637	2443
macro avg	0.7854	0.7954	0.7902	2443
weighted avg	0.8662	0.8637	0.8649	2443

LR Classification report:

	precision	recall	f1-score	support
Non Adherent	0.8628	0.9621	0.9098	1954
Adherent	0.7197	0.3885	0.5046	489
accuracy			0.8473	2443
macro avg	0.7912	0.6753	0.7072	2443
weighted avg	0.8341	0.8473	0.8287	2443

RF Classification report:

	precision	recall	f1-score	support
Non Adherent	0.908	0.941	0.924	1954
Adherent	0.724	0.618	0.667	489
accuracy			0.876	2443
macro avg	0.816	0.779	0.795	2443
weighted avg	0.871	0.876	0.873	2443

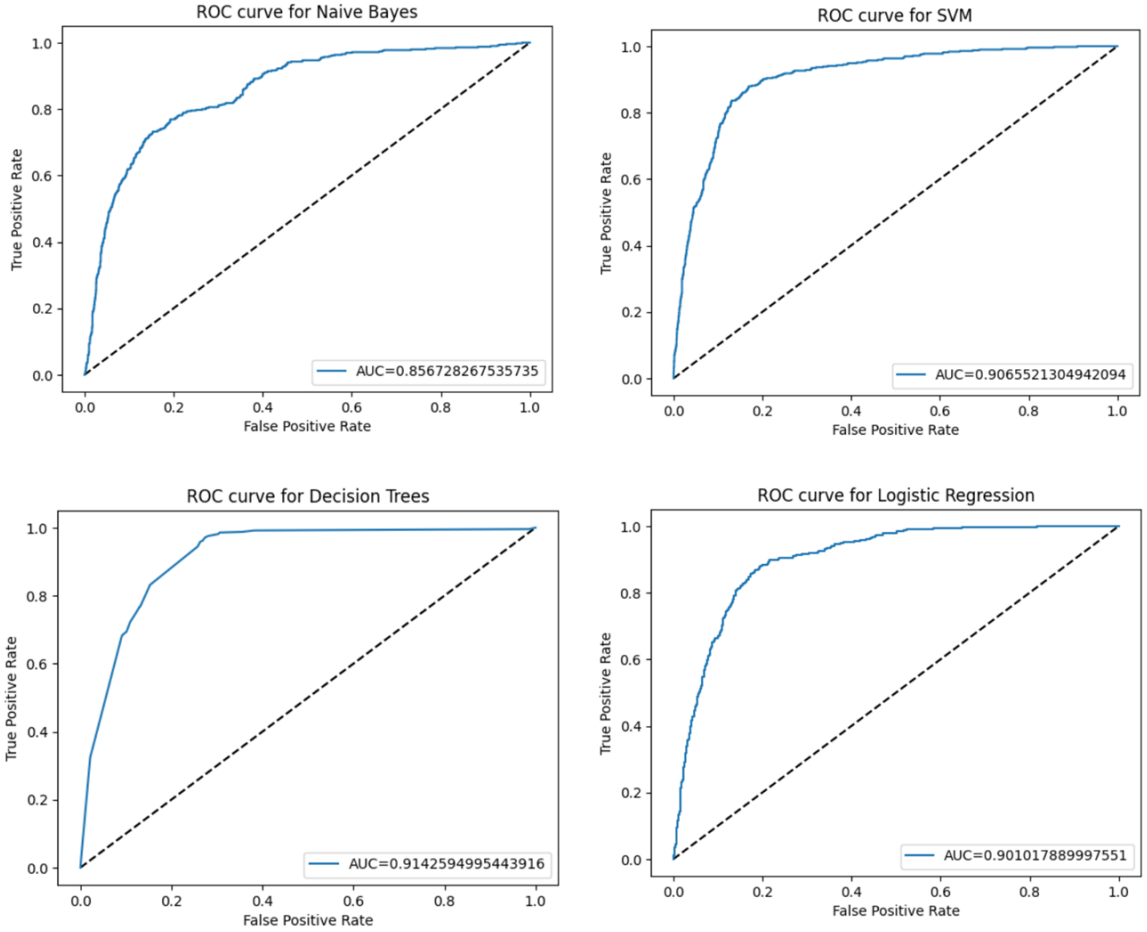
k-NN Classification report:

	precision	recall	f1-score	support
Non Adherent	0.8800	0.9498	0.9136	1954
Adherent	0.7066	0.4826	0.5735	489
accuracy			0.8563	2443
macro avg	0.7933	0.7162	0.7436	2443
weighted avg	0.8453	0.8563	0.8455	2443

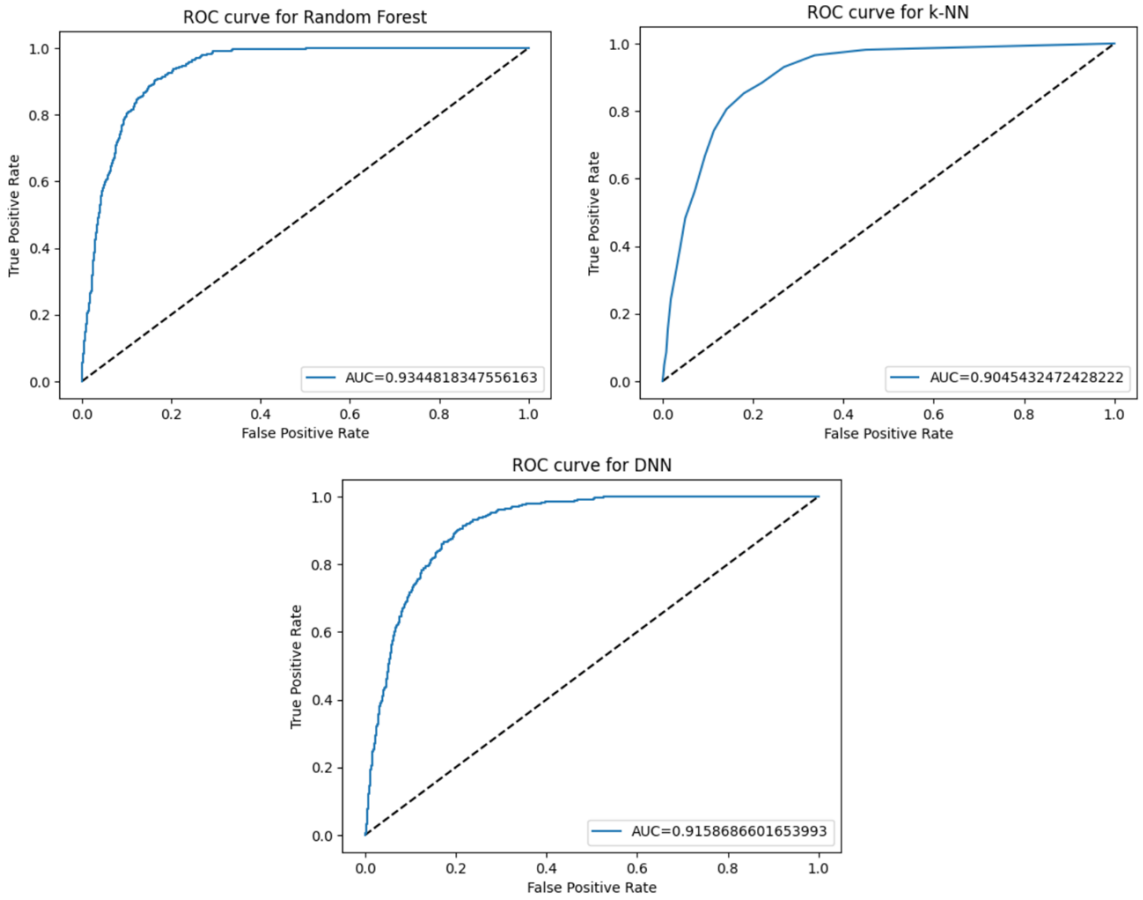
DNN Classification report:

	precision	recall	f1-score	support
Non Adherent	0.9010	0.9360	0.9182	1954
Adherent	0.6973	0.5890	0.6386	489
accuracy			0.8666	2443
macro avg	0.7992	0.7625	0.7784	2443
weighted avg	0.8602	0.8666	0.8622	2443

Supplementary Figure S2: ROC curves for the naïve bayes, SVM, DT and LR. These are ROC curves produced for the medication adherence models built using four different machine learning models (Naïve bayes, SVM, DT and LR). These curves demonstrate the classification models' performance at all classification thresholds. The curves' plots have two parameters: True Positive Rate. False Positive Rate.



Supplementary Figure S3: ROC curves for the RF, k-NN and DNN. These are the ROC curves for medication adherence models developed using three different machine learning models (RF, k-NN, and DNN). These visualisations represent the classification model's performance across all classification thresholds. The curve plots feature two parameters: True Positive Rate. False positive rate.



Supplementary Figure S4: Logistic Regression Analysis Results. The outputs are for a medication adherence model generated with logistic regression (LR). The output consists of the dependent and independent variables, the number of observations in the training set, variable coefficients, p-values, and the model's odds ratio.

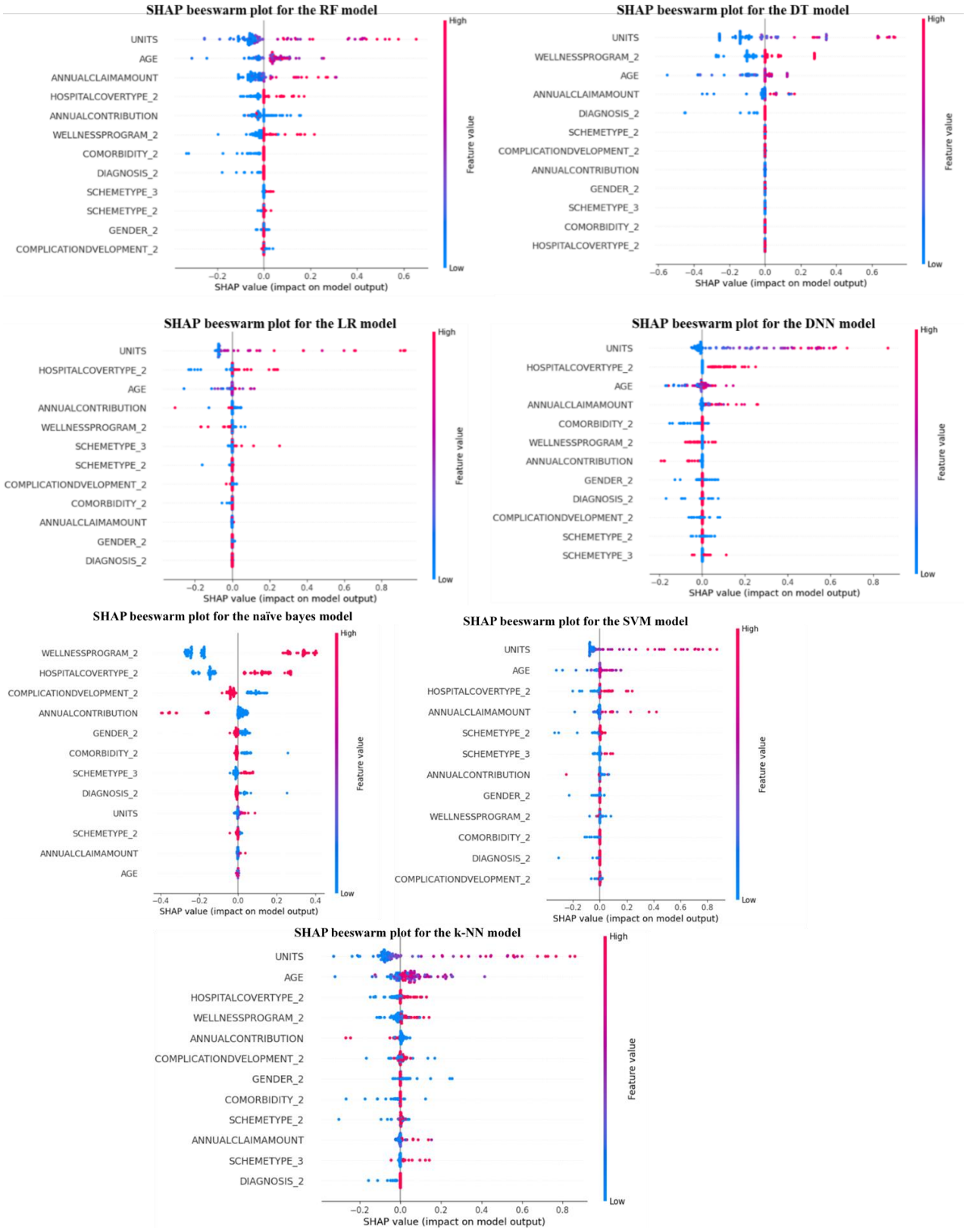
Logit Regression Results

Dep. Variable:	MED_ADHERENCE	No. Observations:	5698
Model:	Logit	Df Residuals:	5685
Method:	MLE	Df Model:	12
Date:	Sun, 17 Sep 2023	Pseudo R-squ.:	0.2876
Time:	15:52:44	Log-Likelihood:	-1941.6
converged:	True	LL-Null:	-2725.6
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.9056	0.306	-9.496	0.000	-3.505	-2.306
AGE	0.3780	0.045	8.442	0.000	0.290	0.466
ANNUALCONTRIBUTION	-1.2569	0.661	-1.900	0.057	-2.553	0.039
ANNUALCLAIMAMOUNT	0.0895	0.034	2.623	0.009	0.023	0.156
UNITS	1.4384	0.053	27.086	0.000	1.334	1.542
DIAGNOSIS_2	-0.0259	0.118	-0.220	0.826	-0.257	0.205
SCHEMETYPE_2	0.1062	0.156	0.679	0.497	-0.200	0.413
SCHEMETYPE_3	0.4263	0.230	1.856	0.063	-0.024	0.876
HOSPITALCOVERTYPE_2	3.9598	2.159	1.834	0.067	-0.272	8.192
WELLNESSPROGRAM_2	-2.8579	2.144	-1.333	0.183	-7.061	1.345
GENDER_2	-0.0437	0.084	-0.522	0.602	-0.208	0.120
COMPLICATIONDVELOPMENT_2	-0.1230	0.084	-1.460	0.144	-0.288	0.042
COMORBIDITY_2	0.2561	0.106	2.419	0.016	0.049	0.464

	OR	Lower CI	Upper CI
const	0.054714	0.030037	0.099665
AGE	1.459407	1.336776	1.593286
ANNUALCONTRIBUTION	0.284526	0.077824	1.040225
ANNUALCLAIMAMOUNT	1.093627	1.022872	1.169276
UNITS	4.213755	3.797239	4.675959
DIAGNOSIS_2	0.974434	0.773419	1.227693
SCHEMETYPE_2	1.112084	0.818356	1.511237
SCHEMETYPE_3	1.531636	0.976461	2.402461
HOSPITALCOVERTYPE_2	52.447023	0.761585	3611.799093
WELLNESSPROGRAM_2	0.057392	0.000858	3.839174
GENDER_2	0.957226	0.812380	1.127898
COMPLICATIONDVELOPMENT_2	0.884250	0.749659	1.043004
COMORBIDITY_2	1.291929	1.049816	1.589880

Supplementary Figure S4: SHAP beeswarm plots. The beeswarm plots display an information-dense summary of how the top features in a dataset are impacting the models' output. The SHAP plots show the features with the biggest impact on the models' predictions.



5.2 Chapter Summary

Before developing and comparing the various Ml models, significant predictors of MA were identified through the RF feature importance scoring mechanism. These included the annual quantity of medical supplies, annual claim amount, patient age, wellness program subscription, medical aid cover, contribution towards medical aid cover, comorbidity, diagnosis, hospital cover type, complications development, gender, and medical aid scheme. The total units of medical supplies dispensed annually emerged as the most significant predictor of MA. Classification accuracy of the ML classifiers ranged from 84.7% to 87.6%, with AUC values from 0.857 to 0.934. RF, an ensemble learning technique, was the most robust classifier, achieving 87.6% accuracy, an AUC of 0.9351, and superior precision, recall, and F1-score. This model shows promise as a prognostic tool for enhancing MA, aiding in identifying adherence levels among patients. These findings contribute to addressing disparities in medication refilling and adherence rates among patients with NCDs. The ML model holds potential for the development of intelligent MA and intervention applications to improve patient adherence to medication in the chronic disease domain. The subsequent chapter entails developing twinned machine learning algorithms for medication adherence analytics in patients with diabetes and hypertension.

CHAPTER 6

Development and validation of a twin machine learning algorithm for medication adherence analytics in patients with diabetes and hypertension

(Paper 5)

6.1 Introduction

The paper in this chapter aimed to improve model performance in classifying medication adherence by developing two hybrid machine learning models: TWINNED-SVM-RF and TWINNED-SVM-DT. This innovation involved integrating Support Vector Machines (SVM) with tree-based algorithms—Decision Trees (DT) and Random Forests (RF)—to enhance the classification of medication adherence in patients with diabetes and hypertension. The study used a publicly available dataset containing 8,141 records of patients with these conditions, covering 13 variables. The dataset was split into a training set (70%) and a testing set (30%). Initially, independent SVM and RF models were created, followed by hyperparameter tuning through GridSearchCV with k-fold cross-validation to optimize performance. The best-performing models were then hybridized using a twinning procedure, including weight tuning, to produce the TWINNED-SVM-RF and TWINNED-SVM-DT classifiers. A comparative analysis was conducted between the classical SVM model and the newly developed TWINNED-SVM-RF and TWINNED-SVM-DT models, evaluating their performance using metrics such as accuracy, AUC, precision, recall, and F1-score. Additionally, LIME (Local Interpretable Model-agnostic Explanations) was employed to provide insights into the decision-making process of the best-performing twinned ML classifier, with the LIME results explained for two selected instances in the dataset.

A twin machine learning algorithm for medication adherence analytics in patients with diabetes and hypertension

Wellington Kanyongo[1], Absalom E. Ezugwu[2[0000-0002-3721-3400],*]
[1] School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg Campus, Pietermaritzburg, 3201, South Africa
[2] Unit for Data Science and Computing, North-West University, 11 Hoffman Street, Potchefstroom, 2520, South Africa

Abstract. In an era where prediction and classification accuracy are crucial for medication adherence (MA) analytics, leveraging hybrid machine learning (ML) models holds the potential to revolutionize data-driven decision-making. This research seeks to enhance model performance in classifying medication adherence by developing two hybrid ML models, named TWINNED-SVM-RF and TWINNED-SVM-DT. These models integrate Support Vector Machines (SVM) with tree-based algorithms—specifically Decision Trees (DT) and Random Forests (RF)—for MA analytics in patients with diabetes and hypertension. The study utilized a publicly available dataset consisting of 8,141 records of patients with diabetes and hypertension, encompassing 13 variables. The dataset was divided into a training set (70%) and a testing set (30%). Initially, independent SVM and RF models were developed, followed by hyperparameter tuning using GridSearchCV with k-fold cross-validation to optimize performance. The best-performing models were then hybridized through a twinning procedure that included weight tuning, resulting in the TWINNED-SVM-RF and TWINNED-SVM-DT classifiers. Model performance was assessed using metrics such as accuracy, AUC, precision, recall, and F1-score. The results indicated that both hybrid models outperformed the classical SVM, with the TWINNED-SVM-RF model achieving an AUC score of 0.923, compared to 0.895 for SVM. Additionally, the application of LIME (Local Interpretable Model-Agnostic Explanations) enhanced the explainability of the models, providing valuable insights into their decision-making processes. Overall, the TWINNED-SVM-RF model demonstrates superior performance, making it a powerful tool for MA analytics and other diverse applications.

Keywords: Machine learning, medication adherence, non-communicable diseases, diabetes, hypertension.

1. Introduction

Global health statistics reveal that non-communicable diseases (NCDs) are responsible for the deaths of 41 million people annually, accounting for approximately 74% of total global mortality (WHO, 2022). According to the World Health Organization (WHO, 2023), around 1.28 billion people aged 30 to 79 worldwide suffer from hypertension, with two-thirds of this population residing in low- and middle-income countries. In the case of diabetes, at least 422 million individuals are affected, with the majority also living in low- and middle-income nations. Diabetes directly contributes to 1.5 million deaths each year (WHO, 2022). Amid these alarming statistics, non-adherence to NCD medications has emerged as a widespread issue with serious consequences (Algabbani & Algabbani, 2020; Hamrahian, 2020). Research by WHO on medication adherence (MA) for chronic conditions shows that patients in developed countries adhere to only 50% of their prescribed medications, with adherence rates even lower in developing nations (Ahmed et al., 2018). Given the severity of the non-adherence crisis, it is crucial to accurately detect non-adherence at an early stage. This is essential for ensuring optimal resource allocation to those who need more care and for tailoring interventions that address the unique factors contributing to medication non-adherence.

The widespread availability and accessibility of electronic health records (EHR) have facilitated efforts to detect and analyze medication adherence or non-adherence. Given the vastness of EHR datasets and the shortage of medical professionals in the healthcare sector, manually examining a patient's medical history and assessing outcomes can be time-

consuming (Berwick & Gaines, 2018). However, machine learning (ML) in medicine and public health offers the potential to overcome these challenges by efficiently managing large and complex datasets and providing real-time, cost-effective evidence (Abujaber et al., 2022). Goecks et al. (2020) emphasize that modern clinical decision support tools incorporating ML algorithms have shown considerable promise in areas such as disease classification, risk prediction, dose recommendation, pattern recognition, precision treatment, and health monitoring.

The growing complexity of machine learning tasks and the increasing availability of large-scale datasets demand sophisticated models capable of managing these challenges and delivering optimal results efficiently. While no single algorithm is universally superior across all tasks, the limitations of individual models can be overcome by combining their strengths through hybridization. As a result, hybridization techniques are widely regarded as a highly effective approach for tackling complex machine-learning problems. By harnessing the complementary strengths of different methods, hybrid models can enhance performance, increase accuracy, and provide more robust solutions to intricate data challenges (Azevedo et al., 2024). However, many researchers have primarily relied on individual classical ML algorithms to build stand-alone models for medication adherence (MA) classification.

Koesmahargyo et al. (2020) explored the accuracy of ML in predicting medication adherence (MA) using extreme gradient boosting. Hezarjaribi et al. (2016) monitored medication adherence solely with decision trees. Wu et al. (2020) applied various ML approaches, including decision trees (DT), logistic regression (LR), Bayesian networks, KNN, SVM, and neural networks, to evaluate MA. Chen et al. (2024) developed a prediction model for medication refill adherence using extreme gradient boosting. Mirzadeh et al. (2022) employed decision-tree-based methods and baseline models like linear regression to predict MA in atherosclerotic cardiovascular disease patients individually. Gu et al. (2021) developed MA models using large-scale healthcare data by separately applying tree-based and deep-learning algorithms. Warren et al. (2022) analyzed medication adherence using four distinct binary classifiers: LR, DT, random forests (RF), and XGBoost.

While various machine learning techniques have been employed to build medication adherence (MA) models, there is a notable lack of research on hybrid machine learning models specifically designed for MA classification and analytics. Son et al. (2010) utilized support vector machines (SVM) to develop a predictive model for medication adherence in heart failure patients, achieving a maximum accuracy rate of 78%. The current study seeks to improve the performance of the classical SVM algorithm by developing a novel twinned algorithm that combines SVM with tree-based machine learning techniques for MA analytics. Recent systematic reviews by Kanyongo and Ezugwu (2023) have confirmed that tree-based algorithms such as Random Forest (RF) and Decision Trees (DT) are among the most popular and top-performing ML algorithms for MA analytics. In response, this study uniquely hybridizes SVM with these tree-based algorithms to build MA classifiers for patients with diabetes and hypertension. The study is guided by the following research objectives:

i. To develop a hybrid machine learning algorithm that integrates Support Vector Machines (SVM) with Decision Trees (DT) and Random Forests (RF) for medication adherence analytics in patients with diabetes and hypertension.

ii. To evaluate the performance of the classical SVM compared to the newly developed TWINNED SVM-DT and TWINNED SVM-RF models for medication adherence classification.

2. Methods and Materials

This section describes the dataset used, model development procedures, the tools for development, and the architectural design showing the systematic development and evaluation of the twinned ML models.

2.1 The Dataset

The dataset used, formatted in Comma-Separated Values (CSV), is publicly available online and originally sourced from CIMAS, a medical insurance provider based in Zimbabwe. It includes patient-level information and medical claims data, with numerical variables scaled and categorical data encoded. The dataset also contains medication refill thresholds, which were transformed into a categorical medication adherence variable with two classes: "adherent" and "non-adherent." The data pertains to members aged 16 to 101 who had been diagnosed with essential diabetes (ICD-10 Code: E10-) and hypertension (ICD-10 Code: I10-). The dataset covers a full calendar year, specifically from January 1 to December 31, 2022.

Adherence metrics measure the percentage of patients covered by prescription claims for the same medication within the same therapeutic class during the measurement year. This dataset comprises data exclusively from 2022, spanning the entire year. It includes 13 variables and 8,141 rows of data. The input features are age, annual contribution (subscription payments to the medical aid company), annual claim amount (amount claimed to cover medical bills), units, diagnosis (the NCD condition, either diabetes or hypertension), scheme type (the medical aid package subscribed to), hospital cover type, wellness program, gender, complication development (whether complications related to diabetes or hypertension developed during the year), and comorbidity (diagnosis of both diabetes and hypertension). The target variable is "medication adherence".

2.2 Model Development

Before model development, exploratory data analysis (EDA) was performed to report the descriptive statistics of the dataset features before any variable transformations and data encoding. Descriptive statistics included frequency counts, frequency percentages, mean, mode, standard deviation, and range. Python 3.11, in conjunction with Jupyter Notebook 6.5.4, was used for model development in this study. The dataset was divided into features (X) and the target variable (y), and then split into training (70%) and testing (30%) sets. The primary focus was on developing hybrid machine learning algorithms by integrating classical Support Vector Machines (SVM) with tree-based algorithms, specifically Decision Trees (DT) and Random Forests (RF). The objective was to create TWINNED-SVM-DT and TWINNED-SVM-RF models for classifying medication adherence among patients with diabetes and hypertension. The development process involved preliminary modeling with individual classifiers, namely classical SVM, DT, and RF models. Following this, the two hybrid models (TWINNED-SVM-DT and TWINNED-SVM-RF) were constructed, incorporating hyperparameter tuning. Details and parameters of the model development are described in the following subsections.

Support Vector Machines (SVM)

This study utilized the Radial Basis Function (RBF) kernel due to its flexibility and capability to handle various types of data distributions. The parameter C manages the trade-off between maximizing the margin and minimizing classification errors and was tuned using GridSearchCV with values [50, 10, 1.0, 0.1, 0.01] to balance bias and variance. The gamma parameter, which determines the influence of a single training example, was set to 'scale' by default for the RBF kernel, adjusting automatically based on the data. Cross-validation, specifically 5-fold, was employed in GridSearchCV to ensure robust and unbiased hyperparameter tuning. This process systematically explored different parameter combinations to optimize the model's performance. GridSearchCV identified the best parameters, and the SVM classifier was initialized with the RBF kernel and these optimal C and gamma values. The model was then retrained using these best parameters.

Random Forest

The development of the Random Forest (RF) model involved tuning several parameters. The n_estimators parameter, which determines the number of trees in the forest, was evaluated with values of [100, 200, 300] to find the optimal number. The max_depth parameter, which controls the maximum depth of each tree, was tested with values of [None, 10, 20, 30] to prevent overfitting by limiting the model's complexity. For the min_samples_split parameter, which specifies the minimum number of samples required to split an internal node, values of [2, 5, 10] were used to ensure that splits are only made when they are statistically significant. In this study, these parameters were tuned using GridSearchCV, which systematically explored the parameter grid to identify the best combination of hyperparameters. The RF model development process employed 5-fold cross-validation within GridSearchCV to ensure robust and unbiased tuning. The Random Forest model was initialized and trained using GridSearchCV with 5-fold cross-validation and the specified parameter grid.

TWINNED-SVM-DT

Among the two hybrid machine learning models, TWINNED-SVM-DT was developed by combining Support Vector Machines (SVM) and Decision Trees (DT) to assess whether this new hybrid model improves medication adherence (MA) classification accuracy and other metrics. The development of TWINNED-SVM-DT began with independently tuning the hyperparameters of the SVM and DT models using GridSearchCV with 5-fold cross-validation to ensure robust and unbiased optimization. After identifying the best hyperparameters for each model, SVM and DT were trained separately. The TWINNED-SVM-DT model was then created by combining the best estimators from GridSearchCV results for both SVM and DT using a VotingClassifier with soft voting. This hybrid approach allows the model to leverage the complementary strengths of SVM, which excels in high-dimensional and non-linear classification, and DT, which offers interpretability and handles feature correlations effectively. To further enhance the performance of the TWINNED-SVM-DT model for MA analytics, additional tuning was performed using a parameter grid to determine the optimal weights for integrating the predictions from SVM and DT.

TWINNED-SVM-RF

In another instance of hybridizing the classical SVM with a tree-based machine learning algorithm, the TWINNED-SVM-RF model was developed. The creation of TWINNED-SVM-RF began with tuning the hyperparameters of both SVM and Random Forest (RF) models using GridSearchCV with 5-fold cross-validation to ensure a thorough and unbiased optimization process. After determining the optimal hyperparameters for each model, the SVM and RF models were trained separately. Subsequently, TWINNED-SVM-RF was constructed by combining the best estimators from the GridSearchCV results for both SVM and RF using a VotingClassifier with soft voting. This hybrid model harnesses the strengths of SVM in handling high-dimensional and non-linear data and the advantages of RF in managing large datasets and reducing overfitting through ensemble learning. The model twinning process also involved fine-tuning by exploring different weight combinations to achieve the optimal integration of predictions from both SVM and RF. This approach aims to enhance overall model performance and prediction accuracy by effectively leveraging the capabilities of both classifiers.

2.3 Architectural design for twinning SVM and a tree-based ML algorithm

The architectural design of the TWINNED-SVM-RF model consists of five layers to systematically develop and evaluate the hybrid machine learning model.

- Data Layer: The dataset is first processed through feature selection, followed by splitting into training and test sets.
- Model Building Layer: In this layer, the SVM and Random Forest (RF) models are developed and optimized independently using GridSearchCV with 5-fold cross-validation to identify the best hyperparameters.
- Models Twinning Layer: The optimized SVM and RF models are then combined through a twinning process. This includes weight tuning to refine the integration of predictions from both models.
- Evaluation Layer: The performance of the hybrid model is assessed using various evaluation metrics to measure its effectiveness.
- Explainability Layer: To ensure transparency and understanding, techniques such as Local Interpretable Model-Agnostic Explanations (LIME) or SHapley Additive exPlanations (SHAP) are employed to provide insights into the model's decision-making process.

This comprehensive architectural design ensures that the TWINNED-SVM-RF model is robust, accurate, and interpretable. Figure 1 illustrates the diagrammatic architectural design for twinning SVM with a tree-based ML algorithm.

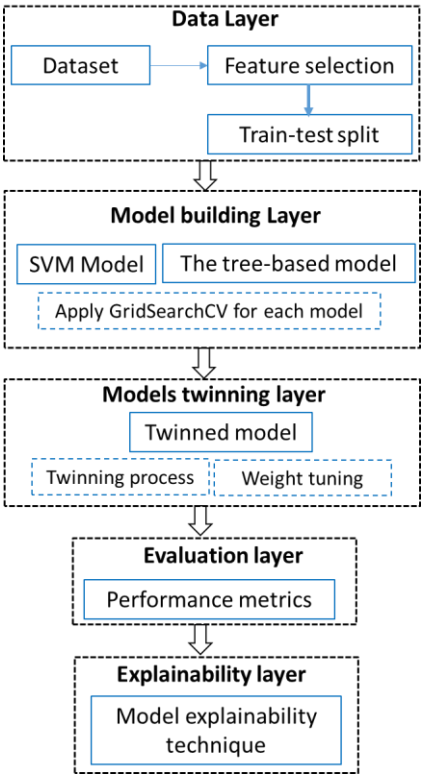


Fig. 1: Architectural design for the TWINNED-SVM-RF/DT model

3. Results

This section presents the research findings that directly address the research objectives. The study aimed to develop a hybrid machine learning algorithm by integrating Support Vector Machines (SVM) with Decision Trees (DT) and Random Forests (RF) for medication adherence classification in patients with diabetes and hypertension. For this classification problem, patients were considered adherent if their medication adherence level met or exceeded a predetermined refill threshold of 75%, which corresponds to at least 9 out of 12 months. Patients falling below this threshold were classified as non-adherent. This setup defines the problem as a binary classification task. The performance of the developed TWINNED-SVM-DT and TWINNED-SVM-RF models was compared against that of an individual SVM model for medication adherence classification. Before presenting the model results, the findings from the exploratory data analysis (EDA) are described.

3.1 Exploratory data analysis

Table 5 provides an overview of the dataset, which includes 6,461 observations for hypertensive patients and 1,680 observations for diabetic patients. Of the total 8,141 patients, 4,768 (59%) are female and 3,373 (41%) are male. Insurance coverage details show that 53% of the Medical Aid Society members are enrolled in basic-to-grade-B hospital health insurance plans, while 47% have private-to-grade-A coverage.

Regarding comorbidity, 30% (2,402 out of 8,141) have both diabetes and hypertension simultaneously, while the remaining 70% (5,739) have only one of the conditions. Within this cohort, 70% of patients have either diabetes or hypertension, but not both, and 30% have a dual diagnosis. During the study period, 3,417 individuals (42% of the sample) sought specialized medical care for complications related to either diabetes or hypertension, while 5,724 patients (58%) did not report any related issues.

Adherence metrics reveal that only 19% of participants adhered to their prescribed medication regimen, defined as refilling prescriptions monthly for at least nine months in 2022, meeting 75% of the total refill requirement. The remaining 81% did not achieve this adherence level, failing to refill their medications at least 75% of the 12 times prescribed over the year. A detailed summary of the categorical and continuous variables relevant to the study is provided in Table 1 below.

Table 1 The summary statistics of the categorical features before data encoding (n = 8141)

Feature	Attributes	Value	Percentage (%)
Diagnosis	Diabetes	1680	21
	Hypertension	6461	79
	Mode	Hypertension	-
Gender	Female	4768	59
	Male	3373	41
	Mode	Female	-
Scheme type	Basic Level	1453	18
	Medium Level	5180	64
	Premium Level	1508	18
	Mode	Medium Level	-
Hospital cover type	Basic to Grade B	4281	53
	Private to Grade A	3860	47
	Mode	Basic to Grade B	-
Wellness program	Subscribed	3160	39
	Not subscribed	4981	61
	Mode	Not Subscribed	-
Complication Development	Yes	3417	42
	No	4724	58
	Mode	No	-
Cormobidity	Yes	2402	30
	No	5739	70
	Mode	No	-
Adherence	Adherent	1541	19
	Non-Adherent	6600	81
	Mode	Non-Adherent	-
Age	Mean + SD	53 + 14	-
	Age range	16-101	-
	Median	52	-
Annual contribution (US$)	Mean + SD	3206 + 7307	-
	Range	84-25452	-
	Median	768	-
Annual claim amount	Mean + SD	222 + 1128	-
	Age range	0.10-40325	-
	Median	69	-
Units	Mean + SD	363 + 461	-
	Range	1-13140	-
	Median	216	-

In additional descriptive statistics, Table 6 presents the mean values and standard deviations (SDs) for the continuous variables within the dataset before any encoding or scaling procedures. Regarding the age variable, the dataset reveals a span from a minimum age of 16 years to a maximum of 101 years. The mean age of the sampled population stands at 53 (SD = 14). Additionally, the average annual financial contribution per member to the Medical Aid Society was $3,206.00, while the average annual claim amount was $222.00.

3.2 Model Results

The results of twinning SVM with DT and RF taking into account hyperparameter tuning of the individual and twinned models are depicted in Table 2.

Table 2 SVM, TWINNED SVM-RF, and TWINNED-SVM-DT Results

Model	Precision (%)	Recall (%)	F-Score	Accuracy	AUC
	Weighted Average Statistics				
SVM	0.8442	0.8551	0.8464	0.8551	0.8947
TWINNED SVM-RF	0.8484	0.8592	0.8499	0.8610	0.9230
TWINNED SVM-DT	0.8443	0.8559	0.8457	0.8559	0.9194

SVM (Support Vector Machine)

The SVM model exhibits a precision of 84.42%, indicating that out of all the patients classified as adherent, 84.42% were correctly identified. The recall is slightly higher at 85.51%, reflecting that 85.51% of all actual adherent patients were correctly classified. The F-Score, which balances precision and recall, stands at 0.8464, suggesting a robust performance in both precision and recall. The SVM classifier achieves an accuracy of 85.51%, which signifies that 85.51% of the overall classifications for both the adherent and non-adherent classes were correct. SVM also demonstrates a good level of ability to discriminate between adherent and non-adherent NCD patients, as reflected by the AUC of 0.8947.

TWINNED-SVM-RF

The newly built TWINNED-SVM-RF shows an improved level of performance, as reflected by 84.84% for the precision metric. This means that the TWINNED-SVM-RF classifier performs better than SVM in correctly identifying patients in the adherent class. The results for recall for the TWINNED-SVM-RF also reveal better classification of the actual adherent patients, as evidenced by the 85.92% statistics. The F-score at 0.8499 reflects an overall better balance of precision and recall than the SVM. In the MA classification task, the accuracy of the TWINNED-SVM-RF is 86.10%, compared to the SVM's 85.51%. This implies that the TWINNED-SVM-RF has a greater proportion of correct classifications overall in comparison to the SVM. Again, the AUC of the newly built TWINNED-SVM-RF significantly improved to 0.9230, indicating the classifier's superiority in distinguishing between adherent and non-adherent patients.

TWINNED-SVM-DT

The other classifier, TWINNED-SVM-DT, which was made by twinning SVM and DT, performed better, as evidenced by its accuracy of 84.43%, which is better than the SVM but slightly lower than the TWINNED-SVM-RF. The TWINNED-SVM-DT recall stands at 85.59%, which is higher than the SVM, indicating improved identification of the positive instances (in this case, adherent patients). The F-Score is 0.8457, closely mirroring the SVM but slightly behind TWINNED-SVM-RF. The accuracy is 85.59%, again better than the SVM but not as high as TWINNED-SVM-RF. The AUC for TWINNED-SVM-DT is 0.9194, significantly higher than the SVM, indicating a strong discriminatory capability.

3.3 Comparative Analysis of the Twinned ML Models versus SVM

When comparing the classical SVM to the newly developed hybrid models (TWINNED-SVM-RF and TWINNED-SVM-DT), the superior performance of the twinned models is evident. Specifically, TWINNED-SVM-RF achieves a precision of 84.84%, surpassing the SVM's precision of 84.42%, indicating better accuracy in identifying adherent patients. TWINNED-SVM-DT also shows slightly higher precision at 84.43%. In terms of recall, TWINNED-SVM-RF has a recall rate of 85.92%, outperforming the SVM's 85.51% and demonstrating improved identification of true adherent patients. TWINNED-SVM-DT also exceeds the SVM with a recall of 85.59%. For the F-Score, TWINNED-SVM-RF scores 0.8499, which is higher than the SVM's 0.8464, reflecting a better balance between precision and recall. Although TWINNED-SVM-DT's F-Score of 0.8457 is comparable to the SVM's, it still shows a marginal improvement.

In terms of accuracy, TWINNED-SVM-RF achieves 86.10%, exceeding the SVM's accuracy of 85.51%, indicating a higher proportion of correct classifications overall. The TWINNED-SVM-RF also demonstrates a significant improvement in the AUC score, with a value of 0.9230 compared to the SVM's 0.8947, highlighting its enhanced ability to distinguish between adherent and non-adherent patients. TWINNED-SVM-DT also shows a relatively higher AUC of 0.9194 compared to the SVM. Overall, these findings confirm that the twinned machine learning models exhibit superior performance over the classical SVM model across multiple metrics, with TWINNED-SVM-RF particularly demonstrating enhanced effectiveness in classifying medication adherence among patients with diabetes and hypertension.

3.4 Lime explanation for the best-twinned ML classifier (for instance– adherent instance)

The LIME (Local Interpretable Model-agnostic Explanations) analysis for the first instance in the dataset provides a detailed view of the model’s prediction. The LIME results are visually depicted in Figure 2, which displays a bar graph illustrating the two target classes for medication adherence: "Non-adherent" (blue) on the left and "Adherent" (orange) on the right. For this specific instance, where the patient is classified as "Adherent," the model predicts a slight preference for the "Adherent" class (51%) over the "Non-adherent" class (49%).

The feature-value table further contextualizes the prediction, showing the importance of various features in influencing the model's output for this instance. For the "Non-adherent" class (blue), significant features include ANNUALCLAIMAMOUNT (2.19), UNITS (1.61), COMPLICATIONDEVELOPMENT_2 (1.00), and HOSPITALCOVERTYPE_2 (1.00). High values for ANNUALCLAIMAMOUNT and UNITS suggest that larger claim amounts and specific unit counts are strong indicators of non-adherence.

Conversely, for the "Adherent" class (orange), important features are SCHEMETYPE_2 (1.00), COMORBIDITY_2 (1.00), and AGE (0.10). SCHEMETYPE_2 and COMORBIDITY_2 are notable positive contributors, indicating that patients with better medical schemes and comorbidities are more likely to adhere to their medication. Although AGE has a smaller weight (0.10), it still positively influences adherence, suggesting that older patients may be more compliant with their medication regimes.

Features WELLNESSPROGRAM_2 and SCHEMETYPE_3, despite being categorized under the "Adherent" class, have a value of 0.00, indicating minimal influence on the prediction for this specific instance. This analysis highlights the importance of understanding how individual features and their weights affect model predictions, offering valuable insights into patient adherence behavior.

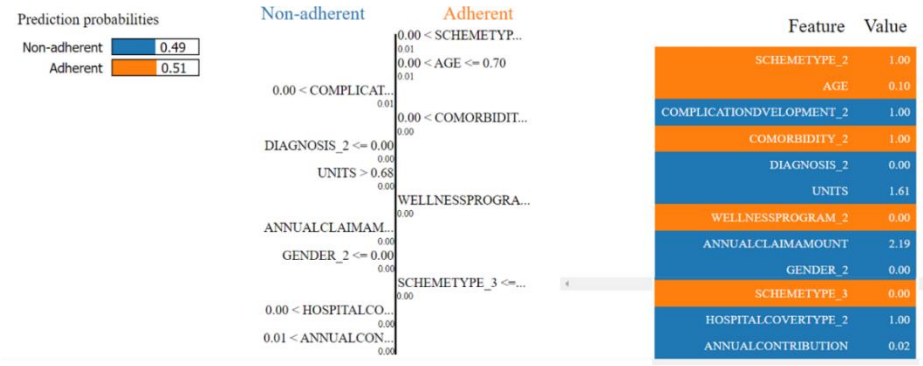


Fig. 2: Lime explanation for the best-twinned ML classifier: instance three of the dataset

3.5 Lime explanation for the best twinned ML classifier (for instance three - non-adherent instances)

The LIME (Local Interpretable Model-agnostic Explanations) results for the third instance in the dataset offer insights into the model's prediction for medication adherence. Figure 3 illustrates the two target classes: "Non-adherent" (blue) on the left and "Adherent" (orange) on the right. For this instance, where the patient is classified as "Non-adherent," the model predicts a slightly higher probability of "Non-adherent" (51%) compared to "Adherent" (49%).

Key features influencing the "Non-adherent" classification include WELLNESSPROGRAM_2, AGE, and HOSPITALCOVERTYPE_2. These features contribute positively, suggesting that patients not involved in wellness programs, older patients, and those with specific types of hospital coverage are more likely to be non-adherent.

For the "Adherent" class, influential features include SCHEMETYPE_2, DIAGNOSIS_2, COMORBIDITY_2, ANNUALCLAIMAMOUNT, GENDER_2 (Female), ANNUALCONTRIBUTION, SCHEMETYPE_3, and UNITS. SCHEMETYPE_2, DIAGNOSIS_2, COMORBIDITY_2, GENDER_2, and HOSPITALCOVERTYPE_2 have the highest positive weights (1.00), indicating that patients with specific medical schemes, diagnoses, comorbid conditions, and certain genders (female) are more likely to adhere to their medication. Financial factors such as ANNUALCLAIMAMOUNT (0.41) and ANNUALCONTRIBUTION (0.82) also positively influence adherence predictions, highlighting their role in medication adherence.

Conversely, UNITS and AGE have negative weights (-0.84 and -0.90, respectively), suggesting that higher values for these variables contribute to the likelihood of being classified as "Non-adherent." Overall, the LIME results underscore the complex interplay

of demographic, medical, and financial factors in predicting medication adherence, with distinct characteristics distinguishing adherent from non-adherent patients..

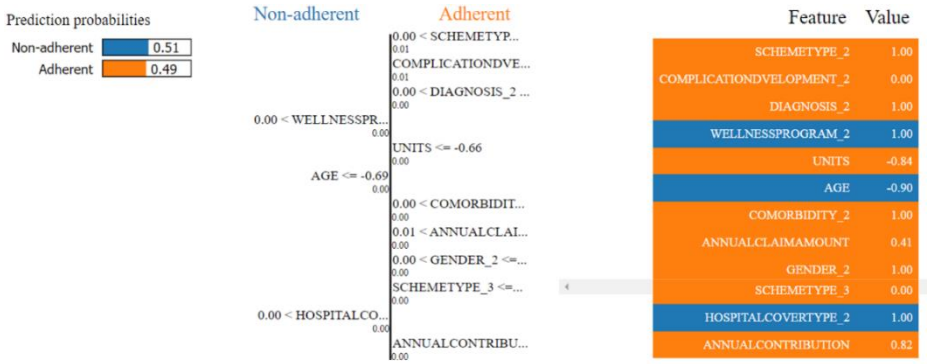


Fig. 3: Lime explanation for the best-twinned ML classifier: instance three

4. Discussion

The research demonstrated that developing a novel machine learning (ML) algorithm for medication adherence analytics by hybridizing classical Support Vector Machines (SVM) with tree-based ML algorithms significantly improves the classification of individuals with diabetes and hypertension based on their medication refill adherence. Specifically, the study combined SVM with Decision Trees (DT) and Random Forests (RF) to create the TWINNED-SVM-DT and TWINNED-SVM-RF models, respectively. The results indicate that both twinned models outperformed the classical SVM in key performance metrics, including precision, recall, F-Score, accuracy, and AUC. This evidence supports the effectiveness of hybridizing SVM with tree-based ML algorithms for medication adherence classification.

Notably, the TWINNED-SVM-RF achieved the highest overall performance, highlighting the significant contribution of the RF component. The RF's strength in handling imbalanced data and capturing complex patterns through ensemble learning contributed to its superior performance, consistent with findings by Mohindru, Mondal, and Banka (2021). The TWINNED-SVM-DT also improved model performance, though not to the same extent as the RF hybridization, reinforcing the idea that integrating various classical ML techniques can leverage their individual strengths, as discussed by Azevedo, Rocha, and Pereira (2024).

The model explainability technique used, LIME, was crucial in interpreting the predictions. It provided insights into the features influencing the model's decisions. For instance, better medical aid coverage (SCHEMETYPE_2) was positively associated with adherence predictions, as opposed to lower coverage schemes. Additionally, the presence of comorbidities among patients with non-communicable diseases (NCDs) also predicted higher adherence, suggesting that patients with better support systems and additional health conditions are more likely to adhere to their medication regimens.

5. Conclusion

The research findings conclusively validate the enhanced effectiveness of the twinned ML classifier, TWINNED-SVM-RF, in accurately classifying medication adherence (MA)

among patients with non-communicable diseases (NCDs), outperforming the classical SVM model. The superior performance of the TWINNED-SVM-RF, combined with the model explainability provided by LIME, offers valuable insights for healthcare providers. Understanding the features that significantly influence MA or non-adherence allows for the development of targeted interventions to address specific patient barriers, ultimately enhancing health outcomes. The study highlights the advantages of hybridized ML approaches and model explainability in creating effective healthcare solutions. The improved performance, robustness, and interpretability of the twinned models provide actionable insights for healthcare providers. However, potential biases within the dataset need to be addressed in future research.

Future studies should focus on mitigating these limitations and integrating these ML techniques into clinical practice. Recommendations include developing intelligent applications powered by the twinned model for at-risk patients and further exploring the incorporation of advanced ML models into healthcare decision-making processes to optimize patient outcomes.

Ethical Approval

N/A

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interest

The authors assert that they do not possess any identifiable conflicting financial interests or personal relationships that may have potentially influenced the findings presented in this research article.

Data and Code Availability

The used datasets can be found in a publicly accessible folder uploaded to Kaggle at https://www.kaggle.com/datasets/wellingtonkanyongo/cimas-medication-adherence-patient-and-claims-data

References

1. Abujaber, A. A., Nashwan, A. J., & Fadlalla, A. (2022). Enabling the adoption of machine learning in clinical decision support: A total interpretive structural modeling approach. Informatics in Medicine Unlocked, 33, 101090.
2. Ahmed, I., Ahmad, N. S., Ali, S., Ali, S., George, A., Danish, H. S., Uppal, E., Soo, J., Mobasheri, M. H., & King, D. (2018). Medication adherence apps: review and content analysis. JMIR MHealth and UHealth, 6(3), e6432.
3. Algabbani, F. M., & Algabbani, A. M. (2020). Treatment adherence among patients with hypertension: findings from a cross-sectional study. Clinical Hypertension, 26(1), 1–9.
4. Azevedo, B. F., Rocha, A. M. A. C., & Pereira, A. I. (2024). Hybrid approaches to optimization and machine learning methods: a systematic literature review. Machine Learning, 1–43.
5. Berwick, D. M., & Gaines, M. E. (2018). How HIPAA harms care, and how to stop it. Jama, 320(3), 229–230.

6. Chen, Y.-L., Nguyen, P.-A., Chien, C.-H., Hsu, M.-H., Liou, D.-M., & Yang, H.-C. (2024). Machine learning-based prediction of medication refill adherence among first-time insulin users with type 2 diabetes. Diabetes Research and Clinical Practice, 207, 111033.
7. Goecks, J., Jalili, V., Heiser, L. M., & Gray, J. W. (2020). How machine learning will transform biomedicine. Cell, 181(1), 92–101.
8. Hamrahian, S. M. (2020). Medication non-adherence: a major cause of resistant hypertension. Current Cardiology Reports, 22, 1–7.
9. Hezarjaribi, N., Fallahzadeh, R., & Ghasemzadeh, H. (2016). A machine learning approach for medication adherence monitoring using body-worn sensors. 2016 Design, Automation & Test in Europe Conference & Exhibition (DATE), 842–845.
10. Koesmahargyo, V., Abbas, A., Zhang, L., Guan, L., Feng, S., Yadav, V., & Galatzer-Levy, I. R. (2020). Accuracy of machine learning-based prediction of medication adherence in clinical research. Psychiatry Research, 294, 113558.
11. Mirzadeh, S. I., Arefeen, A., Ardo, J., Fallahzadeh, R., Minor, B., Lee, J.-A., Hildebrand, J. A., Cook, D., Ghasemzadeh, H., & Evangelista, L. S. (2022). Use of machine learning to predict medication adherence in individuals at risk for atherosclerotic cardiovascular disease. Smart Health, 26, 100328.
12. Mohindru, G., Mondal, K., & Banka, H. (2021). Different hybrid machine intelligence techniques for handling IoT-based imbalanced data. CAAI Transactions on Intelligence Technology, 6(4), 405–416.
13. Son, Y.-J., Kim, H.-G., Kim, E.-H., Choi, S., & Lee, S.-K. (2010). Application of support vector machine for prediction of medication adherence in heart failure patients. Healthcare Informatics Research, 16(4), 253–259.
14. Warren, D., Marashi, A., Siddiqui, A., Eijaz, A. A., Pradhan, P., Lim, D., Call, G., & Dras, M. (2022). Using machine learning to study the effect of medication adherence in Opioid Use Disorder. Plos One, 17(12), e0278988.
15. World Health Organisation. (2021). Hypertension. https://www.who.int/news-room/fact-sheets/detail/hypertension
16. World Health Organisation. (2022). Noncommunicable diseases. https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/ncd-mortality#:~:text=Noncommunicable diseases (NCDs) kill 41,- and middle-income countries.
17. Wu, X.-W., Yang, H.-B., Yuan, R., Long, E.-W., & Tong, R.-S. (2020). Predictive models of medication non-adherence risks of patients with T2D based on multiple machine learning algorithms. BMJ Open Diabetes Research and Care, 8(1), e001055.
18. Kanyongo, W., & Ezugwu, A. E. (2023). Machine learning approaches to medication adherence amongst NCD patients: A systematic literature review. Informatics in Medicine Unlocked, 38, 101210.
19. Kanyongo, W., & Ezugwu, A. E. (2023). Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives. Informatics in Medicine Unlocked, 38, 101232.

6.2 Chapter Summary

This chapter presents a study on enhancing machine learning model performance in classifying medication adherence (MA) by developing two hybrid models: TWINNED-SVM-RF and TWINNED-SVM-DT. The research aimed to improve the traditional Support Vector Machine (SVM) by integrating it with tree-based algorithms—Random Forests (RF) and Decision Trees (DT). Using a dataset of 8,141 records from patients with diabetes and hypertension, the study developed these hybrid models to enhance classification accuracy. The models were evaluated using accuracy and Area Under the Curve (AUC) metrics. The results showed that the TWINNED-SVM-RF model achieved the highest performance with an accuracy of 0.8610 and an AUC of 0.9230, outperforming the classical SVM model, which had an accuracy of 0.8551 and an AUC of 0.8947. The TWINNED-SVM-DT also performed well, with an accuracy of 0.8559 and an AUC of 0.9194. Additionally, the study applied LIME (Local Interpretable Model-Agnostic Explanations) to the TWINNED-SVM-RF model, enhancing its explainability by providing insights into its decision-making process. This research demonstrates that the TWINNED-SVM-RF model is a powerful tool for MA analytics, offering superior performance and valuable insights, making it highly applicable in healthcare and other fields requiring accurate and interpretable predictions. The next and last chapter of the study submit conclusions and recommendations based on the findings reported across the entire study.

CHAPTER 7:

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

7.1 Introduction

This chapter furnishes a summary of the entire thesis and includes brief explications of the findings associated with each chapter. The summary provides a comprehensive overview of the study's outcomes and contributions, encompassing results relating to all the research objectives. The key findings of each chapter are accentuated to showcase their alignment with the research objectives as well as the accomplishment of the research objectives. Furthermore, this chapter provides definitive conclusions and recommendations that stem from the main findings of the study. The conclusions consolidate study insights, highlighting the potential implications of the findings. The chapter also addresses the theoretical implications, elucidating how the study enhances a more comprehensive comprehension of ML applications in healthcare. The chapter's practical contributions demonstrate how predictive models can assist medical professionals in identifying patients at risk of non-adherence and implementing targeted interventions. To improve healthcare delivery, policy considerations are discussed and recommendations are made regarding how integrating ML techniques might be encouraged broadly. Finally, this study proposes specific areas for future research to build upon the current study's findings. This final chapter of the thesis ensures that readers understand the impact of the study and potential advancements in MA analytics.

7.2 General Summary

With a particular focus on diabetic and hypertensive patients in Zimbabwe, the study aimed to leverage ML techniques to build models for classifying MA among patients with NCDs. The study was predicated on the premise that adhering to prescribed medication is a crucial element in managing NCDs. However, non-adherence continues to be an epidemic, resulting in adverse health effects and higher healthcare expenses. Understanding and accurately classifying MA allows healthcare providers to intervene more efficiently and improve patient outcomes. The research motivation arises from the urgent necessity to

tackle the high rates of medication non-adherence among patients with NCDs. Existing approaches for measuring adherence are frequently insufficient, presenting a substantial opportunity to improve existing systems by employing advanced ML approaches. By making precise predictions about which patients are prone to non-adherence, healthcare systems can more effectively focus interventions, thus lessening the impact of NCDs. The study constitutes an expansive literature review on ML methods for MA analytics. It also aimed to ascertain and assess important predictors of MA using feature selection techniques. The study also conducted rigorous pre-processing and DW in generating the datasets necessary for the ML modelling tasks. Finally, the study developed and validated ML models to classify MA among diabetic and hypertensive patients in Zimbabwe.

The research methodology comprises several key steps. Firstly, a systematic literature review was conducted to identify any gaps or opportunities in the existing research on ML applications in MA analytics. Secondly, feature selection was performed to identify the most relevant predictors of MA and assess their importance. Next, data wrangling and generation were carried out, which involves identifying and acquiring data from various sources, importing it, inspecting it, preprocessing it, enriching it, integrating it, profiling it, transforming it, and reducing its dimensionality. Finally, model development and validation were conducted by applying various ML algorithms to develop the classification models for MA analytics, which were then evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC. The study was based on data obtained from a Zimbabwean medical aid firm, primarily analysing patients with diabetes and hypertension throughout 2022. The dataset consists of individual patient data and medical claims data, which have been preprocessed and enriched to ensure the model development's reliability and robustness. The subsequent sub-sections provide a summary of each chapter, including its findings and conclusions.

7.2.1 Summary of findings related to Review on ML Approaches to Medication Adherence Analytics

The comprehensive literature review conducted in Chapter Two (2) reveals a range of ML approaches used for predicting MA. These algorithms include classical ML algorithms, ensemble, and deep learning techniques. Algorithms such as LR, DT, RF, support vector machines, and deep learning techniques have shown different levels of performance. These algorithms have shown different levels of efficacy in predicting medication compliance among individuals with NCDs. Each algorithm has distinct advantages and drawbacks, while certain research revealed the higher efficacy of particular algorithms over others. Again, the systematic literature review revealed paucity in past studies, such as the requirement for more rigorous, widely applicable models and the incorporation of various data types. Typical data sources used in prior research included EHRs, claims data, and administrative databases found in health institutions. Moreover, the review reported the commonly used ML model evaluation metrics, such as accuracy, precision, recall, and the area under the receiver operating characteristic (ROC) curve which have been applied in previous research work. It is always imperative to establish standardised evaluation procedures to enable comparisons between studies and guarantee the replicability of findings. By and large, the review chapter highlights the capacity of ML to revolutionise MA analytics. However, it emphasises the need for more extensive and unified research endeavours to fully exploit this potential.

7.2.2 Summary of findings related to feature selection and importance of predictors of medication adherence from ML research perspectives

Chapter three's (3) systematic literature review focused on feature selection and collating the important predictors in MA analytics from the perspective of M research. The review found that the selection of important features greatly influences the efficacy of prediction models. Typically, indicators of MA include demographic variables (age, gender), socio-economic factors (income, education level), health status indicators (comorbidities, severity of disease), medication-related factors (type of medication, dosage, frequency), and behavioral factors (past adherence history, lifestyle habits). Advanced techniques for feature selection, such as RFE, principal component analysis, and regularization methods

(including Lasso and Ridge), have been used to improve model performance by lowering dimensionality and enhancing interpretability. The review highlighted that the inclusion of a diverse set of predictors can greatly improve the accuracy and resilience of adherence models. Nevertheless, the selection of predictors differs significantly among studies, indicating the intricate and diverse nature of MA behaviour.

The feature selection review chapter emphasised the importance of domain knowledge in feature selection, as the inclusion of clinically significant predictors can improve the model's comprehensibility and applicability. The chapter reports the need for accurately identifying and selecting the appropriate features for creating ML models in MA analytics. The review underscored the importance of employing advanced feature selection techniques, such as the RF feature importance mechanism, to incorporate the most relevant variables for classifying MA or non-adherence. The necessity for further investigation into the comparative significance of various predictors and their interrelationships was emphasised. The results indicate that employing a more sophisticated and evidence-based method for choosing features significantly enhances the precision and reliability of MA classification models. The review reveals a diverse range of possible predictors and the essential role of feature selection in enhancing model performance. These insights are crucial for directing the advancement of more efficient and interpretable ML models in MA analytics.

7.2.3 Summary of results related to data wrangling and generation for ML models in medication adherence analytics

The chapter describes the development of a high-quality state-of-the-art dataset using medical and pharmaceutical claims as well as patient-level data obtained from the TPS of a health insurance company in Zimbabwe. The raw sub-datasets, one for diabetics and one for hypertensive patients were obtained and divided according to the International Classification of Diseases 10th Revision Code (ICD-10). These sub-datasets served as a starting point for generating the data. Given that the adherence metrics assess the proportion of patients who have prescription claims for the same medication within the same therapeutic class, the generated dataset solely encompassed data from January 1 to

December 31, 2022. This dataset includes pharmacy refill records, patient-level data, and medical claims data, providing a comprehensive view of patient adherence behaviour and related information.

The chapter generated and published a dataset containing patient-level and medical claims data to analyse MA. The dataset was created using a pragmatic approach that resembled an agile methodology for dataset development. This involved employing nine fundamental data-wrangling building blocks and tasks, progressing in incremental stages. The practical approach to generating the dataset involved nine components: data source identification and acquisition, data importation and inspection, data pre-processing, data enrichment, data integration, data profiling, data transformation, dimensionality reduction, and technical validation. Data wrangling tasks were applied iteratively and incrementally across the whole data wrangling process, without being limited to a single step. Following the meticulous process of organising and cleaning the data, the generated dataset was used to build a machine-learning model to validate the quality of the dataset. The model exhibited a commendable level of performance, as indicated by an accuracy rate of 81%. The study asserts that meticulous data wrangling and generation procedures are important for building robust and reliable machine-learning models for MA analytics. The chapter emphasises the significance of meticulous DW and the incorporation of multiple data sources to ensure that generated datasets that are both all-inclusive and rich, serving as the foundation of optimal ML modelling.

7.2.4 Summary of results related to the development and validation of ML models for medication adherence analytics

Chapter five (5) aimed to build and validate ML models for classifying MA in diabetic and hypertensive patients. The study involved the development, evaluation, and comparison of multiple models, namely k-NN, SVM, RF, DNN, DT, LR, and Naïve Bayes. The performance of each model was thoroughly evaluated using a variety of evaluation metrics. The empirical results identified several significant predictors of MA, including the number of medical supplies dispensed, the annual number of claims, the patient's age, enrollment in a wellness programme in addition to medical aid coverage, the annual contribution

towards medical aid coverage, comorbidity status, type of diagnoses, type of hospital coverage, the occurrence of diabetes or hypertension-related complications, gender, and type of medical aid scheme. These predictors were ascertained using the RF feature importance scoring mechanism. The extensive range of features was subsequently utilised to develop ML models for classifying MA.

In Chapter Five (5), it was established that the ML models that were created successfully achieved a high level of accuracy in predicting MA. The RF model, in particular, demonstrated the highest levels of precision, recall, F-Score, accuracy, and AUC, which indicates its reliability and robustness in classifying MA. The RF model's great efficacy can be attributed to its capacity to handle large datasets with high dimensionality and its effectiveness in capturing intricate interactions among features without overfitting. This model has the potential to be an intelligent tool for improving MA, helping to identify patients' adherence levels. These findings contribute to the resolution of discrepancies in medication refills and adherence rates among patients with NCDs. The robustness of this algorithmic approach makes it highly suitable for practical use in healthcare environments, where accurate predictions are of utmost importance. Other models such as DNN, SVM, and DT also demonstrated strong performance, making them ideal alternatives for MA analytics. These results highlight how crucial it is to take into account a variety of evaluation metrics and model attributes to make sure the ML model selected fits the particular requirements and constraints involved in MA analytics application.

7.3 Implications for Practice

The contributions of this study to practice are diverse and critical.

The development and validation of ML models to classify MA in patients with NCDs, such as diabetes and hypertension proffer practical solutions for healthcare professionals and institutions. By using models such as RF, DNN, and SVM, clinical processes may effectively detect patients who are likely to not adhere to their treatment plans. This enables healthcare providers to take proactive precautions and create personalised care plans for these patients.

Moreover, the application of advanced feature selection approaches enables healthcare institutions and providers to ascertain significant predictors of MA, thereby delivering practical and actionable insights. By comprehending the patient-level features, treatment variables, and socio-economic variables that impact MA behaviour, health practitioners can customise interventions and support mechanisms to successfully address unique patient requirements.

Furthermore, the study aids in enhancing the efficiency and effectiveness of healthcare delivery systems. The utilisiation of ML to classify MA enables healthcare providers to improve resource allocation, focus patient outreach efforts, and deploy remedies in areas where they are most needed. This, in turn, improves overall patient outcomes and lowers healthcare expenses.

In addition, integrating the ML-based classification models for MA analytics into clinical decision support systems has the potential to simplify decision-making procedures and enable the delivery of care based on solid evidence. Healthcare providers can use predictive analytics to use up-to-date data to make well-informed decisions regarding drug administration, treatment plans, and patient interventions.

In another practical implication, the ML models serve as a foundation for the creation of intelligent MA and intervention apps to promote patient MA among chronic disease patients. The apps can be configured to incorporate personalised intervention plans, such as tailored reminders.

In summary, this research has practical value because it effectively applies theoretical knowledge to develop strategies that increase patient care, optimise healthcare delivery, and eventually lead to better health outcomes for patients with chronic diseases.

7.4 Implications for Policy

The policy implications emerging from this study are useful for changing healthcare systems and plans that aim to improve patients' ability to take their medications as prescribed and their overall health, especially when it comes to NCDs such as diabetes and hypertension.

Policymakers can use the findings emerging from the study to guide the formulation of focused interventions and efforts designed to enhance MA rates among people with chronic diseases. Policymakers can develop evidence-based policies that target the underlying causes of non-adherence and offer customised support to vulnerable patient populations by comprehending the features that impact adherence behaviour and the efficacy of various predictive models.

Furthermore, policymakers should use this research's valuable insights to actively promote the incorporation of predictive analytics into healthcare delivery systems and reimbursement models. Policymakers can motivate healthcare institutions to invest in data science tools and integrate them into regular clinical practice by showcasing the benefits of ML in enhancing MA and patient outcomes.

In summary, this research highlights the significance of using data-driven insights to impact policy decisions and bring about positive modifications in healthcare delivery and patient outcomes. Through the adoption of predictive analytics and evidence-based interventions, policymakers can strive to achieve higher rates of adherence, decrease healthcare expenses, and enhance the quality of care for people with chronic diseases.

7.5 Theoretical Implications

Key and diverse contributions to knowledge by using ML techniques to advance the field of MA in patients with NCDs were submitted under this section.

The systematic reviews conducted in the study offer an in-depth review of the current body of literature on the use of ML in MA analytics, specifically focusing on feature selection. The study provides an in-depth analysis of the current state of the subject of MA analytics

from the perspective of ML by synthesising and evaluating literature from multiple studies, emphasising prevailing trends, research designs, and research gaps that require further investigation.

The research enhances the development of predictive, classification, clustering, and association models for MA by incorporating insights from the conducted systematic reviews. An exhaustive assessment of several ML algorithms and feature selection methods establishes a strong basis for developing more highly accurate and robust models, ultimately enhancing patient management and healthcare outcomes.

This study enhances awareness of the predictors that impact MA in patients with NCDs such as diabetes and hypertension. The study also enhances understanding of adherence dynamics by examining the key predictors and assessing intricate connections among patient-level features, medical history, and adherence behaviour.

The study demonstrates the capabilities of ML algorithms in classifying MA, emphasizing the strengths and weaknesses of various models, such as classical ML models, deep learning models, and ensemble learners. This comparative investigation enhances the current understanding of the suitability and efficacy of various ML methods in healthcare data analytics.

The study used advanced feature selection approaches, such as the RF feature importance scoring mechanism, to discover significant predictors of MA. The predictors include variables such as the number of medical supplies dispensed, the total number of claims made per year, the patient's age, involvement in a wellness program, the type of medical aid coverage, and the presence of comorbidities. This full feature set analysis enhances current knowledge by identifying essential features that impact adherence. By carefully selecting and identifying important features, the research not only enhances understanding of adherence behaviour but also provides valuable information on predictive modelling in healthcare.

The study demonstrates the significance and practicality of merging patient-level and medical claims data to improve the accuracy of classifications and predictions. This method emphasises the importance of integrating extensive data to create efficient ML models for healthcare systems applications.

Moreover, the results of the study highlight the capability of ML models to enhance patient care techniques and improve treatment approaches for chronic illnesses. This study adds to the existing theoretical discourse on how to use data-driven insights to assist medical doctors in making informed decisions by employing sophisticated analytical techniques. This results in a more profound comprehension of how patient attributes, treatment strategies, and adherence results mutually influence one another in complex ways.

In conclusion, these theoretical contributions create opportunities for more efficient treatments, customised healthcare solutions, and enhanced patient outcomes in the management of chronic diseases.

7.6 Future Research Directions

Future studies in the field of MA among patients with NCDs offer various interesting opportunities for further investigation and progress, especially by utilising ML approaches.

i. Future research practitioners can perform longitudinal research to analyse MA patterns over time and determine the temporal variables that impact adherence behavior. Machine learning algorithms could be applied to analyse the longitudinal data, revealing patterns and predicting future adherence habits based on historical data.
ii. Future research should incorporate longitudinal data to analyse the dynamic nature of medication adherence and improve predictive accuracy by tracking changes in patient behavior and health status over time.
iii. Future research examining the effectiveness of personalised adherence therapies that are tailored to the unique traits, preferences, and needs of each patient can be conducted. Future research can utilise emerging ML methods, such as transfer

learning, reinforcement learning, and personalised recommendation systems, to create adaptive solutions that alter according to patient feedback and outcomes.

iv. In the future, researchers can pursue implementation science research to analyse the scalability, practicality, and long-term viability of adherence strategies in actual healthcare environments. The research can employ ML techniques to evaluate the variables that impact the effective adoption and execution of adherence interventions, as well as to pinpoint solutions for overcoming implementation obstacles.

v. There is a need to analyse the impact of healthcare policies, reimbursement models, and financial incentives on the results of MA. In that future study, ML models to assess the cost-effectiveness of adherence strategies, offering data-driven insights into healthcare expenditure, resource allocation, and population health outcomes can be employed.

References

Abdul Basith Khan, M., Hashim, M. J., King, J. K., Govender, R. D., Mustafa, H., & Al Kaabi, J. (2020). Epidemiology of type 2 diabetes—global burden of disease and forecasted trends. *Journal of Epidemiology and Global Health*, *10*(1), 107–111.

Ahmed, I., Ahmad, N. S., Ali, S., Ali, S., George, A., Danish, H. S., Uppal, E., Soo, J., Mobasheri, M. H., & King, D. (2018). Medication adherence apps: review and content analysis. *JMIR MHealth and UHealth*, *6*(3), e6432.

Ai, L., Muggleton, S. H., Hocquette, C., Gromowski, M., & Schmid, U. (2021). Beneficial and harmful explanatory machine learning. *Machine Learning*, *110*, 695–721.

Al-Noumani, H., Al Omari, O., & Al-Naamani, Z. (2023). Role of health literacy, social support, patient-physician relationship, and health-related quality of life in predicting medication adherence in cardiovascular diseases in Oman. *Patient Preference and Adherence*, 643–652.

Al-Ganmi, A. H. A., Alotaibi, A., Gholizadeh, L., & Perry, L. (2020). Medication adherence and predictive factors in patients with cardiovascular disease: A cross-sectional study. *Nursing & Health Sciences*, *22*(2), 454–463.

Algabbani, F. M., & Algabbani, A. M. (2020). Treatment adherence among patients with hypertension: findings from a cross-sectional study. *Clinical Hypertension*, *26*(1), 1–9.

Almutiri, R., Alhabeeb, S., Alhumud, S., & Khan, R. U. (2022). A Survey of Machine Learning for Big Data Processing. *Journal on Big Data*, *4*(2).

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., & Badreldin, H. A. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*, *23*(1), 689.

American Heart Association. (2021). *A tough pill to swallow medication adherence and cardiovascular disease*.

Aziz, F., Malek, S., Ali, A. M., Wong, M. S., Mosleh, M., & Milow, P. (2020). Determining hypertensive patients' beliefs towards medication and associations with medication adherence using machine learning methods. *PeerJ*, *8*, e8286.

Babel, A., Taneja, R., Mondello Malvestiti, F., Monaco, A., & Donde, S. (2021). Artificial Intelligence Solutions to Increase Medication Adherence in Patients With Non-communicable Diseases. *Frontiers in Digital Health*, 69.

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *Jama*, *319*(13), 1317–1318.

Benziger, C. P., Roth, G. A., & Moran, A. E. (2016). The global burden of disease study and the preventable burden of NCD. *Global Heart*, *11*(4), 393–397.

Berger, M. L., & Doban, V. (2014). Big data, advanced analytics and the future of

comparative effectiveness research. *Journal of Comparative Effectiveness Research*, *3*(2), 167–176.

Berwick, D. M., & Gaines, M. E. (2018). How HIPAA harms care, and how to stop it. *Jama*, *320*(3), 229–230.

Bohlmann, A., Mostafa, J., & Kumar, M. (2021). Machine learning and medication adherence: scoping review. *JMIRx Med*, *2*(4), e26993.

Bourdès, V., Ferrières, J., Amar, J., Amelineau, E., Bonnevay, S., Berlion, M., & Danchin, N. (2011). Prediction of persistence of combined evidence-based cardiovascular medications in patients with acute coronary syndrome after hospital discharge using neural networks. *Medical & Biological Engineering & Computing*, *49*, 947–955.

Brown, M. T., & Bussell, J. K. (2011). Medication adherence: WHO cares? *Mayo Clinic Proceedings*, *86*(4), 304–314.

Chen, Y.-L., Nguyen, P.-A., Chien, C.-H., Hsu, M.-H., Liou, D.-M., & Yang, H.-C. (2024). Machine learning-based prediction of medication refill adherence among first-time insulin users with type 2 diabetes. *Diabetes Research and Clinical Practice*, *207*, 111033.

Christopoulou, S. C. (2024). Machine Learning Models and Technologies for Evidence-Based Telehealth and Smart Care: A Review. *BioMedInformatics*, *4*(1), 754–779.

Comer, S., Cunningham, C., Fishman, M. J., Gordon, F. A., Kampman, F. K., Langleben, D., Nordstrom, B., Oslin, D., Woody, G., & Wright, T. (2015). National practice guideline for the use of medications in the treatment of addiction involving opioid use. *Am Soc Addicit Med*, *66*, 39–42.

Deo, R. C. (2015). Machine learning in medicine. *Circulation*, *132*(20), 1920–1930.

Fernandez-Lazaro, C. I., Adams, D. P., Fernandez-Lazaro, D., Garcia-González, J. M., Caballero-Garcia, A., & Miron-Canelo, J. A. (2019). Medication adherence and barriers among low-income, uninsured patients with multiple chronic conditions. *Research in Social and Administrative Pharmacy*, *15*(6), 744–753.

Franchi, C., Ludergnani, M., Merlino, L., Nobili, A., Fortino, I., Leoni, O., & Ardoino, I. (2022). Multiple medication adherence and related outcomes in community-dwelling older people on chronic polypharmacy: a retrospective cohort study on administrative claims data. *International Journal of Environmental Research and Public Health*, *19*(9), 5692.

Franklin, J. M., Shrank, W. H., Lii, J., Krumme, A. K., Matlin, O. S., Brennan, T. A., & Choudhry, N. K. (2016). Observing versus predicting: initial patterns of filling predict long-term adherence more accurately than high-dimensional modeling techniques. *Health Services Research*, *51*(1), 220–239.

Galozy, A., & Nowaczyk, S. (2020). Prediction and pattern analysis of medication refill adherence through electronic health records and dispensation data. *Journal of Biomedical Informatics*, *112*, 100075.

Galozy, A., Nowaczyk, S., Sant'Anna, A., Ohlsson, M., & Lingman, M. (2020). Pitfalls of medication adherence approximation through EHR and pharmacy records: definitions, data and computation. *International Journal of Medical Informatics*, *136*, 104092.

Gardezi, S. K. M., Aitken, W. W., & Jilani, M. H. (2023). The Impact of Non-Adherence to Antihypertensive Drug Therapy. *Healthcare*, *11*(22), 2979.

Gottlieb, A., Yatsco, A., Bakos-Block, C., Langabeer, J. R., & Champagne-Langabeer, T. (2022). Machine Learning for Predicting Risk of Early Dropout in a Recovery Program for Opioid Use Disorder. *Healthcare*, *10*(2), 223.

Gu, Y., Zalkikar, A., Liu, M., Kelly, L., Hall, A., Daly, K., & Ward, T. (2021). Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data. *Scientific Reports*, *11*(1), 18961.

Guerra, R. (2024). Enhancing risk management in hospitals: leveraging artificial intelligence for improved outcomes. *Italian Journal of Medicine*, *18*(2).

Guo, M., Wang, Y., Yang, Q., Li, R., Zhao, Y., Li, C., Zhu, M., Cui, Y., Jiang, X., & Sheng, S. (2023). Normal Workflow and Key Strategies for Data Cleaning Toward Real-World Data. *Interactive Journal of Medical Research*, *12*(1), e44310.

Hamrahian, S. M. (2020). Medication non-adherence: a major cause of resistant hypertension. *Current Cardiology Reports*, *22*, 1–7.

Hasan, M. M., Young, G. J., Shi, J., Mohite, P., Young, L. D., & Weiner, S. G. (2021). A machine learning based two-stage clinical decision support system for predicting patients' discontinuation from opioid use disorder treatment: retrospective observational study. *BMC Medical Informatics and Decision Making*, *21*(1), 1–21.

Hu, F., Warren, J., & Exeter, D. J. (2020). Predicting lipid-lowering medication persistence after the first cardiovascular disease hospitalization. *Methods of Information in Medicine*, *59*(02/03), 61–74.

Huber, C. A., Meyer, M. R., Steffel, J., Blozik, E., Reich, O., & Rosemann, T. (2019). Post-myocardial infarction (MI) care: medication adherence for secondary prevention after MI in a large real-world population. *Clinical Therapeutics*, *41*(1), 107–117.

Jacobson, A. M. (2004). Impact of improved glycemic control on quality of life in patients with diabetes. *Endocrine Practice*, *10*(6), 502–508.

Kanyongo, W., & Ezugwu, A. E. (2023). Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives. *Informatics in Medicine Unlocked*, 101232. https://doi.org/https://doi.org/10.1016/j.imu.2023.101232

Kassahun, T., Gesesew, H., Mwanri, L., & Eshetie, T. (2016). Diabetes related knowledge, self-care behaviours and adherence to medications among diabetic patients in Southwest Ethiopia: a cross-sectional survey. *BMC Endocrine Disorders*, *16*, 1–11.

Kleinsinger, F. (2018). The unmet challenge of medication nonadherence. *The Permanente*

Journal, *22*.

Lam, W. Y., & Fresco, P. (2015). Medication adherence measures: an overview. *BioMed Research International*, *2015*.

Lappan, S. N., Brown, A. W., & Hendricks, P. S. (2020). Dropout rates of in-person psychosocial substance use disorder treatments: a systematic review and meta-analysis. *Addiction*, *115*(2), 201–217.

Lee, S. K., Kang, B.-Y., Kim, H.-G., & Son, Y.-J. (2013). Predictors of medication adherence in elderly patients with chronic diseases using support vector machine models. *Healthcare Informatics Research*, *19*(1), 33–41.

Li, M., Lu, X., Yang, H., Yuan, R., Yang, Y., Tong, R., & Wu, X. (2022). Development and assessment of novel machine learning models to predict medication non-adherence risks in type 2 diabetics. *Frontiers in Public Health*, *10*, 1000622.

Li, Y., Jasani, F., Su, D., Zhang, D., Shi, L., Yi, S. S., & Pagán, J. A. (2019). Decoding nonadherence to hypertensive medication in New York City: a population segmentation approach. *Journal of Primary Care & Community Health*, *10*, 2150132719829311.

Lin, L.-K., Sun, Y., Heng, B. H., Chew, D. E. K., & Chong, P.-N. (2017). Medication adherence and glycemic control among newly diagnosed diabetes patients. *BMJ Open Diabetes Research and Care*, *5*(1), e000429.

Malo, S., Aguilar-Palacio, I., Feja, C., Lallana, M. J., Rabanaque, M. J., Armesto, J., & Menditto, E. (2017). Different approaches to the assessment of adherence and persistence with cardiovascular-disease preventive medications. *Current Medical Research and Opinion*, *33*(7), 1329–1336.

Masiero, M., Spada, G. E., Sanchini, V., Munzone, E., Pietrobon, R., Teixeira, L., Valencia, M., Machiavelli, A., Fragale, E., & Pezzolato, M. (2023). A Machine Learning Model to Predict Patients' Adherence Behavior and a Decision Support System for Patients With Metastatic Breast Cancer: Protocol for a Randomized Controlled Trial. *JMIR Research Protocols*, *12*(1), e48852.

Mills, K. T., Stefanescu, A., & He, J. (2020). The global epidemiology of hypertension. *Nature Reviews Nephrology*, *16*(4), 223–237.

Náfrádi, L., Nakamoto, K., & Schulz, P. J. (2017). Is patient empowerment the key to promote adherence? A systematic review of the relationship between self-efficacy, health locus of control and medication adherence. *PloS One*, *12*(10), e0186458.

Odunyemi, A., Rahman, T., & Alam, K. (2023). Economic burden of non-communicable diseases on households in Nigeria: evidence from the Nigeria living standard survey 2018-19. *BMC Public Health*, *23*(1), 1563.

Ogurtsova, K., da Rocha Fernandes, J. D., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N. H., Cavan, D., Shaw, J. E., & Makaroff, L. E. (2017). IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Research and Clinical Practice*, *128*, 40–50.

Organization, W. H. (2018). Global health estimates 2016: deaths by cause, age, sex, by country and by region, 2000–2016. *Geneva: World Health Organization*, 1242–1247.

Owusu-Adjei, M., Hayfron-Acquah, J. Ben, Twum, F., & Abdul-Salaam, G. (2023). Machine Learning Modeling of Disease Treatment Default: A Comparative Analysis of Classification Models. *Advances in Public Health*, *2023*.

Pradeep, Y. (2023). Prescription-Based Medication Management System with Automated Adherence Notification. *International Journal of Research Publication and Reviews*, *4*(12), 3912–3916.

Pulice, E., & Coustasse, A. (2024). AI-Driven Solutions Promote: Medication Adherence High nonadherence in the US leads to avoidable mortality. *Pharmacy Times ·*, 42–44.

Sahu, M., Gupta, R., Ambasta, R. K., & Kumar, P. (2022). Artificial intelligence and machine learning in precision medicine: A paradigm shift in big data analysis. *Progress in Molecular Biology and Translational Science*, *190*(1), 57–100.

Son, Y.-J., Kim, H.-G., Kim, E.-H., Choi, S., & Lee, S.-K. (2010). Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthcare Informatics Research*, *16*(4), 253–259.

Stafford, I. S., Gosink, M. M., Mossotto, E., Ennis, S., & Hauben, M. (2022). A systematic review of artificial intelligence and machine learning applications to inflammatory bowel disease, with practical guidelines for interpretation. *Inflammatory Bowel Diseases*, *28*(10), 1573–1583.

Stolar, M. (2010). Glycemic control and complications in type 2 diabetes mellitus. *The American Journal of Medicine*, *123*(3), S3–S11.

Swed, S., Alibrahim, H., Bohsas, H., Hafez, W., Surma, S., Rais, M. A., Abuelsaoud, H. M., Elshazly, R. M., Shoib, S., & Sawaf, B. (2023). Assessment the awareness toward hypertension and diabetes mellitus: Syrian cross sectional study. *BMC Public Health*, *23*(1), 789.

Thanh, P. T., & Duong, P. B. (2022). The economic burden of non-communicable diseases on households and their coping mechanisms: Evidence from rural Vietnam. *World Development*, *151*, 105758.

Unni, E., & Bae, S. (2022). Exploring a new theoretical model to explain the behavior of medication adherence. *Pharmacy*, *10*(2), 43.

von Wyl, V., Ulyte, A., Wei, W., Radovanovic, D., Grübner, O., Brüngger, B., Bähler, C., Blozik, E., Dressel, H., & Schwenkglenks, M. (2020). Going beyond the mean: economic benefits of myocardial infarction secondary prevention. *BMC Health Services Research*, *20*, 1–13.

Walsh, C. A., Cahir, C., Tecklenborg, S., Byrne, C., Culbertson, M. A., & Bennett, K. E. (2019). The association between medication non-adherence and adverse health outcomes in ageing populations: a systematic review and meta-analysis. *British Journal of Clinical Pharmacology*, *85*(11), 2464–2478.

Wang, L., Fan, R., Zhang, C., Hong, L., Zhang, T., Chen, Y., Liu, K., Wang, Z., & Zhong, J. (2020). Applying machine learning models to predict medication nonadherence in Crohn's disease maintenance therapy. *Patient Preference and Adherence*, 917–926.

Warren, D., Marashi, A., Siddiqui, A., Eijaz, A. A., Pradhan, P., Lim, D., Call, G., & Dras, M. (2022). Using machine learning to study the effect of medication adherence in Opioid Use Disorder. *Plos One*, *17*(12), e0278988.

World Health Organisation. (2021). *Hypertension*. https://www.who.int/news-room/fact-sheets/detail/hypertension

World Health Organisation. (2022). *Noncommunicable diseases*. https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/ncd-mortality#:~:text=Noncommunicable diseases (NCDs) kill 41,- and middle-income countries.

Xu, Y., Zheng, X., Li, Y., Ye, X., Cheng, H., Wang, H., & Lyu, J. (2023). Exploring patient medication adherence and data mining methods in clinical big data: A contemporary review. *Journal of Evidence-Based Medicine*, *16*(3), 342–375.

Yang, J., Li, Y., Liu, Q., Li, L., Feng, A., Wang, T., Zheng, S., Xu, A., & Lyu, J. (2020). Brief introduction of medical database and data mining technology in big data era. *Journal of Evidence-Based Medicine*, *13*(1), 57–69.

Yang, X., Yan, J., Lai, S., Shen, C., & Duan, N. (2023). What affects the direct economic burden of non-communicable diseases on middle-aged and older adult people in Shaanxi Province? *Frontiers in Public Health*, *11*, 1219199.

Zakeri, M., Sansgiry, S. S., & Abughosh, S. M. (2022). Application of machine learning in predicting medication adherence of patients with cardiovascular diseases: a systematic review of the literature. *Journal of Medical Artificial Intelligence*, *5*.

Zullig, L. L., Jazowski, S. A., Wang, T. Y., Hellkamp, A., Wojdyla, D., Thomas, L., Egbuonu-Davis, L., Beal, A., & Bosworth, H. B. (2019). Novel application of approaches to predicting medication adherence using medical claims data. *Health Services Research*, *54*(6), 1255–1262.

Appendices

Appendix A: Letter of Approval for Data Collection

connect@cimas.co.zw
cimas.co.zw
facebook.com/cimasmedicalaid
East Block, Borrowdale Office Park
08677 400 500 | 08080001-3

08 June 2023

Wellington Kanyongo -PhD. Student
School of Mathematics, Statistics and Computer Science
University of KwaZulu-Natal
Pietermaritzburg Campus, Pietermaritzburg, 3201
South Africa

Dear Wellington Kanyongo and Dr. Tsitsi Moyo

Re: Granted request for all data related to medication adherence amongst hypertension & diabetes patients for research.

This letter is in direct response to the letter you wrote to us on 20 May 2023 requesting for dataset related to medication adherence amongst hypertension & diabetes patients which you need to use for your academic research and publications with regards to medication adherence preparation and analytics using machine learning techniques. We are pleased to inform you that your dataset request has been officially approved and now you go ahead to use for your academic research and publications.

The dataset that we have approved for your use include have medical claims data, patient-level data and internal processing data with at least 44 variables including age, gender, and any additional demographics of the members or patients, race, scheme type, service state (count), claim code, diagnosis, units, threshold showing level of medication adherence based medical claims count, amount claimed, total paid, tariff, shortfall, race, subscription default variable, complications development variable, among other relevant variables that can be used to predict medication adherence level. We understand, you can also make use of data (such as subscription amounts or contributions, different types of our medical aid packages and the associated hospital cover types among other variables) available in the public domain which can be accessed via our company brochures and websites.

We have examined your request and found we are proving you the data in an anonymized form that the identities of the patients or your members are kept confidential. As such the data and your research finding will not be tied to any single individual but will be reported in general terms. We wish you the best in your research work as you contribute to research around data preparation for medication adherence and publications related to the dataset thereof. We kindly ask you to share your results for the benefit of our day-day business operations.

Yours faithfully

[redacted]

Thando Kembo
Chief Operating Officer

Board Members:
Mrs. E. Fundira (Chairman), D. G. Bain (Vice Chairman), V. L. Ndlovu (Chief Executive Officer), N. Frost, Ms. R. Magorimbo, B. Mswaka, Mr C. Chiketa, Mrs. M. L. Ndachena, V. Nkomo
Cimas Health Group (Private) Limited

connect@cimas.co.zw
cimas.co.zw
facebook.com/cimasmedicalaid
East Block, Borrowdale Office Park
08677 400 500 | 08080001-3

08 June 2023

Wellington Kanyongo -PhD. Student

School of Mathematics, Statistics and Computer Science University of KwaZulu-Natal

Pietermaritzburg Campus, Pietermaritzburg, 3201 South Africa

Dear Wellington Kanyongo and Dr. Tsitsi Moyo

Re: Granted request for all data related to medication adherence amongst hypertension & diabetes patients for research.

This letter is in direct response to the letter you wrote to us on 20 May 2023 requesting for dataset related to medication adherence amongst hypertension & diabetes patients which you need to use for your academic research and publications with regards to medication adherence preparation and analytics using machine learning techniques. We are pleased to inform you that your dataset request has been officially approved and now you go ahead to use for your academic research and publications.

The dataset that we have approved for your use include have medical claims data, patient-level data and internal processing data with at least 44 variables including age, gender, and any additional demographics of the members or patients, race, scheme type, service state (count), claim code, diagnosis, units, threshold showing level of medication adherence based medical claims count, amount claimed, total paid, tariff, shortfall, race, subscription default variable, complications development variable, among other relevant variables that can be used to predict medication adherence level. We understand, you can also make use of data (such as subscription amounts or contributions, different types of our medical aid packages and the associated hospital cover types among other variables) available in the public domain which can be accessed via our company brochures and websites.

We have examined your request and found we are proving you the data in an anonymized form that the identities of the patients or your members are kept confidential. As such the data and your research finding will not be tied to any single individual but will be reported in general terms. We wish you the best in your research work as you contribute to research around data preparation for medication adherence and publications related to the dataset thereof. We kindly ask you to share your results for the benefit of our day-day business operations.

Yours faithfully

[redacted]

Thando Kembo

Chief Operating Officer

CIMAS HEALTH GROUP (PRIVATE) LIMITED
08 JUN 2023
EAST BLOCK
BORROWDALE OFFICE PARK
BORROWDALE. HARARE

Board Members:
Mrs. E. Fundira (Chairman), D. G. Bain (Vice Chairman), V. L. Ndlovu (Chief Executive Officer), N. Frost, Ms. R. Magorimbo, B. Mawaka, Mr C. Chikata, Mrs. M. L. Ndlachena, V. Nkomo
Cimas Health Group (Private) Limited