Deep Learning Framework for Speech Emotion Classification

by

Samson Adebisi Akinpelu
222068579

Submitted in fulfilment of the academic requirements for the degree of

Doctor of Philosophy in Computer Science in the

School of Mathematics, Statistics, and Computer Science

University of KwaZulu-Natal

Durban South Africa, 2024

Supervisor: Prof. Serestina Viriri

Declaration of Authorship

I, Samson Adebisi AKINPELU, declare that this thesis titled, "Deep Learning Frameworks for Speech Emotion Classification" and the work presented in it are my own. I declare that:

1. The research reported in this thesis, except where otherwise indicated or acknowledged, is my original work;
2. This thesis has not been submitted in full or in part for any degree or examination to 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, their writing has been placed 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 being detailed in the thesis and in the References sections.

Candidate: Samson Adebisi Akinpelu

Signature: 

Date: 13/06/2024

As the candidate's supervisor, I approve the submission of this thesis for examination.

Supervisor: Prof. Serestina Viriri

Signature: 

Date: 13/06/2024

Abstract

A robust deep learning-based approach for the recognition and classification of speech emotion is proposed in this research work. Emotion recognition and classification occupy a conspicuous position in human-computer interaction (HCI) and by extension, determine the reasons and justification for human action. Emotion plays a critical role in decision-making as well. Distinguishing among various emotions (angry, sad, happy, neutral, disgust, fear, and surprise) that exist from speech signals has however been a long-term challenge. There have been some limitations associated with existing deep learning techniques as a result of the complexity of features from human speech (sequential data) which consists of insufficient label datasets, Noise and Environmental Factors, Cross-cultural and Linguistic Differences, Speakers' Variability and Temporal Dynamics. There is also a heavy reliance on huge parameter tunning, especially for millions of parameters before the model can learn the expected emotional features necessary for classification emotion, which often results in computational complexity, over-fitting, and poor generalization. This thesis presents an innovative deep learning framework-based approach for the recognition and classification of speech emotions. The deep learning techniques currently in use for speech-emotion classification are exhaustively and analytically reviewed in this thesis.

This research models various approaches and architectures based on deep learning to build a framework that is dependable and efficient for classifying emotions from speech signals. This research proposes a deep transfer learning model that addresses the shortcomings of inadequate training datasets for the classification of speech emotions. The research also models advanced deep transfer learning in conjunction with a feature selection algorithm to obtain more accurate results regarding the classification of speech emotion. Speech emotion classification is further enhanced by combining the regularized feature selection (RFS) techniques and attention-based networks for the classification of speech emotion with a significant improvement in the emotion recognition results. The problem of misclassification of emotion is alleviated through the selection of salient features that are relevant to emotion classification from speech signals. By combining regularized feature selection with attention-based mechanisms, the model can better understand emotional complexities and outperform conventional ML model emotion detection algorithms.

The proposed approach is very resilient to background noise and cultural differences, which makes it suitable for real-world applications. Having investigated the reasons behind the enormous computing resources required for many deep learning-based methods, the research proposed a lightweight deep learning approach that can be deployed on low-memory devices for speech emotion classification. A redesigned VGGNet with an overall model size of 7.94MB is utilized, combined with the best-performing classifier (Random Forest). Extensive experiments and comparisons with other deep learning models (DenseNet, MobileNet, InceptionNet, and ResNet) over three publicly available speech emotion datasets show that the proposed lightweight model improves the performance of emotion classification with

minimal parameter size. The research further devises a new method that minimizes computational complexity using a vision transformer (ViT) network for speech emotion classification. The ViT model's capabilities allow the mel-spectrogram input to be fed into the model, allowing for the capturing of spatial dependencies and high-level features from speech signals that are suitable indicators of emotional states. Finally, the research proposes a novel transformer model that is based on shift-window for efficient classification of speech emotion on bi-lingual datasets. Because this method promotes feature reuse, it needs fewer parameters and works well with smaller datasets. The proposed model was evaluated using over 3000 speech emotion samples from the publicly available TESS, EMODB, EMOVO, and bilingual TESS-EMOVO datasets. The results showed 98.0%, 98.7%, and 97.0% accuracy, F1-Score, and precision, respectively, across the 7 classes of emotion.

Declaration

The work described in the thesis was carried out in the School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal from February 2022 to May 2024. This dissertation was completed under the supervision of Professor Serestina Viriri.

This study represents original work by the author and has not been submitted in any form for any degree or diploma to any other tertiary institution. Where use was made of the work of others it has been duly acknowledged in the text.

List of Publications

1. **Samson Akinpelu and Serestina Viriri**, "Lightweight Deep Learning Framework for Speech Emotion Recognition", *IEEE Access*, vol. 11, pp. 77086-77098, **(2023)**, DOI: https://doi.org/10.1109/ACCESS.2023.3297269.

2. **Samson Akinpelu, Serestina Viriri**, "Speech Emotion Classification Using Attention-Based Network and Regularized Feature Selection", *Scientific Reports*, vol. 13(11990), **(2023)**, DOI: https://doi.org/10.1038/s41598-023-38868-2.

3. **Samson Akinpelu and Serestina Viriri**, "Robust Feature Selection-Based Speech Emotion Classification Using Deep Transfer Learning", *Applied Sciences*, vol. 12(16), 8265, **(2022)**. DOI: https://doi.org/10.3390/app12168265.

4. **Samson Akinpelu, Serestina Viriri**, "An Enhanced Speech Emotion Recognition using Vision Transformer", *Scientific Reports*, vol. 14(13126), **(2024)**, DOI: https://doi.org/10.1038/s41598-024-63776-4.

5. **Samson Akinpelu, Serestina Viriri**, "A Robust Deep Transfer Learning Model for Accurate Speech Emotion Classification", *Advances in Visual Computing*, LNCS Springer, vol. 13599, pp. 419-430, **(2022)**. DOI: https://doi.org/10.1007/978 -3-031-20716-7-33.

6. **Samson Akinpelu, Serestina Viriri**, "Speech Emotion Classification: A Survey of the State-of-the-Art", *Pan-African Artificial Intelligence and Smart Systems*, LNICST Springer, vol. 459, pp. 379–394, **(2023)**, DOI: https://doi.org/10.1007/978-3-031-25271-6-24.

7. **Samson Akinpelu and Serestina Viriri**, "Deep Learning Framework for Speech Emotion Classification: A Survey of State-of-the-Arts", *Journal of Big Data*. (resubmitted with minor corrections).

8. **Samson Akinpelu, Serestina Viriri**, "An Improved Bilingual Speech Emotion Recognition Using Shift Window Transformer", (under review).

Acknowledgements

All gory to the most precious God, the author of life who has bestowed upon me the rare grace and right to undertake this research work. I thank Him for the wisdom, knowledge, inspiration, and insight that He bestowed upon me to see to the completion of this work. I equally appreciate my Heavenly father His provision and preservation throughout my journey in this research work. Indeed, I have obtained His mercy in the course of this academic career, all adoration is due to Him alone.

My profound gratitude also goes to my able supervisor, Professor Serestina Viriri, who has been my mentor and guidance all through the period of this research. His undoubted confidence in me has boosted my morale throughout this research. Your kindness will forever remain indelible in my heart. I also thank God for my late parents (MR. & MRS ISAAC AKINPELU) who laid a solid foundation for my education, even amid uncertainty. I appreciate all my Spiritual mentors whom I cannot be mentioned one after the other.

I equally Appreciate the Federal University of Oye-Ekiti (FUOYE), Nigeria for allowing me to pursue this career to a meaningful conclusion at one of the best Universities in Africa (University of KwaZulu-Natal).

I thank all my colleagues at the Department of Computer Science, Federal University Oye-Ekiti, Dr. Adewole (HOD), Assoc. Prof. Ogunleye (Immediate past HOD), DR. Folorunsho Olaiya (International mentor), Dr. Fagbola, Dr. Fagbuagun, Dr. (Mrs.) Daramola, Mr. Bature Hassan, Mr. Odufuwa, and Mrs. Adelekun (Admin.), to mention but a few. I also appreciate Pastor Akinkunmi (Deputy Registrar & Head Personnel Affairs, FOUYE), Dr. Abimbade (Pos Doc Fellow, Nelson Mandella University), Dr. Adediran (immediate past Director, LANDMARK University), Pastor Jerry, Pastor Ilesanmi and Pastor Obajaja C. K. for their support. I also thank Dr. Adekanmi Adegun (UK) and Dr. Ekundayo (UP, SA) for their invaluable contribution and support in the course of this research. My brethren and friends, Br. Victor Uzor, Br. Fave, Mrs. Ijeoma (Computer Science, UNIZULU), Dr. Lizzy, Dr. Vincent, Dr. Folawewo David and Dr. Reginald (Edgewood Campus) thank you very much.

I am also grateful to my adorable wife, Mrs. Oluwadimimu Patience Akinpelu, for her support and her Godly stand all-through the journey of this PhD programme. You are mostly appreciated, my First Lady. My children, Goodness and Godliness Akinpelu, you are wonderful for your endurance. I also thank all my family members, especially, Mr. and Mrs J. O Akinpelu for their immense contribution towards achieving the desired success in my career.

Dedication

This thesis is dedicated to Almighty God.

Contents

List of Tables

List of Figures

List of Abbreviations

1. ADAM- Adaptive Momentum
2. AESDD - Acted Emotional Speech Dynamic Database
3. ANN- Artificial Neural Network
4. AUC- Area Under Curve
5. BiLSTM- Bi-directional Long Short Term Memory
6. CNNs- Convolutional Neural Networks
7. CREMA-D- Crowd-sourced Emotional Multimodal Actors Dataset
8. CRNN- Convolutional Recurrent Neural Network
9. DBN- Deep Belief Network
10. DCNNs- Deep Convolutional Neural Networks
11. DCT - Discrete Cosine Transform
12. DESS- Danish Emotional Speech
13. DTL- Deep Transfer Learning
14. EEG – Electroencephalogram signal
15. EMD - Empirical Mode Decomposition
16. Emo-DB - German emotional database
17. FCBF- Fast Correlation-Based Feature Selection
18. FCN- Fully Convolutional Network
19. FFT- Fast Fourier Transform
20. FKNN- Firey with K-Nearest Neighbour
21. FLOOPS- Floating-Point Operations per Seconds
22. FN- False Negative
23. FP- False Positive
24. FPR- False Positive Rate
25. GANs- Generative Adversarial Networks
26. GELU- Gaussian Error Linear Unit

27. GMM- Gaussian Mixture Model
28. GPU- Graphical Processing Unit
29. HCI- Human-Computer Interaction
30. HMM- Hidden Markov Model
31. IEMOCAP- Interactive Emotional Dyadic Motion Capture Database
32. LOO- Leave One Out
33. LPCC- Linear Prediction Cepstral Coefficients
34. LSTM- Long Short-Term Memory
35. MELD- Multimodal Emotion Lines Dataset
36. MFCC- Mel-frequency Cepstral coefficient
37. MLP- Multilayer Perceptron
38. MLT- Multi-Learning Trick
39. MSHA- Multi-Head Self-Attention
40. NBA- Naive Bayesian Algorithm
41. NCA- Neighbourhood Component Analysis
42. NLP- Natural Language Processing
43. RAVDESS- Ryerson Audio-Visual Database of Emotional Speech and Song
44. RELU- Rectified Linear Units
45. RESNET- Residual Network
46. RF- Random Forest Algorithm
47. RNCA- Regularized Neighbourhood Component Analysis
48. RNN- Recurrent Neural Networks
49. ROC- Receiver Operating Characteristic
50. SAIL- Speech Analysis and Interpretation Laboratory
51. SAVEE- Surrey Audio-Visual Expressed Emotion
52. SEC- Speech Emotion Classification
53. SER – Speech Emotion Recognition

54. SNR - Signal-to-Noise Ratio
55. STFT- Short-Time Fourier Transform
56. SVM- Support Vector Machine
57. TESS - Toronto Emotional Speech Set
58. TKEO- Teager-Kaiser Energy Operator
59. TP- True Positive
60. TPR- True Positive Rate
61. UAR- Unweighted Average Recall
62. VACNN- Visual Attention Convolutional Neural Network
63. VGGNet- Visual Graphics Group Network
64. ViT-Vision Transformer

Chapter 1

Introduction

1.1 Research Background

Emotions are states of mind resulting from neurophysiological alterations that are connected in diverse ways to ideas, sentiments, actions, and a level of happiness or unhappiness. Dynamic emotion-cognition [23] interactions are the most common emotion experiences. These interactions might involve fleeting responses or persistent personality traits that develop over time. The development of consciousness and the functioning of every mental process depend heavily on emotions. The remarkable quality that sets humans apart is their ability to adjust conversations according to the emotional states of both the speaker and the listener. Humans are the most advanced species in terms of civilization because they can express their emotions in a variety of ways. These expressions can be verbal, facial, gestural, or based on other physiological mechanisms. Individual relationships and involvement, however, are best maintained through human speech-based communication. Both in direct and indirect communication, the vast paralinguistic [30] content of human speech can disclose an individual's emotional condition.

Human speech is a potent means of communication, capable of communicating both knowledge and emotions. With the increasing integration of HCI and AI systems in our lives, there is a rising demand for them to proficiently comprehend and react to human emotions[49]. Speech-emotion classification, the process of discerning the emotional state based on spoken language, presents a potential pathway for developing human-computer interactions that are more genuine and compassionate. This thesis introduces a sophisticated deep learning system designed specifically for the classification of speech-emotion. It aims to tackle the difficulties and possibilities that arise in this intricate endeavour. AI systems that can effectively identify human emotions have the potential to enhance personalized, supportive, and engaging experiences in different fields.

Several technologies involving Human-Computer Interaction (HCI) have been made possible by advances in artificial intelligence (AI) [46]. Important discoveries about how emotions affect opinions and decisions throughout human history have emerged as a result of HCI advancements [2]. As HCI is the front end of AI that millions of users interact with, it is critical to create and enhance HCI methodologies. Several of the HCI techniques currently in use include touch, motions with the hands, voice, and facial expressions [49]. Out of all the techniques, voice-activated intelligent gadgets are becoming increasingly common in a variety of uses. Significantly, speech emotion classification occupies a key position among other speech-related tasks. It is the recognition of emotional traits through a computing device, as against the handcrafted approach of human recognition of emotion, that is most often subjective. Most research in psychology and science has shown that speech emotion classification is not only fundamental to human existence and activity but can also be studied with the help of modernist scientific computing devices [31]. For a computing device to correctly interpret commands in a speech-based system, it must fully understand the human speaker's speech perception which involves speech signal processing.

The speech signal carries both linguistic and paralinguistic features that serve various purposes in the development of speech-based systems. The emotional cue from speech resides in the paralinguistic constituent of human speech utterance as shown in Figure 1.1. Paul Ekan categorizes the basic human emotions into six sub-divisions which are anger, disgust, fear, happiness, sadness, and surprise [16]. He argues that each of these basic emotions has distinct qualities that are linked to it, enabling different degrees of expression. Anger is a strong emotional state characterized by an intense, uneasy, and uncooperative reaction to an event that is seen as a threat or provocation. Disgust is an emotional display of rejection or aversion to something distasteful or perhaps infectious. Fear is an extremely unsettling feeling that arises when one perceives danger. It is an unpleasant feeling that typically results in psychological disturbances that could lead to behavioural reactions like launching an offensive attack or running away. Humans experience fear in response to current stimuli or in anticipation of potential threats that they see as endangering their safety. Happiness is a good emotion that can range from satisfaction to extreme excitement. It is an emotion that can be spurred on by optimistic experiences, but occasionally they might happen for no apparent reason.

Sadness is a painful emotional state that is described by emotions such as disappointment, helplessness, despair, and grief. Severe depression has been linked to mental disease and depressive disorders. The neuroscience of depression has been extensively studied, and the results indicate that sadness is linked to a rise in bilateral activities around the lateral cerebellum [6]. Surprise is a fleeting mental and emotional state that follows an unexpected event. Surprise can be either pleasant or negative, or it can have any valence. All these emotional expressions have both positive and negative impacts, and they can be detected from auditory speech through an efficient deep learning model.

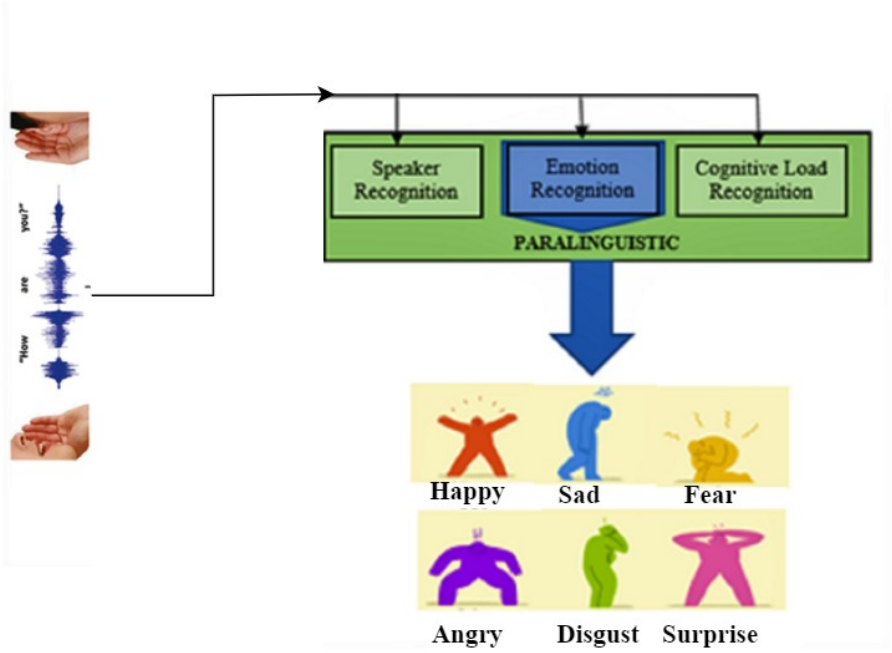


Figure 1.1: Basic Emotion Classification from Speech

From a larger perspective, emotion classification from speech signals has been applied to audio synthesis, healthcare, the automotive industry (drowsy driving), robotics, multimedia, education (e-learning), criminal justice, job recruitment, customer service management, smart home connections (IoT), among other areas. The primary challenge in extracting emotional features from speech signals is that not all speech content contains the relevant emotional expression, making the correct classification of emotional elements a challenging task. Besides, the classical approach to the classification of speech emotion using conventional classifiers suffers from misclassification [42].

1.2 Motivation

Emotions are essential to human communication because they shape our relationships and how we perceive one another. There are various uses for classifying the emotion conveyed in speech perception in this modern era. The study of interactive applications involving people and computers is known as HCI. For an HCI application to be effective, the computer system must comprehend more than mere words [34]. However, many commonly used Internet of Things (IoT) applications, such as Google Home, Amazon Alexa, and Mycroft, rely on speech-based inputs. Speech has a crucial role in Internet of Things applications. Understanding emotion from the speech signal is vital in both cases of these speech interactions, which might be mono- or bi-directional. Additional uses for speech emotion classification include

humanoids, crime agency investigation, and fraud detection systems. Speech emotion is primarily recognized and classified from speech signals through approaches such as perceptual inspection, articulation and pronunciation, vocal quality, empathy, and contextual information [27]. Its application in healthcare (memory disorder), adaptive e-learning and the automobile industry(self-driving cars) cannot be overemphasized [29]. Classifying speech emotion using these human approaches is bewildered with problems and considered to be time-consuming, laborious, inaccurate, and subjected to errors due to the complex features inherent in speech signals. Speech sound is characterized by multi-sizes and noise presence. Accurate classification of speech emotion is particularly important to enhance human-computer interaction in the AI era. These challenges have motivated the development of a computer-based system for speech emotion classification to aid its real-world applications.

Several computer-based systems have been developed for the classification of speech emotion. These systems' primary goal is to extract the dominant emotional features from speech sounds. These systems analyze speech signals using a variety of computer techniques. Pitch, energy, prosody, Mel Frequency Cepstral Coefficient (MFCC), zero crossing rate (ZCR), formants or resonance frequency, articulation rate, and other general features of speech signals were the cornerstones of some of the early systems used for speech emotion classification [50]. These features are selected and detected for analysis using these algorithms. Additional basic algorithms include the following: Prosodic Features-Based algorithm with Gaussian Mixture Models (GMM), Rule-based Emotional Speech Synthesizer, Kernel regression, Likelihood Bayes' (MLB) [19][3], Pitch and Formant Analysis using autocorrelation and Linear Predictive Coding, [38][13] etc. The complex nature and variability of speech signals have undoubtedly hampered these algorithms, which are based on techniques such as speech pre-processing and acoustics feature extraction [19].

Speech emotion classification systems have recently taken a more comprehensive approach, incorporating tasks like speech pre-processing, segmentation (framing), feature extraction, and classification into the analysis and recognition of emotion from speech signals[37]. Preprocessing, the first step, is typically carried out to lessen prominent frames in the speech dataset's audio sample and enhance the quality of the speech signals. The preprocessing methods correct artifacts, normalize amplitudes, and eliminate noise from raw speech signals to improve their quality. Speech segmentation is located next to it. The continuous speech signal is segmented into brief, overlapping frames. This procedure helps to capture temporal dynamics. Common feature extraction approaches convert every speech frame into a collection of relevant features that describe the distinctive features of the signal. Artificial intelligence approaches have been utilized in classification recently [34], wherein emotion is categorized into several groups.

Machine learning techniques are typically employed by SEC systems to learn and extract features from speech signals, both in segmentation and classification tasks. In the recent past, these tasks have been completed using conventional machine learning methods such as support vector machines (SVMs), logistic regression, nearest k-neighbors, decision trees, and neural networks [32] [43] [8]. These methods indicate the relationship between the observed features and a certain emotion using various mathematical equations and functions. These methods continue to rely on handcraft features that were taken from more conventional feature extraction methods which are extremely susceptible to noise and variations in the speech signal. This constitutes a major limitation for the SEC system when analyzing speech signals.

The high levels of noise sensitivity and the limitations of traditional machine learning algorithms in handling imprecise and uncertain data can then be attributed to the difficulties encountered by the different emotion classification systems. Additionally, these methods typically [21]lack substantial supervision, which frequently results in the loss of specific information during training, and they perform poorly in spatial modeling. As a result, they fail to analyze the complex features that are characterized by the presence of noise and artifacts, temporal dynamics, context dependency, cross-cultural and linguistic variances, speakers' variability in emotional expression, and environmental impacts [25][47]. These are the factors that lead to the SEC systems' poor performance, which occasionally causes them to misclassify and yield false positives in emotion classification. For example, a psychoanalyst may make a mistake in providing the appropriate treatment for a patient with depression or mental illness if they interpret a neutral emotional expression as indicating disgust. When a crucial choice, such as a forensic or criminal inquiry, is made based on a recognized emotion, it may also have negative consequences.

Therefore, a pertinent question that never goes away is: Is it possible to enhance the accuracy and reliability of the emotion classification systems to gain the confidence of experts in the neuroscience field? Also, can we develop a system that efficiently and accurately classifies emotion, keeping in mind that emotion affects crucial decisions and that correctly classifying it would enhance human-computer interaction, given the significance of accurate emotion classification in our scenarios?

An accurate, portable, and efficient system for classifying emotions is greatly needed to obtain suitable responses to these questions. The framework that is required to minimize the time and computational cost involved in handling complex features of speech signals must be data-driven, parameter- and memory-efficient [4]. The classification technique will lower the misclassification rate of speech utterances with complex features and classify speech utterances into six basic emotion classes: angry, sad, disgust, happy, neutral, and fear. The primary goal of this research is to develop a speech emotion classification system that is more accurate and specific while minimizing false negatives.

The development of a reliable and efficient speech-emotion classification system through the use of cutting-edge machine learning techniques like deep learning approaches is what then motivates this research. In pattern recognition and sequence operation, deep learning techniques have reached state-of-the-art performance [22][1]. Their ability to learn and extract deep and hierarchical features from complex speech spectrogram data is what makes them effective. For example, a popular deep learning technique known as Deep Convolutional Neural Networks (DCNNs) with attention mechanisms and Long-Short-Term-Memory (LSTM) can process a wide range of tasks involving fine-grained objects[15]. Compared to handcrated techniques, this robust method is far more effective at extracting salient features from audio signals and sequence data [24][45][44]. Additionally, deep learning approaches offer fault tolerance, adaptability, and optimal performance with high sensitivity and specificity for accurate speech signal analysis. They frequently offer solutions to the problems with the conventional SEC model. Compared to traditional SEC methods, deep learning techniques are more effective, particularly when it comes to learning highly discriminative features and long-range dependencies that aid in the classification of emotions. However, the following factors continue to limit their performance:

1. Training deep learning techniques on sparsely labeled data may result in poor generalization and over-fitting.

2. To function well, the majority of deep learning techniques rely heavily on tuning millions of parameters, which increases their memory and computing requirements.

3. When faced with complex situations involving background noise, a wide range of verbal expressions, and cultural variances, they do badly.

4. Since speech utterances are always recorded using various devices with varying sample rates, deep learning approaches also require a strong framework to learn emotional cues from a variety of datasets and languages.

The motivation of this research is to model optimum solutions to the shortfalls of deep learning systems so that they can classify speech emotions accurately and with the fewest possible false positives and false negatives. It explores how the robustness of the emotion classifier improves when deep transfer learning, feature selection, and a natural language processing model are incorporated into the classification task.

1.3 Problem Statement

The research work explores and models a deep learning framework for improved Speech Emotion Classification (SEC) systems using speech signals. When it comes to addressing the issues that conventional SEC systems have when processing raw speech signals, deep learning frameworks are highly promising. A great deal of research in the recent time has attested to this fact. Therefore, an ultimate solution will result from a deep learning framework that can effectively classify emotion from human speech. However, accurately classifying speech emotions and developing effective, efficient feature extraction and selection approaches for speech spectrogram images containing complex features like language, culture, and background noise present challenges. Furthermore, optimizing the selection of significant features that aid in the recognition of emotions from speech signals is also necessary. Additionally, the deep learning framework's computational and storage complexity must be optimized. Another issue is that training with little datasets necessitates a data-driven system.

Owing to the potential benefit, a robust deep learning framework for the analysis of speech signals will be produced by combining feature selection approaches with classification processes. By investigating the potential of Deep Convolutional Neural Network (DCNN) based architectures for emotion recognition and mitigating the effects of currently existing deep learning system limitations, such as the need for millions of parameter tuning and a massive training dataset, overfitting and weak generalization while training with insufficient labeled training data, and common challenges in processing images generated from speech signal with varied features, this study proposes a new deep learning framework for automated speech emotion classification.

The following are the major problematic areas that are taken into consideration in this thesis:

1. Deep learning techniques are mostly used to hierarchically learn the features corresponding to the emotional cues from speech signals by using huge labelled datasets [24]. In speech emotion classification, where there is a scarcity of labelled datasets, using limited labelled data can lead to overfitting and poor generalization. Generally, they require large training datasets to create effective models. Due to security concerns, human speech datasets are more difficult to obtain than image datasets. The low sensitivity of the model to emotional information can also result from training deep learning techniques with a smaller amount of data. This thesis addresses this issue using a deep transfer learning approach. An optimal deep transfer learning approach can manage the differences in emotional utterance representation and class labels between the source and target data, which could lead to a biased feature

extraction network that performs worse on the target data in terms of generalization [45].

2. Moreover, some speech signals exhibit abnormalities in their spectral properties, leading to over-blotted features. This poses a challenge for deep learning techniques that aim to accurately analyze speech and classify emotions [22][1]. For instance, because of their striking similarity in energy and pitch characteristics, it may be incorrect to identify the emotions of surprise, happiness, and anger in spoken utterances [15]. When recognizing emotion, the systems may become confused by these striking similarities. Furthermore, it is impossible to overemphasize the likelihood of many characteristics arising from overlapping emotions in a multiclass problem such as speech emotion classification. However, using unique feature extraction approaches, the traditional and certain early SEC systems sequence models attempted to address this issue. However, due to the heavy reliance on handcrafted features with low model generalization ability and limited capacity to create high-level representations, their effectiveness has been limited [15]. When recognizing emotion, the systems may become confused. This research aims to provide an effective Deep Learning framework that can handle these problems by introducing an extended deep transfer learning with feature selection approaches that minimize misclassification of emotion.

3. This research aims to develop a deep learning framework for accurate and efficient speech emotion classification by addressing the challenges posed by artifacts and noise [17][12][14] in spectrogram images. The framework will incorporate effective speech-processing techniques and capture long-range dependencies to improve classification accuracy and enable early detection of emotional states. Additionally, feelings that lead to depression can result in chronic mental illness or premature death. Human well-being will be enhanced by early and accurate emotion classification from speech that is not limited by place, time, or obstruction, unlike facial expression identification. With the help of this research, a strong and effective framework for quickly identifying emotions from speech signals—such as anger, sadness, happiness, fear, neutrality, and disgust—will be developed.

4. Despite the recent progress in deep learning for speech emotion Classification, models for cross-cultural adaption still require significant improvement. Speakers' diverse cultural backgrounds contribute to the diversity of speech. Speaking utterances are associated with other languages in addition to the cultural background. This will overcome the shortcomings of existing systems that find it difficult to account for language differences when working with sequence data. Examining and creating deep learning methods that allow models to generalize in a variety of languages [11] and cultural contexts, without concentrating on local features alone is needful. This thesis addresses this issue using a deep learning approach that learns global emotional features from speech signals using a large language transformer [44] model.

5. Deep learning approaches mainly depend on the appropriate tuning of millions of parameters, which frequently results in overfitting, subpar generalization, and excessive computational resource usage [20]. Additionally, this process raises the memory and processing resource requirements for an effective system. A time-based emotion classification system must build a strong system that can quickly and efficiently handle computer resources, even though the widespread availability of GPUs has been able to counteract this effect. Computational resources should be managed efficiently to promote system portability on low-memory devices for convenient access. For effective performance speech emotion categorization, deep learning architecture still has an opportunity for several advancements and best practices.

6. Lastly, Though many novel feature extractions have been developed, however, their performance is limited without corresponding features feature selection approach. Besides, not all speech utterances carry relevant emotional features and extraction of local features by deep network layers alone for accurate emotion classification is not sufficient. Therefore, the incorporation of feature selection techniques with an attention mechanism for the selection of emotionally relevant features is proposed in this research on deep learning frameworks.

1.4 Aims and Objectives

The main aim is to model a fully automated deep learning framework for speech emotion classification.

Objectives:

1. To perform a comprehensive and analytical review of the current deep learning approaches for speech emotion classification.

2. To model a Deep Learning technique and dimensionality reduction for speech emotion classification using Deep Transfer Learning.

3. To investigate the performance of deep learning methods based on feature selection algorithms and classifiers.

4. To model an efficient lightweight deep learning framework using various deep neural network architectures.

5. To improve the classification method with an efficient transformer model and its performance on the speech emotion dataset.

6. To model a fully automatic approach using an attention-based network and regularized feature selection for efficient classification of speech emotion.

1.5 Research Approach

This research investigates and develops a novel deep learning approach that utilizes 3 main stages; speech analysis, feature extractions, and classification stages for accurate classification of various categories of emotion from speech signals. This novel approach achieves accurate and efficient classification of speech emotion into various classes. This approach tackles myriad challenges previously highlighted that are confronting the existing deep-learning framework for the SEC system from speech signals. In the approach, an auditory speech utterance, which is the most natural way of human communication and expression of personality traits, was first subjected to standard audio processing of segmentation or framing, filtering, windowing, and silence removal for the extraction of speech features. Two prominent features (mel-frequency cepstral coefficients and mel-spectrogram) that are most suitable for the deep learning method have been utilized in this research work. To extract the mel-frequency cepstral coefficients (MFCCs) [33], the audio signal is first framed into short, overlapping audio frames that last 25–40 ms to calculate the MFCCs. This yields a total of 16x32 = 512 audio samples per frame at a sampling rate of 16 kHz and a frame length of 32 ms.

Furthermore, the overlapping sections contain 16x16 = 256 audio samples if the frame step size (hop length) is 16ms. Next, for every audio frame, the power spectrum's periodogram estimate is computed. Thereafter, a mel-filterbank with a set of 20–40 triangular overlapping filters is utilized for mapping the powers of the spectrum to the mel-scale which results in the mel-spectrogram image. When applied to the power spectrum, these filters—whose spacing is determined by the mel-scale—provide the filter bank with energy. The logarithm function is applied to the filter bank energy once they have been obtained. This is necessary since loudness is not perceived by humans on a linear scale. The features are compressed by the log operation to make them more akin to human speech. To obtain the mel-frequency cepstral coefficients, the log mel-filter bank energy must be subjected to a discrete cosine transform (DCT).

The speech feature extraction output is sent into the various deep learning models proposed before the classification stage, which is made up of the best-performing classifier algorithms. We have therefore developed an automated framework that incorporates speech signal analysis, feature extraction, and selection for emotion classification.

1.6 Thesis Contributions

The main contribution of this research is to develop an automated framework for speech emotion classification. This research contributes to the research area of HCI in the following areas:

1. A deep transfer learning classifier with dimensionality reduction for overblotted features on speech signals with the capability to perform efficiently on a limited training dataset.

2. An extension of deep transfer learning network with a novel feature selection algorithm for complex features inherent in speech utterances for emotion classification.

3. A novel transformer method derived from a natural language processing model has been introduced.

4. A memory and parameter efficient VGGNet framework specially designed for the efficient classification of speech emotion.

5. An attention-based network incorporated with regularized feature selection has been developed for accurate recognition of emotion from speech utterances

1.7 Datasets Sources

Datasets utilized for this thesis were mostly simulated and synthetic speech emotional corpora. These datasets which include, TESS[39], RAVDESS [28], EMODB [10], and EMOVO [18] are open-source and publicly available. The data were utilized directly for validation of all the models created in this thesis and to assist in evaluating and benchmarking results from the models proposed. It must be stated also that all sources for data used in this thesis were duly acknowledged.

Chapter 2

Literature Review

2.1 Introduction

This section of the study provides an extensive overview of the many cutting-edge deep learning methods and techniques applied to speech emotion classification. Many scholars have put forth several learning algorithms, and these have had a significant impact on the results of emotion classification from speech signals. The study examined the common evaluation metrics used for performance evaluations, examined the popular convolutional neural network architectures used for classification and provided a critical analysis of the effectiveness of some deep learning models on public speech emotion classification datasets.

2.2 Speech Emotion Classification: A Survey of the State-of-the-Art

2.2.1 Brief Review

This section presents a review paper on approaches and techniques in speech emotion classification. The paper studied a comparative analysis of the conventional approach and the performance of some of these approaches on the speech emotion benchmark highlighting the strengths and weaknesses of each approach. This review served as a guide in choosing the appropriate algorithms for future work in speech emotion classification.

Paper status: Published in Springer, LNCS, Pan-African AI & Smart Systems, 2023.

Speech Emotion Classification: A Survey of the State-of-the-Art

Samson Akinpelu(✉) and Serestina Viriri

School of Mathematics, Statistics and Computer Science,
University of KwaZulu-Natal, Durban, South Africa
222068579@stu.ukzn.ac.za, viriris@ukzn.ac.za

Abstract. Technological advancement and rapid growth in Artificial Intelligence (AI) with the corresponding non-availability of sufficient dataset for training the machine learning algorithms has paved the way for applying deep learning techniques for classifying human emotion from auditory speech. The study presents a full survey of the state-of-the-art algorithms and approaches for performing speech emotion classification. Comparative analysis of existing methods for extracting features from the speech signal, a critical review of the performance evaluation of specific algorithms developed for carrying out speech emotion analysis, coupled with the study of evaluation metrics used for performance analysis is presented. The major strength and weaknesses of the algorithms examined were highlighted. Ultimately, the best-performing algorithm can be inferred from the comparison. This paper provides a survey with the utmost aim of revealing how most deep learning techniques outperform conventional algorithms for speech emotion classification.

Keywords: Classification · Speech emotion · Performance metrics · Deep learning · Classifiers

1 Introduction

Auditory speech interaction seems to be the most simple and convenient mode of human communication. In addition to linguistic information such as connotation and dialect type, speech signals carry a wealth of non-linguistic information such as facial expressions, speech emotion, and so on. Speech emotion classification [1] (SEC) has become increasingly important in affective computing and human-machine interactions in recent years, as a result of notable advancements in computer vision, artificial intelligence, and machine learning. In many publications, it is also popularly known as speech emotion recognition (SER). Speech emotion classification entails recognizing the emotional feature of speech regardless of the actual meaning. Though individuals can achieve this task as a natural component of verbal communication, the possibility of accomplishing this automatically and more accurately using a computer device is still research in progress [2,3]. The innovation by which a computer can automatically and

T. M. Ngatched Nkouatchah et al. (Eds.): PAAISS 2022, LNICST 459, pp. 379–394, 2023.
https://doi.org/10.1007/978-3-031-25271-6_24

accurately understand human emotion through speech has piqued the interest of several researchers. SEC is particularly beneficial in applications that require normal human-machine interaction, such as e-learning, customer support and online movies, where the response of the user is determined by the detected emotion [4]. It is also helpful for in-car board systems, where the system can use information about the driver's mental state to initiate vital safety measures [5]. It can also be used as a clinical diagnosis for patients who are suffering from a mental disorder. It is applicable in automatic translation systems where the speaker's emotional state is a factor in communication between parties. However, the lack of a sufficient label dataset to train machine learning algorithms, in identifying human emotion has encouraged the application of deep transfer learning [6]. The specific contribution of this study is to compare several cutting-edge and state-of-the-art techniques for classifying emotion from speech signals. We carried out a comprehensive survey of eliciting emotion from human auditory speech using the conventional approach of classification, neural network and deep learning techniques. Possible combinations of the traditional approach of classification and deep learning (Convolutional Neural Network and Recurrent Neural Network) also known as ensemble methods for improved classification accuracy were highlighted. Researchers in the speech emotion domain and affective computing will be thoroughly furnished with the growth of research in emotion classification and enhancement in models for more accurate recognition of emotion from this study.

1.1 Emotion Classification

Emotion is a positive or negative mental state that is linked to a sequence of physiological activities. Emotions describe an individual's psychological condition. It is impossible to separate it from man, as it is exhibited at one point or another. Emotion is a dominant factor in human attitudinal behaviour and comportment, according to scientific findings. Personality theory, which reveals human actions and inactions, has a significant relationship with the emotional state of people.

The obvious reason why emotion classification has attracted so many scholars in the last decade is that man and his emotional traits are inextricably linked. It influences the creation of higher levels of awareness during embryogenesis and determines the content and structure of consciousness throughout a lifetime [7]. Majorly, emotion falls into two categories, which are positive (happiness, surprise, excitement) and negative (sadness, anger, disgust, fear) emotions respectively.

Emotions, like any other neurobiological activity, range in intensity from low to high. SADNESS, HAPPINESS, DISGUST, ANGER, SURPRISE, and FEAR are the six main categories of emotion identified by Paul Ekman's study

as depicted in Fig. 1. Disgust and anger, for example, may combine to generate a new emotion called con-tempt and so when these six basic emotions unified, there is a likelihood for more complex emotions to emerge [8].

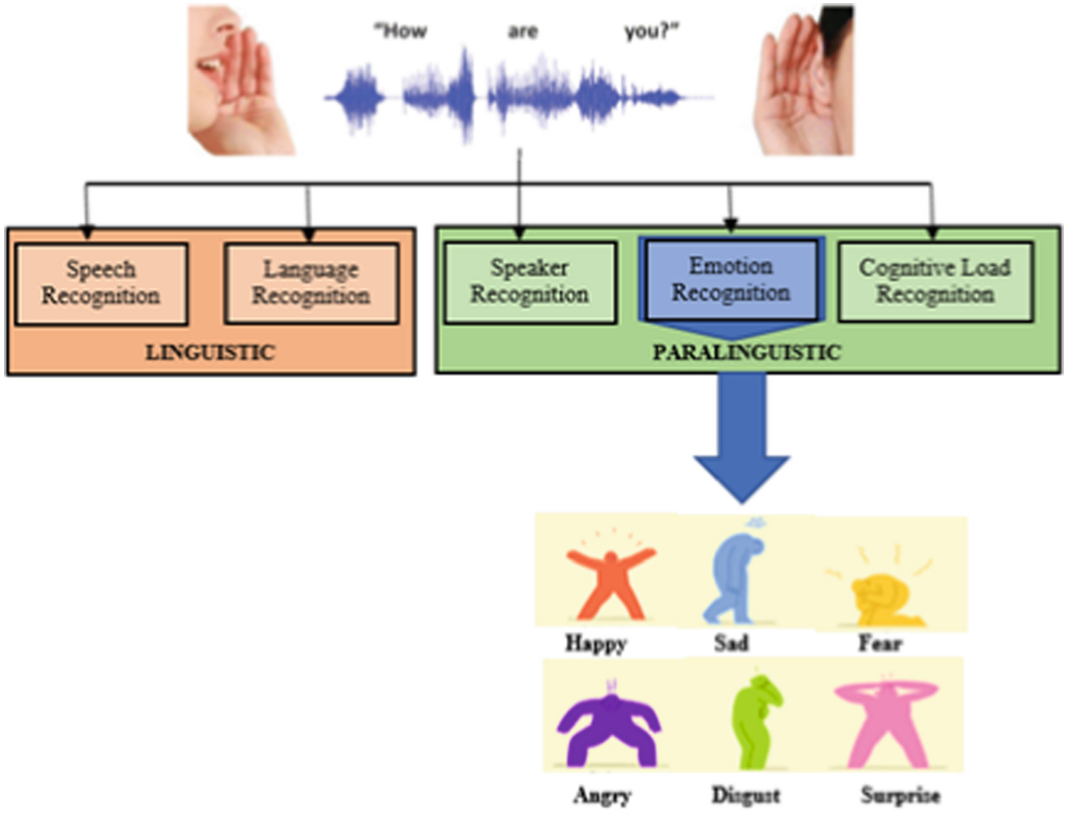


Fig. 1. General framework of emotion classification

Certain real-life scenarios have shown that emotion is transferable (though not always), and the explicit purpose of human social interaction will be meaningless without an accurate classification of emotion [9]. Many relationships and societal groups have been battered because of the inability to manage emotional outbursts. A variety of methods for classifying emotion have been developed, including facial image expression, audio speech, and behavioural traits just to mention a few, however, the focus of this study is on speech utterance [5].

The remaining section of this study is as follows. Section 2 entails a discussion on speech emotion classification with corresponding classification algorithms. Section 3 discusses various evaluation metrics for emotion classification models. A critical review of state-of-the-art methods of emotion classification was carried out in Sect. 4 and Sect. 5 concludes the study.

2 Speech Emotion Classification

The study of emotional attributes from speech signals is known as speech emotion classification (SEC). It is the most prevalent and suitable means of recognizing

human emotion, therefore it's no surprise that it's attracting a growing number of scholars with the potential to expand research in Human-Computer Interaction (HCI). To detect emotional content from a speech signal, various approaches such as Mel-frequency cepstrum coefficient (MFCC), log-mel, prosodic and spectrogram have been applied to extract speech features before classification into different emotions takes place [9,10].

The following reasons have made speech emotion classification a thought-provoking task. Foremost, it is uncertain which features of the speech signal can be optimally used for differentiating between emotions. The variation introduced by different speakers, speaking styles, and tempo has added another obstacle because these properties have a huge impact on commonly extracted features like pitch and formants [11]. Insufficient dataset with which deep learning techniques can be trained and the absence of an indigenous language ascent dataset are subjects of concern in SEC.

A comparison of several techniques and their performance using some well-known evaluation metrics is carried out in this study. The performance of several techniques used in the main conference on emotion classification was also investigated in this work. The task was divided into three main components: dataset preprocessing and augmentation, feature extraction, and emotion classification (sad, happy, angry, etc.). The accuracy of performance evaluations as well as other evaluation measures was critically examined.

2.1 Speech Emotion Classifiers

With a broad range of research articles published a few years ago, the subject of classification algorithms and techniques has been deemed a significant aspect of Machine Learning (ML). The term "classification" in ML has typically been used in a broad sense, encompassing supervised, unsupervised, and semi-supervised learning algorithms. The goal of unsupervised learning is to find and analyze the structure of unlabeled data. Each data input is pre-assigned a class label in supervised learning, to which speech emotion classification also belongs [12,13].

The classification techniques usually require a split of the dataset into training and testing data. In the recent past, there has been a notable development in research for the development of multiclass classification algorithms and techniques for speech emotion classification. One of the popular classification algorithms is rule-based which classifies data by using a group of "if. . . then. . . " rules. It follows a set of conditions having conjunctions of attributes, before arriving at a conclusion. A decision tree is a primary example of rule-based classification [14] (Mulongo Pihlqvist, 2018; Jasmeet-Kaur, 2020) [15]. Other classification algorithms that have been used in the time past are Support Vector Machine (SVM), Hidden Markov Model (HMM), Artificial Neural Network (ANN) and K-Nearest Neighbour (KNN). Majorly, classification algorithms are subdivided into four categories: Binary, Multiclass, Multi-label and Imbalanced classification as depicted in Fig. 2. These algorithms and their combinations have been applied to emotion classification by various researchers and their performance

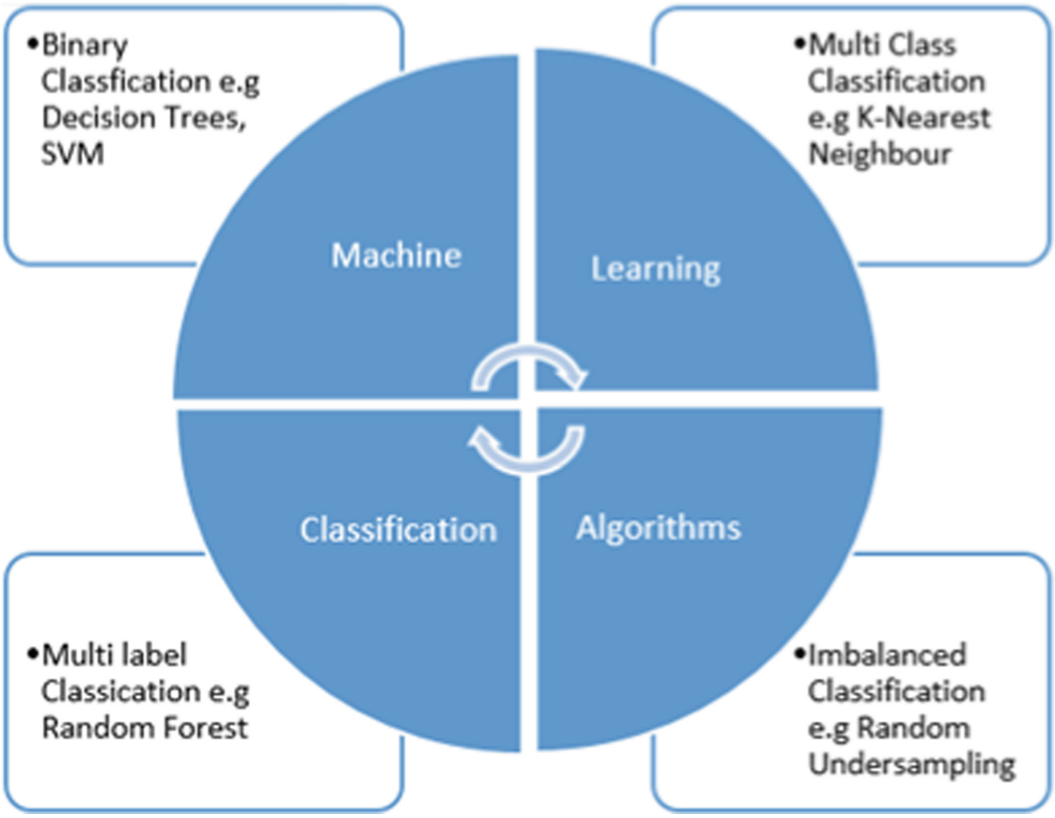


Fig. 2. Categories of classification algorithm

verified using certain standard evaluation metrics [12]. A comparison between these algorithms is shown in Table 1.

2.2 Support Vector Machine (SVM)

SVM is a typical supervised machine learning algorithm that performs well in many classification problems. It uses distinct patterns by splitting hyperplane and kernel functions in modelling non-linear decision margins. SVM ensures that margins between various classes of the dataset (emotional features) are maximized by constructing the best hyperplane. SVM was utilized in [16] to classify three distinct human emotions (happy, sad and neutral) from Berlin Emo-DB and Chinese speech emotion corpus. In their experiments, emotional speech features (Mel-frequency Cepstral Coefficient, energy and pitch) were extracted from the speech signals, and they achieved 91.3% and 95.1% on both datasets respectively. SVM performs classification on input data [17] using Eqs. 1 and 2.

$$k(c, c_i) = (\gamma c^t c_i + m)^d \tag{1}$$

$$c = \sum_{i=1}^{n} w_{gi} k(c_i, c) + b \tag{2}$$

where m is a constant, k represents the kernel function, c is the input data, w_{gi} represents a weight, with b as a bias, c_i is the support vector and d denotes the degree of the polynomial function γ.

2.3 Hidden Markov Model (HMM)

As its name implies, HMM follows the Markov process with an unobserved event in the statistical Markov model in which the probability of a new event relies on the previous event. The word "hidden" denotes the ineptness to recognize the process that produces the event at a certain moment in time. Then, using the framework in conjunction with the target realities of the current event, it is possible to utilize a likelihood to predict the subsequent event. HMM has been successfully used to classify speech emotions in [18] where an accuracy rate of 89.2% was recorded and they established the fact that recognition of emotion with log frequency coefficient features is higher than human vocal tract features.

2.4 K-Nearest Neighbor (KNN)

K-NN is a classification algorithm that is based on propinquity (nearness). It is a supervised classifier known for its convenience and ease of implementation in tackling classification tasks. It performs the classification of data using Euclidean Distance as indicated in Eq. 3.

$$d(x, y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{3}$$

where x, y represents two points in Euclidean space, x_i, y_i represents the vectors of Euclidean and N represents the N-th space. The value of k (k-nearest) determines which class the will data be assigned to, among its neighbours. In [19], KNN and artificial neural networks were applied in classifying emotions from a speech corpus that was centred on two different languages. An accuracy of 69.89% was recorded with KNN after the experiment was carried out using linear prediction and frequency cepstral coefficient (feature) and Hurst parameters.

2.5 Decision Tree (DT)

The decision tree is based on a divide and conquers algorithmic approach most suited for a no-linear classification problem. It can be simply described as a tree structure, having several nodes, leaves, roots and branches. DT in a classification task accepts discrete values based on binary recursive partitioning, which involves dividing the data into subsets and then further dividing those subsets into smaller subsets. The method then terminates the procedure once all the requirements have been successfully completed and the subset data is sufficiently homogeneous. A decision tree approach was applied in [20] for the

Table 1. Summary of comparison of State-of-the-art algorithm

Algorithm	Brief description	Strength	Weakness	Author
SVM	It's a linear model that can be used to solve classification and regression problems. The essential concept is that two classes are separated by a hyperplane defined by its normal vector and bias [21]	Performs well when there is a clear margin of distinction between classes. Highly effective when the dimensionality exceeds the number of samples. It conserves memory	It suffers from over-fitting with a large dataset Fails, when the dataset contains more noise, such as overlapping target classes	Vladimir Vapnik in the 1970s
HMM	This is a statistical model that describes the sequences of events. It consists of a Markov chain whose internal behaviours and states remain hidden from the observer [22]	Its principle can be adapted to many tasks [23]	Highly expensive in terms of computational time and memory consumption [24]	Baum L. E and Petrie, 1966
KNN	This is a form of machine learning algorithm that can handle classification and regression problems. K-represent the number of neighbours [25]	It is not cumbersome to implement. i.e., simple and easy. Also, it has no assumption	It has the major drawback of being substantially slower as the size of the data in use grows. It is a lazy learning algorithm [15]	Evelyn Fix and Joseph Hodges in 1951
Decision Trees	DT is a classifier that recursively partitions data space in such a way that it can be described as a collection of related rules. It is usually made up of various nodes that symbolize a branch of a rooted tree. [26]	Simple, easy to understand and robust to outliers	Not suitable for large datasets because it results in overfitting. Prone to the wrong prediction as a result of noise [27]	Around 1970s s (First Version) Breiman, Stone, Friedman, and Olshen

classification of speech emotion and an accuracy rate of 82.9% was achieved on the EmoDB speech corpus.

3 Performance Evaluation Metrics for the State-of-the-Arts Classification Algorithms

When evaluating and comparing different classification models' multiclass problems, performance metrics are immensely important. Many measures can be used to evaluate a multi-class classifier's performance [28]. The performance of various speech emotion classification algorithms on speech corpus is presented in this paper. Several metrics for analyzing the output quality of speech emotion classification algorithms are employed to evaluate the classification results. Confusion matrix, sensitivity, specificity, accuracy, Mathew Correlation Coefficient (MCC), and average precision are all commonly used metrics for evaluating algorithm performance. There are four essential parameters in getting these metrics:

True positives (TP): Is a key parameter in evaluation metrics which shows whether an observation predicted to belong to a class really belongs to that class i.e., $TP \rightarrow (M \in Y) = 1$, where M is the predicted value and Y is the actual value and 1= +ve, 0 = -ve.

True negatives (TN): Usually indicate that the predicted observation not belonging to a class does not actually belong to that class in the real sense i.e., $TN \rightarrow (M \notin Y) = 0$

False positives (FP): As its name implies, happens when the prediction shows that an outcome belongs to a class when it does not, in the real sense. In other words, both the prediction and the actual are not having the same data point i.e., $FP \rightarrow (M = 1, Y = 0)$

False negatives (FN): This happens when the prediction indicates a false identity of observation of not belonging to a class when it belongs to that class i.e., $FN \rightarrow (M = 0, Y = 1)$
The proper combination of these parameters can then be used as the basis for computing the metrics mentioned above, as follows:

Confusion Matrix: Is a summary of prediction results on a classification task. The number of correct and wrong predictions are reported with count values and simplified by each class. It is one of the easiest means of determining the performance of a classification algorithm.

Accuracy: It can be represented by the ratio of correct predictions to total predictions. The accuracy score function from the "sci-kit learn" python library can be used to compute the accuracy of a classification algorithm or model. Mathematically:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

Specificity: It is the ratio of actual negative to predicted negative

$$Specificity = \frac{TN}{FP + FN} \tag{5}$$

Sensitivity: This is the ratio of actual positives to predicted positives.

$$Sensitivity = \frac{TP}{TP + TN} \tag{6}$$

Precision: It is the ratio of relevant samples that are true positives out of all the samples which were predicted to belong in a certain class.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Mathews Correlation Coefficient (MCC): The quality of multiclass classification in machine learning is measured using MCC. Because, it considers both true and false positives and negatives, regardless of class size, it is commonly viewed as a balanced metric. It stores correlation coefficients between -1 and $+1$, with $+1$ indicating a perfect prediction, 0 indicating an average random prediction, and -1 indicating a reverse prediction. It is defined as follows:

$$MCC = \frac{(TP.TN) - (TP.FN)}{\sqrt{(TP + FP).(TP + FN).(TN + FP).(TN + FN)}} \tag{8}$$

Alternatively, given a matrix C with k classes, MCC can be computed as:

$$MCC = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p^2)(s^2 - \sum_k^K p_k \times t_k^2)}} \tag{9}$$

where;
$c = \sum_k^K C_k k$ represent sum of correctly predicted elements
$s = \sum_i^k \sum_j^K C_i j$ represent cumulative sum of elements
$p_k = \sum_i^K C_k i$ total number of occurrences of class k prediction
$t_k = \sum_i^K C_k i$ number of true occurrence of class k
However, the high performance of a given method or algorithm can be measured by corresponding high sensitivity and specificity. Investigation from this study reveals that MCC has not witnessed huge application as others in both binary and multi-class classification problems.

4 Speech Emotion Classification: A Critical Analysis of State-of-the-Art Techniques and Algorithms

Various Deep Neural Network (DNN) algorithms for speech emotion classification have been developed and applied in the past. DNN has a general framework of an input layer, several hidden layers and an output layer. Speech features are usually extracted through the hidden layer of a typical deep-learning algorithm for speech emotion classification. This section describes how some algorithms have been used to classify emotions using one or more speech corpus, and their

corresponding performance. Table 2 shows the performance evaluation results of these algorithms. In [29] a machine learning approach (KNN, SVM, DT and Random Forest) for classifying speech emotion from the speech signal was presented. MFCC was used in extracting the features from the spectrogram because of its dimensionality and computational time reduction capability it possessed. Average accuracy of 75%, 90% sensitivity and 91% of specificity was reported after an experiment was carried out using TESS and KEEL speech datasets. However, most of the algorithms used in this work in detecting emotion from the speech were conventional and may not be able to handle a large dataset as obtainable in neural network techniques. Also, only MFCC cannot yield a high recognition rate of emotion.

Using a Recurrent Neural Network (RNN) and Multi-Head attention-based mechanism, [30] proposed a multimodal-based approach for SEC. The proposed approach is based on two different forms of speech representations: the MFCC of an audio signal and word embedding from text data. They achieve state-of-the-art performance on the IEMOCAP, MELD, and CMU-MOSEI datasets by training these features in temporal space. Though a higher rate of accuracy was recorded, no fine-tunning in the model was adopted and also other speech features like chroma and prosodic were not captured. The algorithm is susceptible to gradient vanishing and difficulty in training.

In [31], a Multi-scale discrepancy adversarial network for cross-corpus speech emotion classification that uses various timelines of deep speech features to simultaneously train a collection of hierarchical domain discriminators and an emotion classifier in an adversarial training network was proposed. To determine cross-corpus efficiency, the research was bench-marked on three major speech datasets (IEMOCAP, CASIA, and MSP), each of which contained two different languages (English and Chinese). Although there was considerable improvement, no baseline comparison was carried out.

A robust and versatile deep learning emotion recognition system based on the analysis of speech signals using a combination of MFCC, HNR, ZCR and TEO parameters with SVM at first and later with Auto-Encoder (AE) was proposed in [32]. An RML (Ryerson Multimedia Lab) emotion corpus consisting of 720 auditory human emotional expression samples of Anger, Disgust, Happy, Fear, Surprise and Sad were employed. About six languages were contained in the corpus used for this work. They show that auto-encoder dimension reduction can improve recognition rate with an accuracy score of 74.07% and 72.83% for both methods as against another baseline approach of emotion classification. The ability to maintain the same accuracy score and recognition rate with multiple speech databases of larger samples is of major concern in this work.

In the work of [33], an improved speech emotion classification using transfer learning with spectrogram augmentation was presented. Transfer learning is a deep learning technique that leverages a pre-trained neural network model. It usually yields better results even with a small dataset. A pre-trained ResNet model was adopted in their work and features were extracted from high-resolution log-Mel spectro-grams using the convolution layer of the neural

network model. Additional data samples were generated through spectrogram augmentation. The performance evaluation of their work was carried out on the interactive emotional dyadic motion capture (IEMOCAP) dataset and the result indicated that transfer learning with spectrogram augmentation can improve the rate of emotion recognition. However, the mode of fine-tuning of pre-trained neural network model which is the major key in transfer learning was not stated. Also, loss of feature or distortion on the spectrogram image can occur if the augmentation technique of the data sample is not efficient. This work did show a comparison of the original spectrogram data and augmented ones if the quality remains the same before feeding it into the CNN model adopted. Both facial and audio data were utilized in [6] to perform embedding extraction and fine-tuning a transfer learning model for multimodal emotion classification. Despite the poorer performance of the visual modality compared to the speech modality, after fine-tuning the Convolution Neural Network (CNN)-14, the fusion of both inputs yielded 80.08% emotion classification accuracy.

Zhang et al. [34] applied DCNN-BLSTMwA for enhanced speech motion classification. Speech samples were first preprocessed by data enhancement and dataset balancing. Their model was pre-trained on ImageNet architecture to generate the segment-level features. Eventually, the Deep Neural Network (DNN) was fed with learned high-level emotional features in order to predict human emotion. The result of the performance analysis of the proposed model was carried out on two popular speech databases (EMO-DB and IEMOCAP). An average score of 87.86% and 68.50% for Sensitivity and Specificity was recorded respectively. The work shows high performance in recognition, but the model tends to require high computational time and a large dataset. To improve the accuracy of a speech emotion classification, a novel feature reduction approach was proposed in [35]. They used OpenSMILE-2.3.0 to extract speech feature sets, such as the waveform, Mel/Bark spectrum, FFT spectrum, speech quality, signal energy, and formant, to mention but a few for a total of 384-dimensional features, at the feature extraction stage of their work. The result showed an accuracy score of 83.70%. Dangol et al. [36], proposed a 3D-CNN-based LSTM network with a relation-aware attention mechanism for speech emotion classification. The peculiar situation of overfitting was overcome by oversampling and deep learning black box techniques. A higher accuracy score of 81.05% in detecting mood disorders was achieved after the training of the model. Atila & Şengür, [37] applied 3D CNN-LSTM and attention for accurate classification of speech emotion. In their proposed method, six 3D CNN, two batch normalizations, five ReLu activation, three 3D max pooling and one LSTM layer were used. An experiment was carried out on three datasets (RAVDESS, RML, SAVEE). The novel approach achieved 94.17% sensitivity and 99.09% specificity which proved the efficiency of their model.

Table 2. Summary of comparison of State-of-the-art algorithm

Techniques	Year	Dataset	No. Emotion	Reported accuracy
ANN-LSTM [38]	2017	LDC	4	87.5%
ANFIS-MLP [10]	2017	Berlin EOMD	3	72.5%
Multi-SVNN [39]	2018	Berlin EOMD	–	81.37%
Deep RNN [40]	2018	IEMOCAP	4	70.0%
Deep CNN-LSTM [41]	2019	USC-IEMOCAP	3	52.90%
Transfer Learning SoundNet [3]	2019	MASC	6	73.6%
2D, 3D CNN-LSTM [42]	2019	AEFW	7	60.9%
Deep learning 39 MFCC-HNR-ZCR-TEO [32]	2020	RML	6	72.5%
Transfer learning Siamese NN [42]	2020	RAVDESS, CREMA, eENTER-FACE05	4	50.0%
CNN-LSTM [36]	2020	IEMOCAP, Berlin EOMD, SAVEE	4	84.6%
MFCC-NSL [43]	2020	Synthetic	4	84.25%
HSF-DNN, MS-CNN, LLD-RNN [44]	2020	IEMOCAP	4	53.6%
Deep learning GoogleNet	2020	DEAP	3	83.59%
3D CNN-LSTM [37]	2021	RAVDESS, RML, SAVEE	8	96.18%
3D CNN-Esemble learning [45]	2021	DEAP	Valence/Arousal	96.13%
Transfer learning CNN-14 from PAN [6]	2021	RAVDESS	8	76.58%
Deep learning fusion spatial feature [46]	2021	IEMOCAP, RAVEDESS	4, 8	77.5%
Transfer learning Wav2vec2.0 [47]	2021	RAVDESS	8	80.46%
EEG-BiLSTM-DRNN [48]	2022	DEAP, SEED, IDEA	8	59.0%
CNN-BiLSTM-MLP-FSL [49]	2022	Prototype utterance	–	39.0%
DCERNet-SVM [50]	2022	DEAP	4	93.0%

5 Conclusion

Human auditory speech possessed innate features for accurate prediction of emotion, as compared to facial expression, hence the reason why SEC has been attracting huge researchers globally in the past decades. The state-of-the-art techniques for speech emotion classification have been critically and analytically surveyed in this study. The performance of some algorithms as well as evaluation metrics for speech emotion classification were investigated. The strengths and weaknesses of these algorithms were also examined. It was observed that the application of machine learning techniques such as deep learning and transfer learning for performing analysis on speech signals gives a better performance in speech analysis most especially in the classification of speech emotion. The growth of research in speech emotion classification has indeed witnessed a sporadic increase in the last five years and yet is still increasing, through the application of deep learning as it was revealed through some scientific journals, online repositories and archives.

References

1. Pham, N., Dang, D., Nguyen, S.: A method upon deep learning for speech emotion recognition. J. Adv. Eng. Comput. **4**, 273–285 (2020). https://doi.org/10.25073/jaec.202044.311
2. Chenchah, F., Lachiri, Z.: Speech emotion recognition in acted and spontaneous context. Procedia Comput. Sci. **39**(C), 139–145 (2014). https://doi.org/10.1016/j.procs.2014.11.020
3. ElShaer, M.E.A., Wisdom, S., Mishra, T.: Transfer learning from sound representations for anger detection in speech (2019). arXiv:1902.02120
4. Papakostas, M., Giannakopoulos, T.: Speech-music discrimination using deep visual feature extractors. Expert Syst. Appl. **114**, 334–344 (2018). https://doi.org/10.1016/j.eswa.2018.05.016
5. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recogn. **44**(3), 572–587 (2011). https://doi.org/10.1016/j.patcog.2010.09.020
6. Luna-Jiménez, C., et al.: A proposal for multimodal emotion recognition using aural transformer on RAVDESS. Appl. Sci. MDPI **12**, 327 (2022). https://doi.org/10.3390/app12010327
7. Izard, C.: Emotion theory and research: highlights, unanswered questions, and emerging issues. Annu. Rev. Psychol. **60**(3955), 1–25 (2009). https://doi.org/10.1146/annurev.psych.60.110707.163539
8. Ekman, P.: Basic-Emotions by Paul Ekman. Book Chapter, San Franciso, USA (1993)
9. Lu, Y.: Transfer learning for image classification (2019). https://tel.archives-ouvertes.fr/tel-02065405
10. Motamed, S., Setayesh, A., Rabiee, A.: Speech emotion recognition based on a modified brain emotional learning model. Biologically Inspired Cogn. Archit. **19**, 32–38 (2017). https://doi.org/10.1016/j.bica.2016.12.002
11. Wang, Y., Boumadane, A., Heba, A.: A Fine-tuned Wav2vec 2.0/Hubert Benchmark For speech emotion recognition, speaker verification and spoken language understanding (2021). arXiv:2111.02735

12. Pérez-Ortiz, M., Jiménez-Fernández, S., Gutiérrez, P.A., Alexandre, E., Hervás-Martínez, C., Salcedo-Sanz, S.: A review of classification problems and algorithms in renewable energy applications. Energies MDPI **9**(8), 607 (2016). https://doi.org/10.3390/en9080607
13. Vijaya, R., Reddy, K., Ravi-Babu, U: A Review on Classification Techniques in Machine Learning (2018). www.ijarse.com
14. Mulongo, B., Pihlqvist, F: Speech emotion recognition: using rule-based methods and machine learning for short answer scoring. KTH Royal Institute of Technology, trita-eecs-ex (2018). https://www.kth.se/en
15. Jasmeet-Kaur, A.: Databases, features and classification techniques for speech emotion recognition. Int. J. Innovative Technol. Exploring Eng. **9**(6), 185–190 (2020)
16. Seehapoch, T., Wongthanavasu, S.: Speech emotion recognition using support vector machines. In: Proceedings of the 2013 5th International Conference on Knowledge and Smart Technology, KST, vol. 6(2), pp. 101–108 (2013). https://doi.org/10.1109/kst.2013.6512793
17. Farooq, M., Hussain, F., Baloch, N., Raja, F., Yu, H., Zikria, Y.: Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. Sensors **20**(21), 1–18 (2020). https://doi.org/10.3390/s20216008
18. New, T.L., Foo, S.W., De Silva, L.C.: Detection of stress and emotion in speech using traditional and FFT based log energy features. In: ICICS-PCM Proceedings of the 2003 Joint Conference of the 4th International Conference on Information, Communications and Signal Processing and 4th Pacific-Rim Conference on Multimedia, vol. 3, pp. 1619–1623 (2003)
19. Rejith, S., Manju, K. G.: Speech based emotion recognition in Tamil and Telugu using LPCC and Hurst parameters- a comparative study using KNN and ANN classifiers. In: Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies, ICCPCT, pp. 1–6 (2017)
20. Yuncu, E., Hacihabiboglu, H., Bozsahin, C.: Automatic speech emotion recognition using auditory models with binary decision tree and SVM. In: Proceedings of International Conference on Pattern Recognition, pp. 773–778 (2014). https://doi.org/10.1109/ICPR.2014.143
21. Schnall, A., Heckmann, M.: Feature-space SVM adaptation for speaker adapted word prominence detection. Comput. Speech Lang. **53**, 198–216 (2019). https://doi.org/10.1016/j.csl.2018.06.001
22. Mao, S., Tao, D., Zhang, G., Ching, P.C., Lee, T.: Revisiting hidden Markov models for speech emotion recognition. In: ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pp. 6715–6719 (2019)
23. Chakraborty, C., Talukdar, P.: Issues and limitations of HMM in speech processing: a survey. Int. J. Comput. Appl. **141**, 13–17 (2016). https://doi.org/10.5120/ijca2016909693
24. Degirmenci, A.: Introduction to Hidden Markov Models (2014). https://scholar.harvard.edu/files/adegirmenci/files/hmm_adegirmenci_2014.pdf
25. Venkata Subbarao, M., Terlapu, S.K., Geethika, N., Harika, K.D.: Speech emotion recognition using k-nearest neighbor classifiers. In: Shetty D., P., Shetty, S. (eds.) Recent Advances in Artificial Intelligence and Data Engineering. AISC, vol. 1386, pp. 123–131. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-3342-3_10
26. Liu, Z., Wu, M., Cao, W., Mao, J., Xu, J., Tan, G.: Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. Sensors **273**, 271–280 (2018). https://doi.org/10.1016/j.neucom.2017.07.050

27. Kim, M., Yoo, J., Kim, Y., Kim, H.: Speech emotion classification using tree-structured sparse logistic regression. Interspeech **12**, 1541–1545 (2015). https://doi.org/10.21437/Interspeech.2015-337
28. Grandini, M., Bagli, E., Visani, G.: Speech emotion detection using machine learning techniques (2020). arXiv:2008.05756
29. Sundarprasad, N.: Metrics for multi-class classification: an overview (2018). https://doi.org/10.31979/etd.a5c2-v7e2
30. Ho, N., Yang, H., Kim, S., Lee, G.: Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. IEEE Access **8**, 61672–61686 (2020). https://doi.org/10.1109/ACCESS.2020.2984368
31. Wanlu, Z., Wenming, Z., Yuan, Z.: Multi-scale discrepancy adversarial network for cross-corpus speech emotion recognition. Virtual Real. Intell. Hardw. **3**(1), 57–76 (2022). https://doi.org/10.1007/s40747-021-00637-x
32. Aouani, H., Ayed, Y.: Speech emotion recognition with deep learning. Procedia Comput. Sci. **176**, 248–260 (2020). https://doi.org/10.1016/j.procs.2020.08.027
33. Padi, S., Sadjadi, S., Sriram, R., Manocha, D.: Improved speech emotion recognition using transfer learning and Spectro-gram augmentation. In: ICMI- Proceedings of the International Conference on Multimodal Interaction, pp. 645–652 (2021). https://doi.org/10.1145/3462244.3481003
34. Zhang, H., Gou, R., Shang, J., Shen, F., Wu, Y., Dai, G.: Pre-trained deep convolution neural network model with attention for speech emotion recognition. Front. Physiol. **12**, 643202 (2021). https://doi.org/10.3389/fphys.2021.643202
35. Zhang, Z.: Speech feature selection and emotion recognition based on weighted binary cuckoo search. Alex. Eng. J. **60**(1), 1499–1507 (2019). https://doi.org/10.1016/j.aej.2020.11.004
36. Dangol, R., Alsadoon, A., Prasad, P.W.C., Seher, I., Alsadoon, O.H.: Speech emotion recognition UsingConvolutional neural network and long-short TermMemory. Multimed. Tools Appl. **79**(43), 32917–32934 (2020). https://doi.org/10.1007/s11042-020-09693-w
37. Atila, O., Şengür, A.: Attention guided 3D CNN-LSTM model for accurate speech-based emotion recognition. Appl. Acoust. **182**, 108260 (2021). https://doi.org/10.1016/j.apacoust.2021.108260. Frontiers in Physiology, 12
38. Thirukumaran, S., Archana, A.F.C.: Speech emotion classification analysis using short-term features. Fron. Physiol. J. Sci. EUSL **8**(1), 13–22 (2017)
39. Mannepalli, K., Sastry, P., Suman, M.: Emotion recognition in speech signals using optimization based multi-SVNN classifier. J. King Saud Univ. Comput. Inf. Sci. **34**, 384–397 (2018). https://doi.org/10.1016/j.jksuci.2018.11.012
40. Chernykh, V., Prikhodko, P. Emotion recognition from speech with recurrent neural networks (2018). arXiv:1701.08071v2. [CsCL]
41. Cho, J., Pappagari, R., Kulkarni, P., Villalba, J., Carmiel, Y., Dehak, N.: Deep neural networks for emotion recognition combining audio and transcripts (2019). arXiv:1911.00432
42. Ren, M., Nie, W., Liu, A., Su, Y.: Multi-modal correlated network for emotion recognition in speech. Vis. Inform. **3**(3), 150–155 (2019). https://doi.org/10.1016/j.visinf.2019.10.003
43. Uddin, M., Nilsson, E.: Emotion recognition using speech and neural structured learning to facilitate edge intelligence. Eng. Appl. Artif. Intell. **94**, 103775 (2020). https://doi.org/10.1016/j.engappai.2020.103775

44. Yao, Z., Wang, Z., Liu, W., Liu, Y., Pan, J.: Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. Speech Commun. **120**, 11–19 (2020). https://doi.org/10.1016/j.specom.2020.03.005
45. Salama, E.S., El-Khoribi, R.A., Shoman, M.E., Wahby Shalaby, M.A.: A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition. Egypt. Inform. J. **22**(2), 167–176 (2021). https://doi.org/10.1016/j.eij.2020.07.005
46. An, X., Ruan, Z.: Speech Emotion Recognition algorithm based on deep learning algorithm fusion of temporal and spatial features. J. Phys.: Conf. Ser. **1861**(1), 012064 (2021). https://doi.org/10.1088/1742-6596/1861/1/012064
47. Pepino, L., Riera, P., Ferrer, L.: Emotion recognition from speech using Wav2vec 2.0 embeddings (2021). arXiv:2104.03502
48. Joshi, V., Ghongade, R., Joshi, A., Kulkarni, R.: Deep BiLSTM neural network model for emotion detection using cross-dataset approach. Biomed. Signal Proc. Control **73**, 103407 (2022). https://doi.org/10.1016/j.bspc.2021.103407
49. Guibon, G., Labeau, M., Lefeuvre, L., Clavel, C.: Few-shot emotion recognition in conversation with sequential prototypical networks. Softw. Impacts **12**, 100237 (2022). https://doi.org/10.1016/j.simpa.2022.100237
50. Pusarla, A., Singh, B., Tripathi, C.: Learning DenseNet features from EEG based spectrograms for subject independent emotion recognition. Biomed. Signal Proc. Control **74**, 103485 (2022). https://doi.org/10.1016/j.bspc.2022.103485

2.3 Deep Learning Framework for Speech Emotion Classification: A survey of the state-of-the-art

2.3.1 Brief Review

This paper is an extension of the research paper in section 2.2 which presents the paper entitled Deep Learning Framework for Speech Emotion Classification: A Survey of State-of-the-art. The study provides an extensive and critical analysis of many methods—including the most recent state-of-the-art methods—for identifying emotion from speech signals. The effectiveness of state-of-the-art methods, particularly deep learning approaches, was compared to that of traditional procedures. There are several issues with speech emotion classification, which were pointed out and discussed. The effectiveness of the state-of-the-art method for the speech emotion classification tasks was reviewed. This survey's main objective is to identify the shortcomings and difficulties of the current state-of-the-art and suggest a new model to lessen the impact of these challenges on the creation of effective models that reliably and automatically identify emotion from images of human auditory speech.

Paper status: Resubmitted with Minor Correction to Journal Big Data

Deep Learning Framework for Speech Emotion Classification: A Survey of the State-of-the-Art

Samson Adebisi Akinpelu[1] and Serestina Viriri[1*]

[1]School of Mathematics, Statistics and Computer Science
University of KwaZulu-Natal, Durban, South Africa.

*Corresponding author(s). E-mail(s): viriris@ukzn.ac.za;
Contributing authors: 222068579@stu.ukzn.ac.za;

Abstract

The intricate landscape of speech emotion classification poses a captivating yet challenging realm due to emotions being fundamental to human communication. In recent years, deep learning frameworks have emerged as powerful tools, shedding light on the elusive domain of emotion recognition, revolutionizing human-computer interactions, and enhancing the emotional intelligence of artificial intelligence (AI). This survey embarks on an exploratory journey into the forefront of deep learning approaches dedicated to speech emotion classification. The scarcity of extensive speech corpora and the need for high accuracy at low computational cost, deep learning has become the standard approach. One of the main reasons the Deep Convolutional Neural Network (DCNN) is considered the best method is that it is good at extracting important information from spectrogram images. Deep learning has been applied to speech emotion classification by many academics, leading to significant improvements in performance and accuracy. Modern deep learning methods designed for human auditory speech emotion classification are carefully examined in this work. A thorough examination of various deep convolutional neural network designs used in emotion classification is provided, illuminating unique characteristics that capture important data from speech signals for accurate emotion prediction. The research critically analyzes selected deep models using well-established emotion corpora, highlighting their efficacy. This research provides an analysis of typical performance evaluation metrics used to evaluate speech emotion classification models. With this review, we hope to offer a comprehensive overview of the state-of-the-art, potential directions for further investigation, and developing approaches that further the field of speech emotion classification with deep learning frameworks.

Keywords: Human-computer interaction, Deep learning, Speech emotion recognition, Convolutional neural networks, Vision Transformer, mel spectrogram

1 Introduction

Speech emotion classification is a potent, common, and predictable determinant of decision-making. As a result of advancement in HCI, important findings have surfaced that indicate how emotions influence decisions and choices across the entire humanity[129]. A large amount of literature, scholarly publications, and research initiatives devoted to this domain since the end of the first decade of the twenty-first century, have shown that interest in the study of speech emotion classification

continues to grow. Most of the scientific and psychological effort has demonstrated that not only does speech emotion classification form the heart of human life and activity, but it can also be examined utilizing the computing devices of today's modernist concept of science. However, one lingering question is how the use of machine learning approaches to studying the human, normal, and empirically observed dynamics of emotion classification have evolved, changed, or altered their scope [8],[111],[116]. The general overview of emotion classified into five basic categories according to Paul Ekan, (1993) is shown in Figure 1. Speech signal carries both linguistic and para-linguistic information [64].

The role of emotion classification through auditory speech in social interaction cannot be overemphasized. Its application spans self-driving cars, e-learning and online caregiver service [112][171]. As the most natural means of communication, human speech possesses vital information that relates to personal traits, and this information has a major impact on how people interact with each other. This information furnishes us with the required knowledge to recognize racial background, accent, personality trait, and culture; which are from speech utterances [29][94][121]. These speech features are of utmost importance to humanity and significant throughout our life span, but a satisfactory level has not been reached on how to effectively and accurately classify emotion from auditory speech. The fact that speech utterance has a developmental stage from childhood to adulthood poses another key challenge to academia that specializes in speech emotion classification.

However, many researchers and scientific scholars and concerned industries have undauntedly devoted a lot to model design, algorithms and techniques, system testing and quality assurance, among others; yet accurate classification of speech emotion remains a rocky task in this contemporary time. Apart from hereditary factors, there are other notable external contributory factors [230] to speech emotion which are age, gender, culture and sometimes health condition [31] [50].

The conventional approach of classifying emotion from speech utterances is bewildered by some difficulties. Many of these existing approaches such as SVM (Support Vector Machine), HMM (Hidden Markov Model), and GMM (Gaussian Mixture Model) follow automatic speech recognition (ASR) that depends heavily on dataset manipulation and any alteration may require that the entire model is reconstructed [166]. Emotion carries vital information that can either mar or makes an individual personality, the reason why its classification cannot be held with levity. In a bid to avert some of these challenges associated with the traditional approach [109], Deep learning methods have been adopted for the classification of speech emotion and an optimum accuracy has been recorded so far. Transfer learning, a core branch of deep learning which has been proven successful in many computer vision related task, including emotion recognition [112] usually leverages the standardized pre-trained [42] model to overcome the problem of the insufficient training dataset. It is a sub-division of a deep convolutional neural network (DCNN). The DCNN application to emotion classification came to the limelight as a result of its innate capability to extract speech features distinctively and efficiently from speech signals [153][212].

Current findings have revealed the long-standing problems of insufficient label dataset for the classification of speech emotion and high level of parameterization of the field, however, researchers are always on their toes in studying different DCNN techniques [39] that can yield more appreciable results.

Given foregone, this paper presents a survey of different deep learning architectures for speech emotion classification, feature extraction and selection approaches. We study existing speech emotion corpora and standard performance evaluation metrics that are in use. The methodology used in state-of-the-art DCNN and performance evaluation strategy is discussed in detail. We believe that this study is the first to offer a comprehensive review on deep learning framework for speech emotion classification.; previous studies such as Abbaschian et al. [1] and Imani [81] focused on the model, speech database, and algorithmic accuracy. The author in Ruhul et al., [172] and Javier [88] emphasized classical deep learning techniques application for emotion

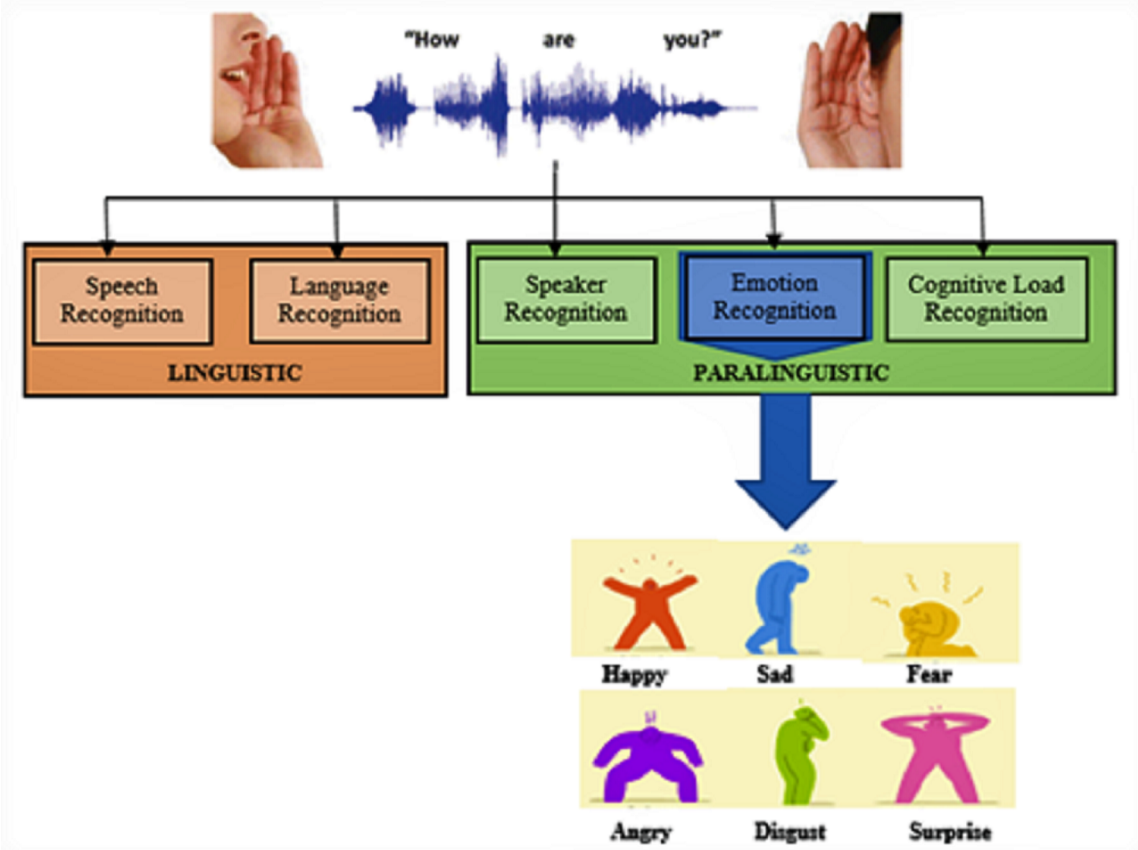


Fig. 1: Basic Speech Emotion Classification Overview

recognition and analysis of peculiar challenges.

The methodology of this review paper goes over a fundamental concept of speech emotion classification and deep learning models. Related publications include a description of the novel deep transfer learning techniques, as well as acoustics and spectral features. We focused on speech emotion corpus, application, and critical analysis of cutting-edge deep transfer learning techniques. To find the most relevant published articles, we searched the Web of Science (WoS) and Scopus citation repositories. Due to the substantial amount of research on this subject, this survey article only reports on a few papers from the most recent six to seven years of thorough study (2017-2023)..

The following is how the rest of the work is organized: we present different deep transfer learning and DCNN architectures as they relate to speech emotion classification in section 2 and 3, popular speech features extraction and selection with an existing speech emotion database in sections 4, previous DCNN approaches on speech emotion classification that has been proposed in section 5, standard performance evaluation metrics in use is highlighted in section 6 and discussion in section 9, with a conclusion in section 10. The main contributions of this review paper are outlined below:

- We present an unambiguous review of various cutting-edge deep learning (CNN) architectures for speech emotion classification, including their benefits and drawbacks.

- We also study the speech emotion corpora and their application in various types of research in speech emotion classification, through which the readers can be thoroughly furnished with the popular speech emotion dataset and why.
- Identification of salient and discriminating features from speech samples using a state-of-the-art deep learning model with a feature selection mechanism was also highlighted.
- We also provide a succinct summary of various deep learning methodologies used in the state-of-the-art speech emotion classification nowadays and their effectiveness.

2 Deep Learning

Deep learning (DL) is a branch of machine learning that can handle nonlinear datasets. It is in most cases interchangeably use as Deep Neural Network(DNN).Deep Layers of stacked nodes make up a DNN, which is often trained through back-propagation and optimization techniques. Each layer has an activation function and corresponding weights [79]. Deep learning has rapidly expanded over the last two decades and is now utilized in many facets of our everyday lives [128]. Since 2011, Convolutional Neural Network (CNN) layers, for example, have enhanced deep learning models for computer vision and pattern recognition related tasks, and as of now, the majority of DLs include CNN[27] layers, which form the basis of Deep Convolutional Neural Network (DCNN).

In a DCNN model (Figure 2, the input data p of each layer is structured in three dimensions: height, width, and depth, or $h \times h \times d$, where the height (h) equals the width. The depth is represented by a corresponding number of a channel. In a survey work by Tan et al.,(2019), deep learning was categorized into four groups, namely: Instances-based, Mapping-based, Network-based and Adversarial-based. They concluded that better results and robustness can be achieved through possible combinations of these various techniques. Deep learning architectures have an in-built feature extraction mechanism for extracting discriminating features from speech signals, which are needed for accurate emotion classification. The major DCNN architectures that are in use are as follows:

2.1 Capsule Network Architecture (CapsNet)

Hinton and his colleagues proposed Capsule Networks as an alternative to CNN [219]. It takes features or output from CNN as its input. The choice of capsule employed will determine how the features will be processed. As opposed to ordinary CNN architecture which only accepts scalar values, CapsNet consists of interconnections of neurons that can receive and outputs vectors[89],[106]. It easily learns features from images and any possible deformations because of this property. Every capsule inside a capsule network is composed of a collection of neurons, and the result of each neuron denotes a distinct property of the same feature, with the advantage of identifying the entire object first before identifying its constituent parts [174]. Traditional CNNs assess the global error that grows toward the rear during training using a specific cost function. In such instances, once the weight between two neurons is zero, the activation of a neuron will not rise anymore [114],[156]. However, CapsNet is referred to as a shallow architecture because it contains only two convolutional layers (one of 256, 9 x 9) and a single fully connected layer as shown in Figure 3. Also, it has not been trained on a large dataset. Hence, it has recorded insignificant applications in the SEC domain.

2.2 AlexNet Architecture

AlexNet was proposed in 2012 by Krizhevsky [103]. It became the first breakthrough of CNN architecture trained on over one million ImageNet datasets. It has a simple design of five convolutional layers with $(11 \times 11, 5 \times 5, 3 \times 3)$ filters, activation layers, max-pooling layer, and dropout (for mitigating overfitting), with three fully-connected layers. Its depth of layers has contributed in no small measure to its increase in performance, though with computational cost. AlexNet won the ILSVRC competition in 2012 as it ranked best among other architecture trained with the largest dataset (ImageNet). It is represented in Figure 4. The original model of AlexNet is trained on two different Graphical Processing Units (GPUs).

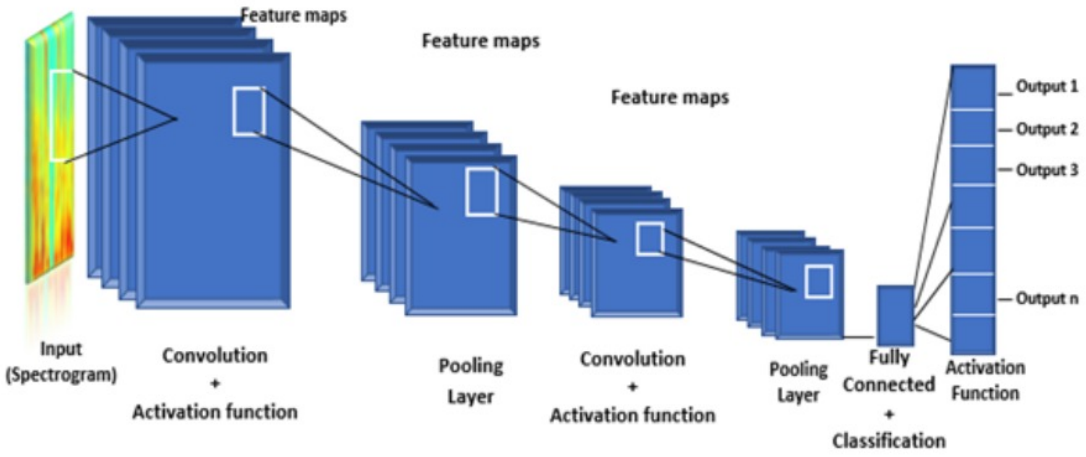


Fig. 2: Overview of Speech Emotion flow with CNN

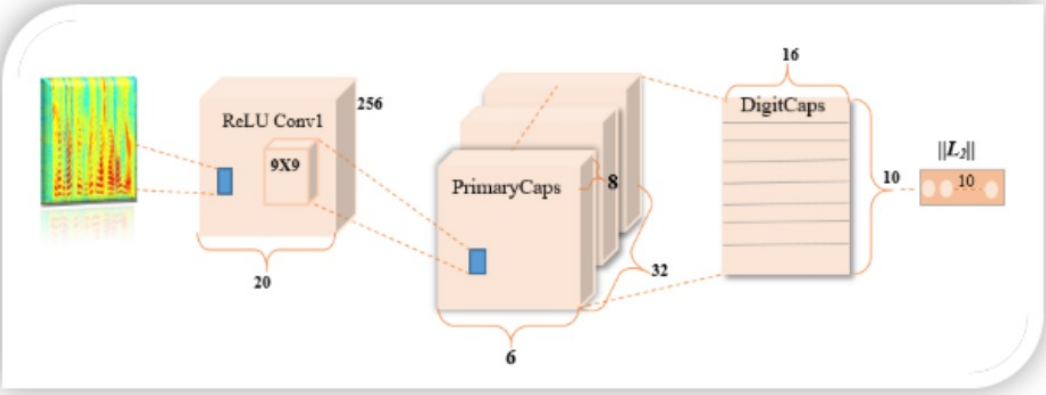


Fig. 3: A simplify three layers CapsNet framework. The DigitCaps layer shows how classification loss can be calculated through the activity vector's length from each capsule [174].

2.3 Visual Geometry Graphics (VGG) Net Architecture

VGGNet is a cutting-edge deep learning architecture proposed after CNN was concluded to be efficient in the pattern recognition domain. The model was developed by Simonyan and Zisserman, (2015). It has a reasonable amount of filters in its architecture (3 x 3) and about 16 to 19 layers [16] for in-depth simulation of network representation capacity. It has an additional max-pooling layer that is used in reducing dimensionality. VGG16 and VGG19 are variants of VGGNet (Figure 5), it is referred to as poor learner convolution network architecture because of the large weight it possessed. It is slow to train VGG from scratch, which invariably increases its high computational cost. This challenge was overcome through an improved design, a pre-trained approach with fewer weights by Simonyan and Zisserman, with the capacity

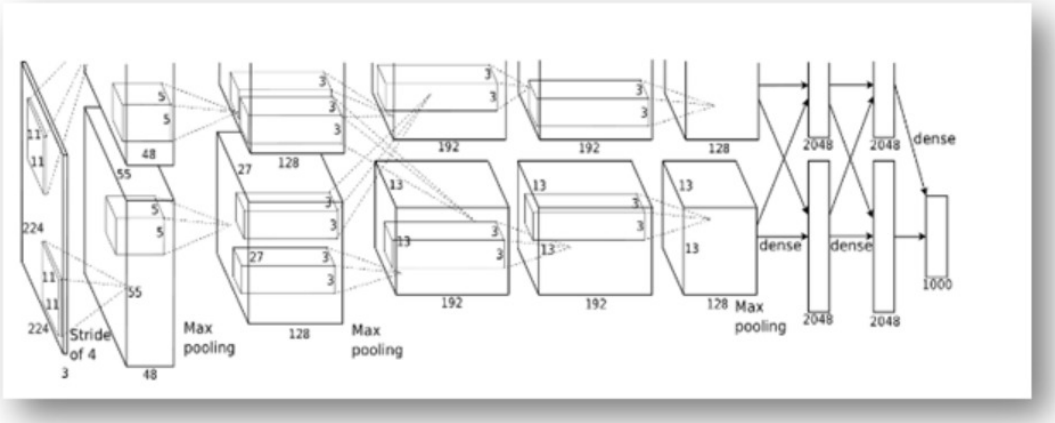


Fig. 4: AlexNet Framework [103] showing how responsibilities are demarcated between two graphic processing units (GPU). The layer at the top is executed by one GPU while the layer at the bottom of the figure is run by another GPU with a network input dimension of 150,528. The communication between the GPUs occurs at a specific layer.

to use randomly initialized layers to learn distinguishing features. VGGNet has been proven to be an effective model in the implementation of a speech emotion classification system [4], [211]. To increase its performance and accuracy, VGG introduced a few designs based on a similar concept (16, 19, and MiniVGGNet). The user has the option of selecting the architecture that best suits their needs. We utilized this technique to propose a robust deep learning speech emotion recognition whereby VGG16 served as the core network in [173].

2.4 ResNet Architecture

A novel architecture called Residual Network (ResNet) was developed by [71], which won the first position at ILSVRC 2015 challenge. In comparison to earlier networks, their objective was to create a mega-network that was devoid of the vanishing gradient problem. Depending on the number of layers, many types of ResNet have been developed (beginning from 34 layers and moving up to 1202 layers). ResNet50 was the most prominent variant of ResNet architecture, with 49 convolutional layers and a few fully connected layers, as depicted in Figure 6. Furthermore, ResNet provided shortcut connectivity within layers to ensure that parameters are free and data is independent of in-depth interconnection. When compared to VGGNet, it possesses a reduced computational complexity, even when the depth is expanded. ResNet has a relative application SEC.

2.5 DenseNet Architecture

To eliminate the difficulty of vanishing gradient, an architecture that is more computationally efficient, called Dense Convolutional Network (DenseNet) was presented in 2017. It was a joint innovation by Facebook AI Research (FAIR), Tsinghua University and Cornwell University. It adopted a feed-forward approach in connecting each layer to all layers in the network. It follows the same concept of an increased network depth as ResNet [76],[212], but with provision to cope with error from back-propagation. Additionally, coping with a large weight is one of the problems associated with ResNet, but DenseNet overcame this challenge by employing an enhanced cross-layer connectivity approach [206]. DenseNet has been widely applied in many speech-related tasks. It is represented in Figure 7 below:

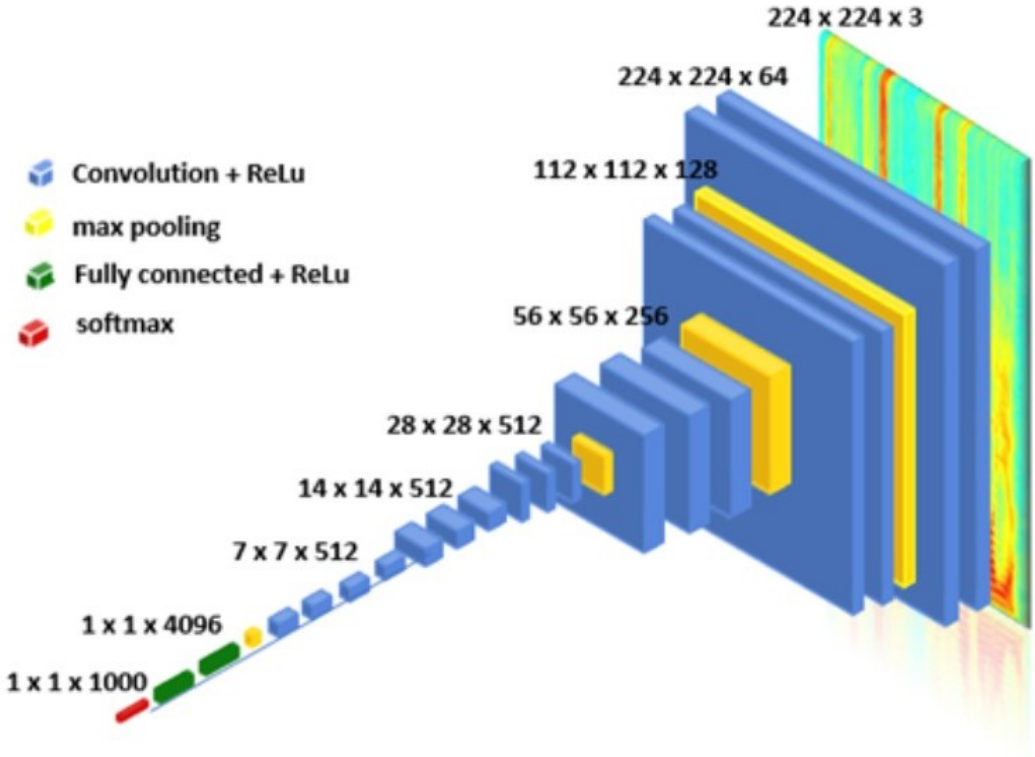


Fig. 5: A simple 5-Conv 2D VggNet Framework with max pooling and ReLu activation function. Standard network input size of 224 X 224 x 3(channel) is accepted

2.6 GoogleNet Architecture

GoogleNet, a 22 layers deep architecture, was proposed by [188] with an input size of 224 × 224. It trained on ImageNet and outdone VGG deep learning model at the SVRC challenge of 2014. It has a model weight of 28MB with network depth reduced by replacing its top fully-connected layers with a global average pooling layer. GoogleNet adopted the concept of micro-architecture (network-in-network) in constructing the micro-architecture [117]. A small building block that efficiently accelerates learning by a network architecture with increasing depth is referred

Fig. 6: ResNet Framework with 3 fully connected layers and one softmax function for multi-class problem

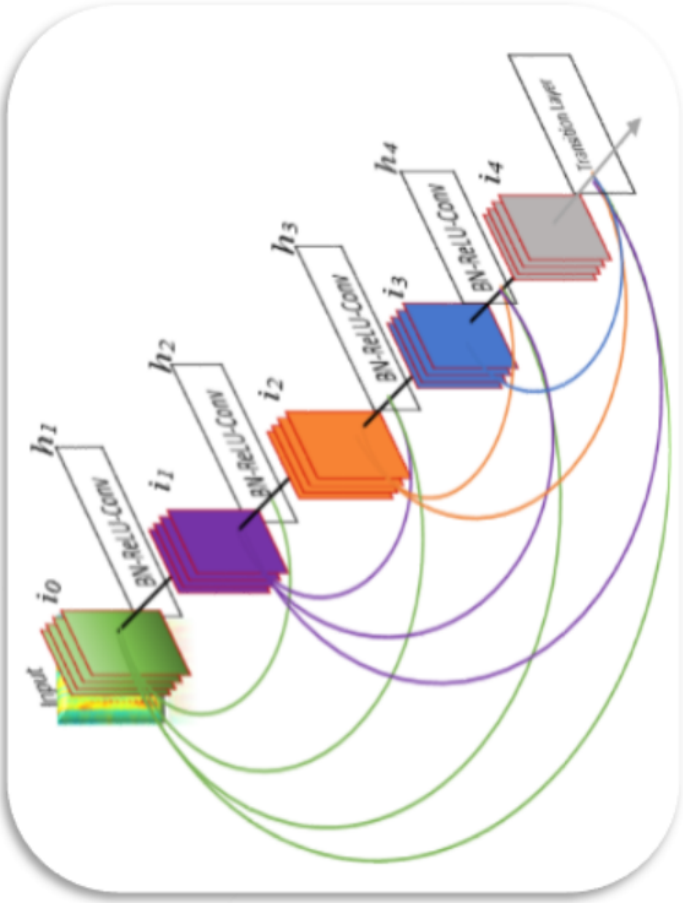


Fig. 7: DenseNet framework comprises of input $i_0 ... i_4$ and Batch Normalization with Reactivation Linear Unit

to as micro-architecture. The target of GoogLeNet was to enhance CNN parameters' effectiveness and learning ability (Dauphin et al., 2015). Additionally, it controls the computation by applying large-size kernels after a bottleneck layer of a 1 x 1 convolutional filter. To overcome the problem of redundant data, sparse connections are utilized in GoogleNet. It saved costs by ignoring useless channels because only some of the input channels are connected to some of the output channels. A major contribution of GoogleNet to the deep learning domain is the concept of the inception module [3]. However, it has a major drawback of mixed topological structure, which requires inter-module adaptation that hampered its rapid use in SEC domain. Figure 8 shows the framework of GoogleNet architecture.

2.7 EfficientNet Architecture

EfficientNet is a powerful DCNN model proposed in 2019 by a team of Google researchers [192] , just as its name suggests. It has variation ranges from EfficientNetV2B0, EfficientNetV2B1, EfficientNetV2B2, EfficientNetV2B3, to EfficientNetV2L with a minimum input size of 20MB and Top-5 accuracy measure of 97.5% [77]. It uses mobile inverted bottleneck convolution (MBConv), while accuracy and efficiency are optimized on floating-point operations per second (FLOPS) basis as shown in Figure 9. The overall goal was to increase accuracy in classification and improve performance through scaling of model, depth, width and resolution balancing. With up to ten times improved efficiency, EfficientNet has reached and surpassed numerous state-of-the-art accuracies in several classification tasks [201]. Furthermore, the main uniqueness of EfficientNet is the most recent feat of intelligent and regulated expansion of a neural network's three dimensions (width, depth, and resolution) using a compound scaling approach [42]. How to increase the dimensions of a neural network in a way that improves accuracy while requiring the fewest number of parameters and making the most of limited resources has proven to be the major challenge over the years of study in the field of deep learning. [95]. These shortcomings have been partly tackled in the EfficientNet deep learning framework.

Due to the vanishing gradients issue, expanding the dimension of a neural network greedily does not produce the expected outputs, even when the minimum operations are not a real target [3][104]. Instead of arbitrarily modifying these dimensions, EfficientNet takes self-introspection into the relationship between the increase in each parameter and applies an exhaustive search under a predefined resource constraint. However, recent findings have shown that EfficientNet performs poorly on hardware accelerators and the huge number of parameters (over 5 million minimum) it possessed may increase computational cost. The overall description of these state-of-the-art deep learning architectures is given in Table 1. The usage of these DCNN architectures for SEC as explored in literature is shown in Figure 10.

3 Deep Transfer Learning (DTL) Architectures

Deep transfer learning architectures are based on the DCNN model that has demonstrated an appreciable learning outcome in many real-life applications. Deep learning usually requires a large data to train with before it can yield an optimum result [5], but deep transfer learning transfers knowledge acquired from a source domain to a target domain [160] [91]. It eliminates the need to train a model (Figure 11) from scratch and thereby reduce time and other computational complexity, especially when label data are not readily available, as it applies to the speech emotion classification domain [208]. DTL has advanced AI in no so measure in the last decades (edge devices) by reducing the computing processing power it will require in carrying out classification and prediction tasks on low-memory devices.

Deep transfer learning differs from other machine learning paradigms such as semi-supervised learning, multitask learning, and multiview learning as it is not mandatory that both the source and target dataset have the same distribution. DTL's focus is on the target domain [80], and source data have already provided the target data with the necessary knowledge, so there is no requirement that they work in tandem or be related. The categorization of DTL is related

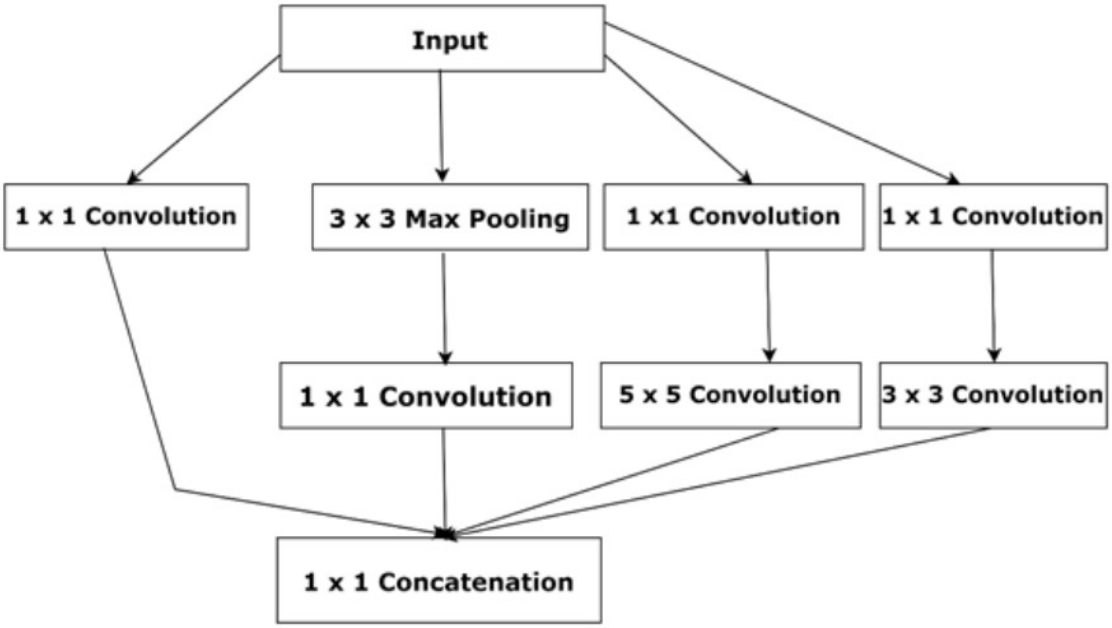


Fig. 8: GoogleNet framework

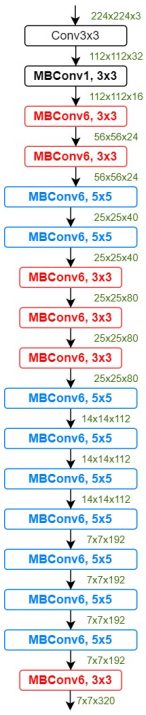


Fig. 9: EfficientNet Structure with huge convolutional layers

Table 1: Summary of state-of-the-art deep learning architectures

Architecture	Brief Description	Layers/Depth	Year	Strength	Weakness
CapsNet [219]	CapsNet consists of interconnections of neurons that can receive and outputs vectors. It easily learns features from images and any possible deformations.	-	2011	It has a robust classification scheme when handling data with inconsistent orientation and size.	Lacks of spatial information and also it is a shallow learner
AlexNet [103]	First breakthrough of CNN architecture trained on over one million ImageNet datasets. It has a simple design of five convolutional layers	107	2012	It is a fast learner architecture with higher depth	Computationally expensive
VGGNet [180]	VGGNet adds more convolutional layers to the network by employing little convolutional filters (3 x 3) while other parameters remain fixed.	16	2015	Highly applicable in deep learning problem	It is slow to train VGG from scratch, thus it requires high computation to evaluate the performance of the model
ResNet [71]	ResNet adopts a stacked convolutional block with skip connection to improve its accuracy and performance. It contains deeper layers than VGGNet.	103	2016	It eliminates the vanishing gradient problem and allows the reuse of features	Highly prone to overfitting because of huge parameters to be trained
GoogleNet [188]	It is a 22 layers deep convolutional neural network architecture, trained on ImageNet, which utilizes a global pooling layer.	22	2015	It trains faster compare to VGGNet.	Computational complexity as a result of large parameter
DenseNet [76]	It is such an architecture that connects each layer with another through a feature map approach to increase performance.	242	2017	It solved the problem of vanishing gradient through a sizeable reduction of parameters	Prone to overfitting and decrease in efficiency as a result of massive connections.
EfficientNet [192]	EfficientNet has a mechanism for better balancing of network depth, weight and resolution.	132	2019	Higher performance and accuracy through the adoption of compound scaling, especially in deep learning applications.	Increased computational cost due to huge parameters (over 20million)

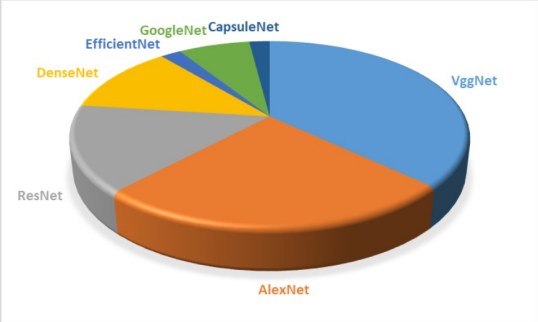


Fig. 10: Chart showing deep learning techniques frequency of usage

to that of transfer learning in terms of problem and solution approach (Figure 12). The problem category can be subdivided into label-dependent or data-dependent only. Transductive, inductive and unsupervised learning fall under the label settings. By description, only the source data is labelled in transductive; but, the source and target data are labelled in inductive; while none of the data is labelled in unsupervised DTL. For Data-dependent, it can either be homogeneous (only image dataset) or heterogeneous(speech and text dataset). Under the solution approach of DTL taxonomy, [190] divided it into four main groups: (i) feature-oriented, (ii) network-oriented,(iii) instance-oriented, and (iv) adversarial-oriented. In instance-oriented, DTL is based on the fact that only selected parts of the main instance are put to use in the source data while adopting a different weight approach to the target data. The feature-oriented performs mapping of a certain set of features(e.g. emotional feature) on both sides(source and target) of the data. The network-oriented approach is interchangeably used with the parameter-oriented which combines knowledge gained from the network model with already trained layers [55]. It operates through the strategies of freezing, fine-tuning, and addition of new layers where necessary. The adversarial-oriented approach implements techniques inspired by generative adversarial networks (GAN).

In literature, it has been found that the network-oriented approach of DTL is the most commonly used approach since it is easier to implement and it quickly adapts both source and target samples through model adjustment. Generally, network-oriented approaches revolve around pre-training, layer freezing, parameter fine-tuning and sometimes the addition of one or more new layers [152]. For instance, with VggNet, the fifth convolutional and top layers can be frozen if a satisfactory speech feature extraction has been actualized from the fourth convolutional layer. The top layers, in this case, may represent fully connected (FC) layers, and can be subjected to replacement with a classifier for the classification of emotion.

4 Speech Emotion Feature Extraction and Selection: An overview

There are several salient features inherent in speech signals. Some features are used to extract linguistic information, and others are used to extract para-linguistic information. The linguistic information represents the language and content of the spoken utterance, while para-linguistic carries emotionally rich information [67]. Similarly, some classifiers perform well with linear data, while others with non-linear data [120]. As a result, providing proper features to be fed into the classifier is critical for implementing an effective and accurate speech emotion classification system [96].

4.1 Speech Emotion Feature Extraction

The quality of feature extraction that is adopted in speech emotion classification will determine the level of accuracy to be recorded. Many feature extraction approach has been used by researchers in classifying emotion. It is accomplished by converting the speech signal (waveform) to a quantitative representation for further processing and analysis at a lower data rate [18][169][231]. Unlike other sub-disciplines in computer vision and image processing, extracting useful features from speech signals is quite challenging, especially for deep learning applications. Traditional and primitive speech features (quality of voice and pitch)[205] can be extracted in a handcrafted manner, but this may not be applicable[155] in deep transfer learning. Majorly, the popularly used feature extraction

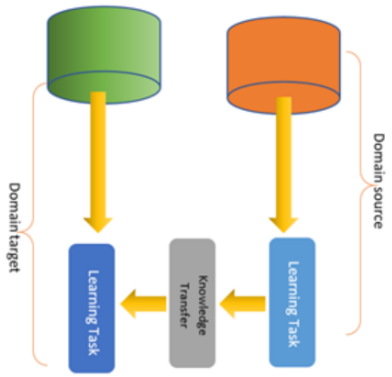


Fig. 11: Structure of Deep transfer learning

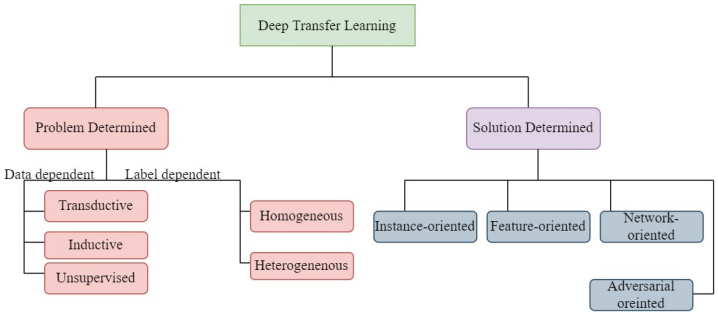


Fig. 12: DTL taxonomy

approach that exists in many research works for the classification of speech emotion is known as acoustics features[127].

Acoustic features can be categorized into spectral, prosodic and voice quality features. Spectral features are commonly used in speech emotion classification. It is based on the conversion of the time-based speech signal to a frequency-based speech spectrum with the Fast Fourier Transform (FFT) approach, as against temporal features (zero-crossing, signal energy, and peak energy), that can be physically interpreted [59]. Mel-frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCCs), and log frequency power coefficients are the most employed spectral features in speech recognition. MFCC is a human hearing system replica that seeks to artificially reproduce the ear's inner workings, with the assumption that a reliable speaker recognition system is the human ear [7]. Windowing the speech signal to disintegrate it into frames is the first step in computing MFCC, after which FFT is used on the frame to locate the power spectrum for every frame. The power spectrum is subsequently processed using a filter bank (mel-scale). After applying the Discrete Cosine Transform (DCT), the MFCC(Figure 13) vector representation is produced. Mathematically, MFCC can be computed using equation 1 [43].

$$\hat{C_n} = \sum_{n=1}^{k} \left(\log \hat{S_k}\right) cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{2}\right] \qquad (1)$$

where $\hat{C_n}$ is the actual MFCC, k denotes the mel-spectrum coefficient and $\hat{S_k}$ denotes the filter bank output. The greater acceptance of MFCC for speech emotion classification is not without its ability to capture information from speech sampled with a peak frequency of 5 kHz, which contains the majority of the energy of sounds produced by humans. It can also detect background

noise that may impair the quality of feature extraction [150].
By simply replicating the vocal tract of a human, the Linear prediction coefficients (LPC) have been providing robust speech features. It uses formant approximation to perform a deep evaluation of the speech signal[34] [75]. LPC is fast becoming an efficient formant estimate method and stronger speech analysis and extraction approach. An offshoot of LPC is Linear prediction cepstral coefficients (LPCC). It can be referred to as the cepstral coefficient that evolves from spectra envelope calculation. In other words, they are the LPC's Fourier transform of the magnitude spectrum [100]. LPCC's ability to represent speech signals in a more perfect manner with a moderate size of features has contributed immensely to their acceptability in many speech processing problems. LPCC is very low in terms of its vulnerability to noise and error rate. It can be computed mathematically using equation 2 below.

$$C_m = a_m + \sum_{k=1}^{m-1} \left[\frac{k}{m}\right] c_k a_{m-k} \tag{2}$$

Where C_m represent the linear prediction cepstral coefficient and a_m is the linear prediction coefficient.

Moreover, in speech feature extraction, the presence of noise has been identified as the major factor that degrades the intelligibility of speech utterances [73]. Therefore, the removal of these noises from raw speech signals while maintaining the emotional content is pertinent. The use of spectral subtraction and MEL filter approaches are the most common techniques for ensuring that the emotional quality of speech signal is retained, even though the background noise is being removed. Mel filter approach is based on Discrete fast Fourier transform to convert the time domain speech signal into the frequency domain. It separates speech signals into a number of components using filter band-pass, with each component of the original speech signal carrying a sub-band frequency. The noise removal technique using spectral subtraction is based on improving the signal-to-noise ratio(SNR) by simple estimation of both the average noise spectrum and signal spectrum and subtracting one from the other. In [134], an enhanced noise removal algorithm was proposed through spectral subtraction and FFT. The noisy speech utterance was segmented into half-overlapped time domain data buffers multiplied by a Hanning window in their spectral subtraction noise removal method, and the output is then transformed into the frequency domain using FFT. The noise is then eliminated by deducting the noisy speech spectrum's average magnitude from it, and the negative values are then zeroed out by using half-wave rectification. Thereafter, an Inverse FFT, was employed to rebuild the noise-reduced speech utterance back into the time domain after eliminating the noise from the noisy speech. The result of their experiment was compared to the original speech signal and it was discovered that noise has been removed. However, in the SEC domain, most publicly available datasets(acted) have been subjected to speech processing, especially, for the removal of noise, but when a non-acted speech dataset is to be used for emotion classification, the removal of background noise will have an effect on the result to be obtained.

4.2 Speech Emotion Feature Selection

To identify speech emotion, it is necessary to choose a small, relevant, and informative set of features. In most cases, there are duplicated features from the speech analysis extraction phase. The correct prediction rates are lowered as a result of these extraneous and redundant features[213]. Using a feature selection approach, the computational load, dimensionality, and the number of features are minimized[131],[59]. The discriminative features for model development are determined using the feature selection technique, to save training time, improve classification performance and minimize overfitting. In real life scenario, the feature selection phase takes input from the feature extraction phase under the deep learning model [89]. Feature selection can be categorized into three major groups: filtered-based (e.g. correlation-based feature selection), wrapper based and intrinsic feature selection[107]. In [204] redundant features were eliminated using contribution analysis feature selection approach with Neural Network(NN). The features selected using this approach was computed using equation 3

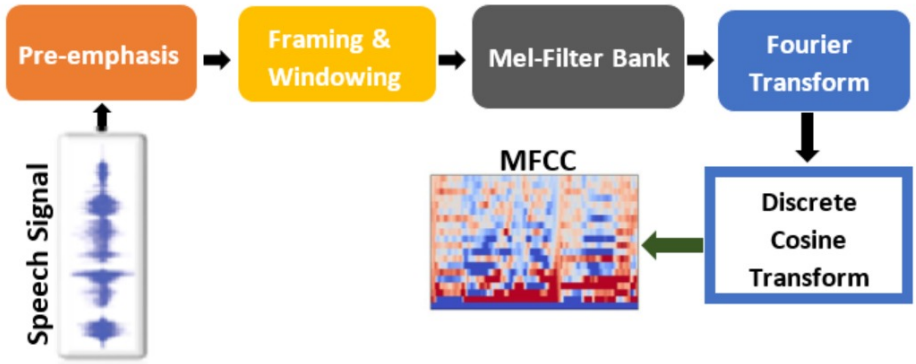


Fig. 13: MFCC feature extraction framework

respectively below.

$$C_k = \sum_{j=1}^{q} u_{jk} v_j, k = 1, ..., p, j = 1, ..., q \quad (3)$$

given that u_{jk} represents the k^{th} contribution from emotional features input to the hidden sublayer j in the neural network, p is the total number of emotional features, while q is the total number of hidden nodes of the layers and v_j denotes the j^{th} output contribution to the neural network. For each emotion class, the set of emotional features selected differs, since the intrinsic feature of each emotion are not the same from speech utterance. Fisher score and Fast Correlation based feature selection (FCBF) were employed in [2] to reduce misclassification rate and improve accuracy of speech emotion classification. Their method proved the significance of feature selection in classification of speech emotion. An accuracy of over 90% was achieved on Berlin Emotional dataset.

5 Overview of Speech Emotion Databases

The success of deep learning in speech emotion classification rests heavily on the availability of speech samples (corpus) that can be used to train the deep learning model. Unlike other machine learning tasks and image processing (e.g. facial emotion), the speech utterance or training dataset requires labelling by hand through a human agent [1] and the mode of perception differs from one person to the other. It is therefore, necessary to have more than one person carrying out this task, to have an accurate label dataset [8][98]. The quality of the dataset is proportional to the result of classification or prediction. In speech emotion classification, three major categories of speech datasets exist: synthetic, semi-natural, and natural datasets of speech samples [109].

The synthetic datasets [149] (RAVDESS, EMO-DB, etc) are the acted speech samples collected from the professional speaker (actors) in a confined environment to extract emotion. They are artificially generated rather than a real-world scene [138][189][216]. The synthetic dataset is the most popular and widely available for speech emotion classification tasks. Semi-natural is close to natural speech dataset, but they have elements of synthesis, e.g. NIMITEK. The last category is natural speech corpora, which capture the real-life scenario of human emotion. They can be gotten from TV shows, call centres, and online videos [207][161][182]. However, there is limited availability of this category of speech datasets because of license issues. It is also prone to environmental noise, which must first be removed before it yields accurate results in emotion classification using deep learning model. Also, it is obvious within the SEC community that, there are a multiplicity of speech emotion dataset but the question of standard measures for these datasets is still lingering to date. Below is a comprehensive description of some of the available speech emotion database and table 2 shows the summary of them all.

5.1 RAVDESS Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a speech corpus released under a creative commons license [163][164]. It has open access for use in research. It captures eight different emotions (angry, sad, disgust, happy, fearful, calm, neutral and surprise) from the speech utterance, of twenty-four (24) professional actors [123]. Twelve of these actors were male and twelve were female as well. Apart from additional neural emotional utterance, each utterance is recorded in two different categories of emotional strength of high or low. It consists of 7356 files, evaluated ten times on emotional authenticity, reliability and validity by mature researchers from North America [137]. Access to it is open for public use. Each file from 7356 total files has a distinct filename (e.g. 02-03-05-01-03-02-10.mp4), that represents the modality, vocal channel, actual emotion, intensity of the emotion and actor's gender (even number for female and odd number for male). RAVDESS [83][85] falls under a synthetic dataset where actors were producing audio speech of different emotional displays.

5.2 IEMOCAP Dataset

The interactive emotional dyadic motion capture database (IEMOCAP) is a speech corpus motivated by the fact that human emotion resides not only in speech but the combination of speech utterance and physiological gesture. The speech samples were recorded by Speech Analysis and Interpretation Laboratory (SAIL) located at Southern California University [37]. Ten actors were recorded in interactional sessions during planned and unscripted verbal communication scenarios with labels on the strategic parts of their bodies (faces and hands), which provided extensive information about their body language and kinesics [54][75]. The actors acted out emotional scripts for over twelve hours while also creating fictitious scenarios meant to elicit particular emotions (happy, furious, sad, frustrated, and neutral mood). This corpus has contributed immensely to the development of multimodal design for emotion classification because of its size, interactivity and holistic elicitation of emotion.

5.3 EMO-DB Dataset

German emotional database (Emo-DB) is a recorded speech utterance (535) that contains 800 files spoken by ten professional actors where five were males and five were females respectively. It was initiated by the ICS, Technical University, Berlin, German [36] [187][183][126]. This corpus captures seven unique human emotions: bore, happy, sad, disgust, fear, angry and neutral. Prior to being resampled (down-sampled) to 16 kHz, the speech sample was originally recorded at a sampling rate of 48 kHz. Each file follows a unique naming pattern like: 02a03Wa.wav which shows that the file is from the second speaker with the emotion "Wut" (Angry)[202][25][221].

5.4 TESS Dataset

Toronto Emotional Speech Set is a publicly available speech emotion database that has been used by many researchers for the classification of emotion. The speech samples were recorded at No 6, Northwestern University Auditory in 2010 [57]. Two actresses were requested to repeat a few hundred words throughout the spontaneous event, and their voices were captured. Seven distinct emotions were captured during the event (happy, angry, fear, disgust, pleasant, surprise, sad and neutral). A total of 2,800 files that depict human emotion were collected. It is audio based only.

5.5 SAVEE Dataset

Surrey Audio-Visual Expressed Emotion database is a unique emotional corpus recorded at CVSSP's 3D vision laboratory, the University of Surrey in 2010. Four actors, who were native English speakers and educationists (student and researchers) whose age ranges from 27 to 31, were involved in the event [69]. The English speakers were labelled as DC, JE, JK and KL. The recording took several months to establish the authenticity of the speech samples and facial gesticulation[113][83]. Altogether, seven emotions were captured by this database which are: anger, disgust, fear, happy, pleasant surprise, sad, and neutral. This corpus has a total size of 480 utterances, with a 44.1 kHz audio sampling rate. However, the database has restricted access, unlike others that are publicly available for research purposes.

5.6 CREMA-D Dataset

Crowd-sourced Emotional Multimodal Actors Dataset (CREMA) is an audio-visual and large dataset (7,442 audio clips) suitable for multimodal human emotion classification [40]. The corpus captures six basic emotional conditions from the facial and spoken utterances of 91 actors, where 48 were males and 43 were females. The corpus is a cross-language comprised of actors from diverse continents (America, Africa, Asia and Europe) to avoid cultural or ethnic barriers that may lead to misclassification of emotion. A total number of 2443 people participated in rating 90 distinct audio clips. It is one of the largest publicly available corpus in use.

5.7 AESDD Dataset

Acted Emotional Speech Dynamic Database (AESDD) is a synthetic and publicly available corpus for the classification of emotion. It is a Greek-based [198] emotional utterance, that captured five basic emotions of anger, disgust, happy, sad and fear. The first version was released in 2018 with an accuracy rate of about 74% by listeners. The level of usage by researchers for this emotional corpus is very low due to solitary language, compared to English based speech emotion corpus.

5.8 eNTERFACE'05 Dataset

eNTERFACE is a German-based audio-visual speech database with open access. It is a multimodal emotion database created for deep learning implementation in emotion classification and human-computer interaction. It was recorded in 2005 by Olivier Martin in collaboration [135] with the eNTERFACE'05 workshop at TCTS Lab, Belgium. The project was fully funded by the Wallon region, Belgium, with contact number EPH3310300R0312/215286. The emotional corpus involved 42 actors (19% female and 81% male) from 14 different countries. A total of 1166 files comprised of six basic emotions were captured. Emotions were elicited by subjecting each actor to an atmosphere or condition where real human emotional outbursts could be captured.

5.9 EMOVO Dataset

An Italian based emotional corpus called EMOVO [49] was developed in 2014 at Fondazione Ugo Bordoni laboratories. It comprised of 6 actors (588 audio files), who synthesized six basic emotions (anger, surprise, joy, fear, disgust and sad) recognized in speech emotion classification research works. The emotions were recorded using standardized equipment that are highly resistant to interference. A total of one hour was used in recording the entire database, with about ten minutes per actor. This audio-based corpus is publicly available for research purposes. EMOVO has contributed greatly to affective computing as it has been applied in Lie detectors, safety management, e-learning and robotics among others.

5.10 MELD Dataset

Multimodal Emotion Lines Dataset (MELD) is one of the recent speech corpora created for emotion classification. In eliciting the emotions, multiple speakers (English speakers) engaged in approximately 1400 conversations and a total of 14,000 utterances were recorded. It is referred to as multimodal [165] because it comprises audio, video and text and is publicly accessible. MELD recordings took another setting as it was recorded from a Television-series programme and was supported by DARPA, National Science and John Templeton Foundation. Emotion classifiers can be trained using MELD and personality indicator systems; however, MELD is not sufficient to train a typical end-to-end conversational system and so it may perform poorly if apply in the deep learning system.

5.11 DES Dataset

The European Union has supported the Danish emotional database (Danish Emotional Speech)[58]. It contains speech recordings from four professional actors, two of whom are male and two of whom are female. The recordings consist of the two isolated words "yes" and "no," nine short sentences, four of which are questions, and two paragraphs. The following five emotional states were imitated in each of the uttered words: neutral, surprise, happiness, sadness, and rage. One of the most popular databases in this sector has been used by scholars. Its applications

include evaluating SVM methodologies both alone and in conjunction with hidden markov models. Additionally, it has been used for cross-corpus validation, unique deep learning techniques [78], and gender-based emotion recognition [101]. As a component of the VAESS project (Voice Attitudes and Emotions in Speech Synthesis), this database was created.

5.12 MSP-PODCAST

A method to efficiently create a sizable, realistic emotional database with well-balanced emotional content is MSP-PODCAST [133]. It depends on already-existing spontaneous recordings that may be found on websites that share audio files and are released under open licenses. With audio pieces ranging from 2.75 to 11 seconds, it offers more than 18,000 natural emotional utterances spread throughout 27 hours from various speakers. The samples were then labelled by a team of over 300 assessors with emotional characteristics (arousal, valence, dominance) and a lengthy list of category emotions. Only labelling based on emotional characteristics is balanced. It has been utilized to evaluate a listener-dependent technique of deep learning models to SER [21].

6 State-of-the-Art Speech Emotion Classification Algorithm

Speech emotion classification (SEC) is a multi-class problem [125] [141]. Various speech emotion classification algorithms that are in use are presented in this section. As depicted in Figure 14, these algorithms fall into four main categories; SEC can be carried out using Traditional Algorithms, Neural networks, Deep Transfer Learning and a hybrid (mixture of two or more algorithms) approach. The description of these algorithms is being presented with the most recommended approach according to our observation.

6.1 Conventional Machine Learning Algorithm

Researchers first used traditional machine learning classifiers to classify these emotions, such as the Support Vector Machine (SVM) [86], [12], [131], the Gaussian Mixture Model (GMM) [47], the k-nearest Neighbour (KNN) [108], and the Hidden Markov Model (HMM) [132]. However, the issue of high noise susceptibility and the inability to effectively handle huge audio speech samples confounds these conventional machine learning classifiers. These algorithms have been used simply for carrying out the classification of speech emotion with a certain level of accuracy recorded. However, several pre-processing and feature engineering needs to be done before an improved result can be obtained with these traditional machine learning algorithms [166]. Slight changes in speech features may lead to a total restructuring of the whole approach. An increase in dataset size has led to poor performance of these conventional algorithms for the classification of speech emotion.

6.2 Neural Network

Improvement in SEC has come through the adoption of the neural network approach. An example includes an Artificial Neural Network that can take several inputs, pass it to a hidden layer and produce output or classification results. Any complex relationship that may exist between the input and expected output can be learned easily with ANN [32]. Long short-term memory (LSTM) is another variant of neural network algorithm that has been widely used for speech emotion classification purposes[87]. These approaches have proven to be a suitable replacement for the conventional algorithm. Though neural network algorithms can easily be implemented, their capability is compromised when it comes to complex problems.

6.2.1 Recurrent Neural Network (RNN)

The nature of speech emotion classification as a sequence data-oriented domain makes RNN an efficient learning model[175]. RNN was proposed in the 1980s as a multi-layer model with memory to remember previous output. The network in RNN follows a set of iterations as shown in Figure 15, which provides means of storing recent information for a specific time, before the actual prediction of output. It uses a backward propagation approach of neural networks [186]. Because of the recursive nature of RNN, the weight of the network is usually updated with possible errors. Therefore, RNN has a major drawback of vanishing gradient problems. This occurs when the

Table 2: Summary of Speech Emotion Database

Speech Corpus	Year	Emotions	Language	Actors	Access	Publication(s)
RAVDESS [123]	2018	angry, sad, disgust, happy, fearful, calm, neutral and surprise	English	24 (12 males and 12 Females) actors	Public	[20][19][62][63][99][102][130][143][177][22][139]
IEMOCAP [38]	2008	happy, angry, sad, frustration, and neutral	English	10 actors	On Request	[30][194] [216][229][161][148][21][72][66][56][35]
EMO-DB [36]	2005	bore, happy, sad, disgust, fear, angry and neural	German	42 actors	Public	[11][145][169][209][53][214][142][162][147][33][9]
TESS[57]	2010	happy, angry, fear, disgust, pleasant, surprise, sad and neutral	English	2 actresses	Public	[167][184][102][63][52][61][92][10][140][184][185]
SAVEE [69]	2010	anger, disgust, fear, happy, pleasant surprise, sad, and neutral	British English	4 actors	Restricted	[44][54][200][130][122][176][214][136][92][139]
CREMA-D [40]	2014	happy, sad, fearful, angry, disgust and neutral	English	91 (48 males and 43 females) actors	Public	[92][90][6][110][193]
AESDD [198]	2018	happy, sad, disgust, anger and fear	Greek	5 actors	Public	[199][217]
eNTERFACE [135]	2006	angry, sad, fear, happy, surprise and disgust	German	42 actors	Public	[97][44][154][203][60]
EMOVO [49]	2014	angry, joy, sad, disgust, fear and surprise	Italian	6 actors	Public	[136][214][157]
MELD [165]	2019	angry, sad, joy, disgust, fear, surprise and neutral	English	Multiple actors	Public	[178][119][41][222][224][179][210][115][105]
DES [82]	Late 90's	rage, sad, happy, surprise and neutral	Danish	4 actors	Proprietary	[78][118][97][197]
MSP-PODCAST [124]	2019	anger, happiness, sadness, disgust, surprised, fear, contempt, neutral and other	Multi-lingua	Multiple actors	Restricted	[21]

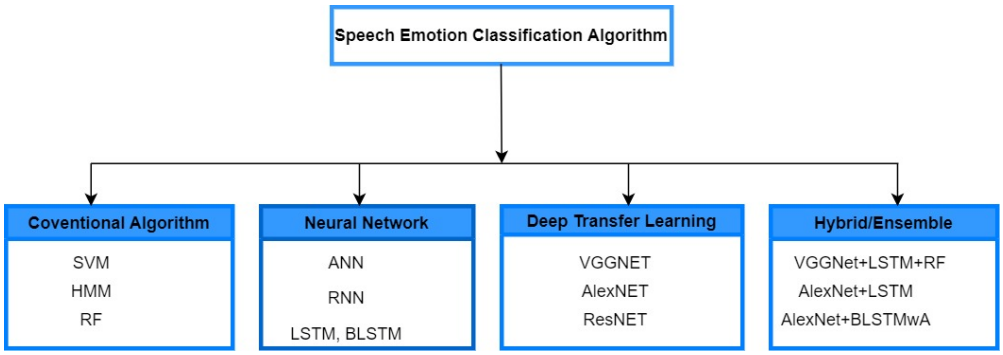


Fig. 14: Categories of Speech Emotion Classification Algorithm

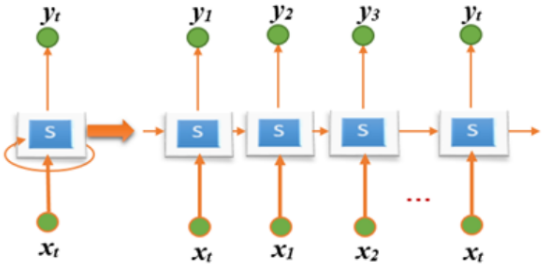


Fig. 15: RNN Framework

network stops learning as a result of the change in output with respect to the change in input. An attempt to solve the major issue associated with RNN resulted in an advanced RNN model called LSTM.

6.2.2 Long-Short-Term-Memory (LSTM)

The long-Short-Term Memory network model was introduced by Hochreiter and Schmidhuber in the late 90s with the sole aim of overcoming the peculiar difficulty with RNN. LSTM incorporated additional memory to learn from long time lag between layers in the network, unlike RNN. It usually incorporates three gates (input, forget and output) [70] which determine whether new input is to be allowed or not, the information is meant to be deleted if it is not essential and finally, allow the information to have a direct impact on the output of the present time step [24]. LSTM has proven to be an effective neural network model in the field of speech emotion classification because of the unique properties of the ability to learn spectral features such as MFCC, LPC and Shifted Delta Cepstral Coefficients (SDCC) from the speech signal. However, LSTM requires huge amount of memory for its effective implementation. Mathematically, suppose that an $X = (x_1, x_2, ..., X_T)$ is given as a finite sequence of input and T denotes the length of the sequence. LSTM can be computed as follows:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1} + b_f) \quad (4)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1} + b_i) \quad (5)$$

$$z_t = \tan h(W_z x_t + U_z h_{t-1} + b_z) \quad (6)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ z_t \quad (7)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t + b_o) \quad (8)$$

$$h_t = o_t \circ \tan h(c_t) \quad (9)$$

where $W_1 \in R^{n\times n}, V_1 \in R^{n\times n}, b_1 R^n, 1 \in \{f, i, z, o\}$ represents weight of the matrices and bias terms respectively. The hyperbolic tangent and sigmoid function are denoted by Tanh and σ, while the Hadamard product is denoted by $\circ$ sign.

6.2.3 Transformer

The transformer model is a recent architecture that is based on an attention mechanism (stack attention layer) for speech sequence data processing in deep learning, with input and output computational representation not relying on the recurrent neural network or convolutional process of sequence data. As a transduction model [196], it has been trained over 4 million sequence data samples (sentences) on 8 NVIDIA P100 GPUs for natural language processing applications. Although, RNN and LSTM in combination with CNN model

have been explored by researchers for efficient performance in speech emotion classification, the introduction of the transformer model is currently a state-of-the-art approach that promises future improvement and excellent performance in speech emotion classification. With the transformer model, the goal has always been to increase the learning dependency of input or output signal positions and reduce the number of operations required in achieving an optimum result with higher accuracy [51]. For speech emotion classification, examples include wav2vec-xlsr [26][13], hubert-base-greek and hubert-large-greek [14].

6.3 DTL Approach

Deep transfer learning is an off-shoot of deep learning, and sometimes they are used interchangeably. Given a typical transfer learning task defined over a set of D_s -source domain, T_s- source Target, D_t- target domain, T_t-target task, and predictive function-$f_T(.)$. The D_s can be expressed as $\{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ annotated samples, and D_t as $\{(x_i^t)\}_{i=1}^{n_t}$ un-annotated samples. DTL occurs when $f_T(.)$ reflects a DNN with $f_T(.)$ as a non-linear function, where $D_s \neq D_t$ [220]. DTL usually consists of several layers of neural networks. They are a pre-trained model that can be used to carry out classification tasks, especially where there is an insufficient dataset, as in the case of the speech emotion domain. The ordinary deep learning model is said to be data hungry (requires a massive dataset).

AlexNet and VggNet have been one of the famous deep transfer learning base models used for speech emotion classification. It is a CNN architecture proposed by Krizhevsky (2012), which was trained on the ImageNet 2010 dataset and has yielded an outstanding result of 1000 object classification. DTL is easy to implement because it does require extra time for training. They can handle nonlinear problems. However, many deep learning algorithms are prone to overfitting and high computational cost because of huge parameters from one layer to the other. As discussed in the previous section, most speech emotion classification DTL model falls under a network-oriented approach in which a pre-trained model is either fine-tuned, frozen (selected layers), or subjected to progressive learning. Progressive in the sense that a new layer can be added to the model, just as continuous learning is associated with human beings until the time of death.

6.4 Hybrid Approach

An innovative approach is the hybridized algorithm, which combines two or more techniques or algorithms to form a robust model. The hybrid classification techniques can also be referred to as ensemble classification techniques. Their strength lies in combining the distinctive characteristics of two or more algorithms for classification purposes. Within the speech emotion classification domain, the different algorithm has been co-join for a hybrid model [195]. An example includes a combination of deep learning algorithms with neural networks and conventional machine learning classifiers. Although the hybrid algorithm has been yielding an improved result in speech emotion classification, the weaknesses of each algorithm that form the hybrid type cannot be eliminated.

7 Speech Emotion Classification Performance Evaluation Metrics

The standard performance evaluation metrics [84] for deep transfer learning models in speech emotion classification include Specificity, Sensitivity, Accuracy, Confusion matrix and Mathews Correlation Coefficient (MCC). Although, MCC is not very pronounced as others, however, it is one of the standardized performance evaluation metrics for the multi-class problem. These metrics are described as follows: *Confusion Matrix:* It is one of the simplest ways to assess a classification algorithm's performance. It summarizes the prediction outcomes for a classification [46] task. Confusion matrix values are used to report the number of correct and incorrect predictions per class. TP-true positive, TN-true negative, FP-false positive and FN-false negative.
Accuracy: The ratio of the correct predictions (TP and TN) to total predictions can be used to measure accuracy as in stated equation 5.

$$Accuracy = \frac{TP + FN}{TP + FN + TN + FP} \tag{10}$$

Sensitivity: This is the proportion of actual positive results to all expected positive results, as given in equation 5.

$$Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

Specificity: It measures the proportion of real negatives to all of the projected negatives, as computed in equation 5.

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

Mathews Correlation Coefficient (MCC): it is a balanced metric that put into consideration, the TP, TN, FN, FP, irrespective of the class size. It stores correlation coefficients ranging from -1 to +1, with +1 denoting a perfect forecast, 0 denoting a mean prediction, and -1 denoting the inverse. It can be computed mathematically as shown in equation 6, where $SN = (TN + FP)(TN + FN)$

$$\text{MCC} = \frac{(TP.TN) - (TP.FN)}{\sqrt{(TP + FP)(TP + FN)SN}} \tag{13}$$

8 A critical review of convolutional neural network and deep learning speech emotion classification studies

In [227], Zang et al., emphasized the importance of feature extraction in speech emotion classification, by developing a robust convolutional neural network model with random forest (CNN-RF) classifier. The features extracted from several convolutional layers of the CNN model were fed into a random forest classifier for emotion classification. A normalized spectrogram image is fed into the CNN input layer before feature extraction was done. The experiment was carried out using a Chinese speech signal record box (NAO). Eventually, the CNN-RF model was tested on the Nao robot, and human emotion through speech signal could easily be figured out by the robot. An overall 75.57% average accuracy was recorded, surpassing conventional CNN model with real-life and efficient human-computer interaction (HCI). A Comparison of the approach in this study with other dataset is expected to be carried out in order to ascertain the performance of the proposed system in this study. Also, CNN alone is not sufficient in extracting global features that carry emotional content.

According to Liu et al., [120], an extreme learning machine (ELM) decision tree with a novel feature selection technique was proposed to alleviate the challenges of redundant features from speech utterance. The feature selection approach used was based on correlation analysis (CA) and the fisher criterion. An experiment for the classification of emotion was carried out with the use of a speech database from Chinese (CASIA), and an 89.6% accuracy report was recorded. By their proposed model, it would be rapid and efficient to distinguish different speakers' emotional states from speech utterances, and it would be able to discover future interactions between robotic machines and speaker-independent. Their method follows a conventional classification approach, and the experiment was carried out on only one speech emotion dataset.

In the work of Gu et al., [65], a more robust and deeper hybrid multimodal framework for the classification of emotions from spoken language (sentence-level) was presented (Gu et al., 2018). Features extracted from text and audio were fused by utilizing three-layer DNN architecture. Fine-tuning of the entire structure was done for the architecture to learn correlation in selecting appropriate features for optimal performance. The framework consists of DNN, CNN-LSTM and ConvNets for extracting both spatial and acoustic features from speech data. The proposed model was evaluated on IEMOCAP public speech dataset and 60.4% weighted accuracy for five emotions was recorded. Also, text and audio features were jointly utilized in their work. However, the model is prone to over-fitting and high computational cost, considering the hybrid nature of the entire framework and the absence of a dimension reduction approach. The addition of feature selection after extraction would probably have increased the accuracy of classification in this study.

Zhang et al., [223] proposed a fully connected CNN architecture with an attention mechanism for speech emotion classification. Their model was based on three hypotheses: emotion from speech utterance is difficult to detect because of the

level of abstraction, the limitation of the labeled dataset for emotion classification and the fact that detection of emotion is time-specific. Their novel model utilized a fully-connected convolutional network for keeping vital information from speech signals without the bottleneck of segmentation. Transfer learning was adopted to eliminate the problem of the limited dataset, and an attention mechanism was used to identify emotion-relevant features from the spectrogram image. A state-of-the-art accuracy was recorded on the IEMOCAP speech corpus with a 70.4% weighted accuracy result, however, accuracy can be improved upon and experiments carried out on other synthetic datasets.

Palo and Mohanty [159] proposed a minimized combined feature for the classification of speech emotion using wavelet decomposition coefficient, linear prediction cepstral coefficient and Mel-frequency cepstral coefficient. The complexity of extracted features was reduced by utilizing the vector quantization approach. The radial basis function network (RBFN) was employed as the classifier for classifying human emotion into five categories: angry, happy, disgust, neutral, and fear. The result and performance of the model were benchmarked on two publicly available speech datasets (EMO-DB and SAVEE). The combined features of WMFC-CVQ+WLPCCVQ were carried out using the MATLAB-R2013a development environment and 93.67%, and 91.82% accuracy was reported on EMO-DB and SAVEE database respectively. Nevertheless, Other emotional classes (sad and surprise) in the dataset were not captured.

Chen et al. [45] utilized three-dimensional attention-based CRNN for selecting discriminating features in the classification of speech emotion. The input layer of their proposed model accepted Mel-spectrogram with delta-deltas. The delta-deltas used, minimized the infiltration of irrelevant features that could lead to poor classification performance while preserving relevant emotional information. Finally, an attention mechanism that could factor in salient features was adopted. The result of their experiment was inspiring, with an accurate report of 82.82% on EMO-DB and 64.74% on the IEMOCAP speech dataset.

Hajarolasvadi and Demirel [68] proposed an enhanced model (3D CNN) for human-computer interaction in speech emotion classification through thorough analysis of speech signals. At first, speech signals were segregated into overlapping frames and Mel-frequency cepstral coefficient, and acoustics features, including an 88-dimensional vector, were extracted. As the spectrogram image was generated, k-means clustering techniques were used on all the frames of the speech signal and the most discriminant k frame was selected. At the final stage, two layers three-dimensional Convolutional Neural Network with 10-fold cross-validation and one FC (fully-connected) layer were applied. The experiment was benchmarked on three popular datasets; Ryerson Multimedia Laboratory (RML), Survey Audio-Visual Expressed Emotion (SAVEE) and eNTERFACE'05. The result shows a state-of-the-art performance over other methods, with an average accuracy report of 81.05%. However, cross-validation tends to increase computational cost.

In averting the challenging nature of automatic emotion recognition and improving human-computer interaction, Ren et al. [170] proposed a novel model that is based on Multi-modal Correlated Network for speech emotion classification. Their sole aim was to draw vital information from the audio-visual channel to accomplish a more accurate emotion prediction. Feature extraction was carried out on both audio and visual signals. Thereafter, the mel-spectrogram image obtained from pre-processing phase was fed into CNN for feature extraction while frames from the visual section were fed into CNN-LSTM for feature extraction. A triplet and correlation loss techniques were utilized for incrementing inter-class and intra-class differentiation reduction, respectively. The experiment was carried out on the AEFW speech database, and the result achieved state-of-the-art speech emotion classification. However, other standard speech emotion datasets were not explored to examine the performance of their model.

Kerkeni et al. [98] designed a global approach for speech emotion classification with the use of empirical mode decomposition (EMD). The efficient analysis of time-frequency on continuous signal attracted the use of Teager-Kaiser

Energy Operator (TKEO) in combination with EMD. Some unique features like modulation spectral, modulation frequency using the AM-FM modulation approach and cepstral features were extracted. At the climax stage, speech emotion classification was carried out using SVM and RNN for seven fundamental human emotions. Berlin and Spanish speech databases were used during the experiment, and the model achieved a 91.16% overall accuracy rate.

Zhao et al. [225] proposed a hybrid model of two CNN-LSTM architectures for learning salient emotional features from the speech signal. 1D and 2D CNN-LSTM were developed to extract emotional contents from the log-mel spectrogram. While the LSTM layer was utilized to extract local features, other techniques of a convolutional layer with max-pooling were used in extracting local correlations. A home-grown feature extraction block was incorporated into the model designed. Their experiment shows state-of-the-art performance compared to ordinary CNN and the deep belief network method. Classification accuracy of 95.33% was recorded on the EMODB database and 89.16% on the IEMOCAP speech corpus respectively. Despite the efficiency of this proposed model, the "black-box" detail of how the network design in classifying emotion was not uncovered. The study leverage the efficiency of LSTM combined with CNN to improve the classification of emotion.

To obtain more accurate performance in model design for classification of speech emotion and lessen redundant features, Jiang et al. [89] proposed a heterogeneous approach that is based on a deep neural network for extracting relevant features from junks of acoustics features. The idea was to bridge the age-long gap between human emotions from speech and acoustics features. A cooperative fusion network was utilized after feature extraction from video and audio datasets, for training the model to learn the distinctive acoustic feature for the classification of emotion. Support Vector Machine (SVM) classifier was used at the topmost layer of the model for the classification of emotion. IEMOCAP speech dataset was used as the benchmark database in experimenting the performance of the proposed DNN model and 64% accuracy was recorded. This study has been able to reduce the level of deterioration associated with emotion classification performance, though, the output result is low, and the experiment was carried out on a single speech corpus. The author established the efficiency of deep learning based on fusion techniques for SEC.

Furthermore, Yao et al. [215], introduced a fusion approach of a convolutional neural network, deep neural network and recurrent neural network (CNN-DNN-RNN) for the classification of four distinct emotions (sad, happy, angry and neutral) from the speech signal. In their framework, three levels of inputs (low-level descriptors, spectrogram, utterance) were fed into CNN, DNN and RNN, at different times. Three models were obtained as a result of this, and outputs were aggregated together. For global features, a multi-task learning approach was adopted across the three models in conjunction with a weighted-pooling attention mechanism for capturing hidden emotional features. To unify the classification of emotional conditions, a fusion approach that is based on confidence ratio was utilized. Eventually, experimentation was carried out using IEMOCAP dataset and the results proved the higher performance of their model as compared to some classifier's-based approaches. The conjoining of three large neural networks as it was proposed in this study can effectively recognize emotion, but the number of parameters that needed to be trained will increase sporadically and this may increase complexity.

To resolve the limitation of the insufficient dataset in speech emotion classification, Zhou and Beigi [228] proposed a transfer learning TDNN (Time-Delay Neural Network) model. Two significant features were used in their architecture, which was MFCC and identity-vector. With transfer learning, it is needless to train a model from scratch and so, it saves time. The initialization process of the transfer learning was done by utilizing the Ted-Lium speech corpus before the actual experiment was carried out using one of the popular synthetic datasets (IEMOCAP). With 5-fold cross-validation, the accuracy result showed improved performance over many state-of-the-art approaches. However, their model has a limitation of the inability to classify emotion at the frame level.

Anvarjon et al. [23], were motivated to propose a lightweight CNN speech emotion classification model to improve the accuracy rate, irrespective of the language barrier. In their approach, the deeper

features were learned to use a convolutional neural network. An enhanced pooling strategy with a filter was used for the extraction of frequency-based features. The simplified CNN model utilized 1 softmax layer (classifier), two batch normalizations, and 3 pooling layers with 8 convolutions. The experiment was carried out using two popular speech databases which are the Berlin emotional speech database and interactive emotional dyadic motion capture respectively. In the end, 77.01% on IEMOCAP was recorded, while 92.02% accuracy was recorded on EMO-DB. The author of this study improved performance in the classification of emotion.

Mustaqeem et al. [146], proposed a deep learning architecture for speech emotion classification that is based on clustering measurement of radial basis function (RBF) and CNN. A Short-term fast Fourier transform (STFT) mechanism was used in converting the speech signal to spectrogram image equivalence. Thereafter, relevant and distinctive features were extracted by feeding the CNN model with a spectrogram generated through a special Librosa (audio processing) tool. Normalization of features extracted from CNN was done before the output was fed into bi-directional long short-term memory (BiLSTM) to reduce complexity. Their model fetched time-space information and produce an accurate classification of emotional state. An experiment was carried out on three widely used speech corpus; RAVDESS, EMO-DB and IEMOCAP respectively. The result produces a state-of-the-art accuracy of 77.02%, 85.57% and 72.25% over the three corpora respectively, however, RBF tends to increase weight of every feature attribute which may invariably lead to an increase in training time.

Furthermore, as effort in research toward improvement in classification accuracy is progressing, Ho et al. [74] proposed another fusion approach of deep learning, RNN and multi-level attention mechanism for emotion classification from auditory speech and text. In their method, extraction of mel-frequency cepstrum (MFCC) from speech signal was carried out using the OpenSMILE toolkit, while transfer learning (pre-trained model with BERT) technique was utilized in encoding text data. Multi-head attention method was used to fuse the feature representation for accurate classification of emotional conditions. An experiment was carried out using three databases which are: IEMOCAP, MELD and CMU-MOSEI. The result of the performance evaluation proved that joint modalities yield a better classification accuracy than one model.

Moreover, as deep transfer learning for speech emotion classification was gaining ascension, Padi et al. [158], proposed anlearnedved approach for emotion recognition by exploring the innate advantage that resides in the pre-trained model. Their work utilized ResNet as a pre-trained model which has been trained on speaker recognition dataset, in conjunction with a specialized pooling layer (statistics). The pooling layer prevents complexity by systematically resizing input. To eliminate the over-fitting of the model and increase classification performance, multiple spectrogram (data) samples were generated through data augmentation techniques. The experiment was carried out on IEMOCAP corpus, and the result indicated a state-of-the-art accuracy. However, implementation of speech emotion classification using ResNet requires heavy computing graphical processing unit (GPU) resources as a result of huge parameters. A deep convolutional neural network with an attention mechanism was proposed by H. Zhang et al., [218] to improve speech emotion classification. Data augmentation and dataset balancing were adopted in pre-processing speech samples. To create segment-level features, their model was pre-trained on ImageNet. To predict human emotion, the Deep Neural Network (DNN) was eventually fed with learned high-level emotional variables. The proposed model's performance was tested on two popular speech databases as a result of the performance analysis (EMO-DB and IEMOCAP). Sensitivity and Specificity received average scores of 87.86% and 68.50%, respectively. The work achieves excellent recognition results; however, the impact of the dataset augmentation on the efficiency of the proposed model in this study was not stated in terms of the eaccuracy of classification. Model implementation parameters were not revealed.

As deep learning is advancing the scope of human-machine interaction through the more accurate model for emotional state identification deepens which made, An and Ruan, [17] proposed a robust, twin and parallel convolutional neural network (CNN), to eradicate the low accuracy rate of other methods. Data augmentation was carried

out by adopting Gaussian white noise (AWGN) techniques, which eventually multiplies the speech samples from 1440 to approximately 6000. Both time-based and frequency-based features were extracted using CNN and transformer encoder. The architecture was trained for 200 epochs before classification into 8 emotions. The experiment was carried out using RAVDESS speech database on Pytorch environment with NVIDIA Tesla V100. An accuracy of 80.46% was recorded after classifying emotions into eight different categories. Although CNN has been proven to be stronger in spatial feature extraction and signal encoding, the model did not make provision for dimensionality reduction, which may be responsible for the increase in computational cost.

Zhao et al. [226] proposed a three tiers deep learning approach for improving emotion classification from speech data. In their model, CNN characteristics in context-based information modeling were exploited to improve the performance of the proposed architecture. Spectrogram in conjunction with delta-deltas was used at the input layer. A specialized convolutional layer arranged in a parallel form combined with a high network representational technique known as Squeeze-and-Excitation Network (Senet) was utilized for the extraction of features from three-dimension spectrograms. Additionally, a reduction in a large disparity between the input and output informed their adoption of connectionist temporal classification (CTC) with the self-attention dilated residual network (SADRN) technique. The overall experiment was benchmarked on IEMOCAP and FAU-Aibo emotion corpora. A weighted accuracy of 73.1% and unweighted accuracy of 66.3% were recorded, which indicated the suitability of the proposed model for discrete speech emotion classification.

Atila and Şengür [28], came up with another innovative deep learning approach for speech emotion classification. They utilized a three-dimension convolutional neural network combined with LSTM and attention in their model. An end-to-end training mode was used in their proposed model. To increase performance and reduce inaccurate prediction, they carried out data resampling, with much focus on high frequencies. Thereafter, images obtained from the conversion of the audio signal to spectrogram and cochleagram served as input to the twenty-eight layered 3D CNN-LSTM architecture. The architecture comprised of convolutional layers with attention, LSTM, flatten, dropout and 2 fully connected layers. The experiment was conducted using three major datasets; Audio-Visual Database of Emotional Speech (RAVDESS), RML, SAVEE and a combination of the three datasets. An evaluation was done using four metrics which are accuracy, specificity, sensitivity, and recall (86%, 93%, 96% and 99%) respectively. The result obtained outperformed most existing approaches, however, their model requires enormous training time as a result of multiple attention and LSTM weights and huge convolutional layers.

A novel deep learning technique was put forth by Chimthankar [48] that applies CNN and LSTM to MFCC features taken from four well-known speech datasets (TESS, RAVDESS, SAVEE, and CREMA-D). On a German-based (audio samples) independent dataset that was not added to the model during training, their model achieved promising result 67.58% validation accuracy and 71.28% highest testing accuracy. However, there is still room for improvement in the recorded experimental results' performance due to the CNN model's unique computational complexity.

Van et al. [195], proposed deep learning that combined convolution neural network, convolution recurrent neural network and gated recurrent unit for speech emotion classification. To increase the classification rate, data augmentation through voice manipulation was carried out. Mel-spectral coefficients features and other features that carried paralinguistic information for emotion classification were extracted in their model. The gated recurrent unit (GRU) approach eliminates the difficulty of vanishing gradient associated with RNN, by updating and resetting the gate. The CNN consists of 5 2D-convolution layers, while double GRU with other primitive layers were used in their architecture. An Interactive Emotional Dyadic Motion Capture (IEMOCAP) speech dataset was used for their experiment, with 97.4% accuracy from the GRU model with shows an outstanding performance compared to other research that has been carried out using the IEMOCAP dataset. However, their data augmentation through voice change and the addition of white noise tends to consume memory space and increase computation

cost as a result of large convolutional layer. IEMOCAP is a long speech dataset, definitely addition of white noise will have some impact, therefore comparison between pre- and post-augmentation of the dataset is expected to be reflected in the study. Training time is another trade-off.

Puri et al. [168] focused on one of the application areas (customer care) of speech emotion classification (SEC) in developing a hybrid convolutional neural network model for SEC. Two major features that comprised mel-spectrogram and MFCC were extracted from the speech signal. Emotion classification was partitioned into three sub-groups which are positive, negative and specific (sad, angry, happy, fearful, disgust, calm and neutral) emotional states. Eight layers of a 2D convolutional neural network model was proposed. An experiment was conducted to evaluate the performance of the proposed model using the RAVDESS speech corpus. The author reported that the model yielded better performance, but no accuracy report or any other evaluation metric to substantiate the performance of their proposed model was stated in the study. The efficiency of the proposed model in this work cannot be established because of missing performance evaluation.

Recently, the authors in Aggarwal et al. [4], proposed a deep learning approach that is based on a two-way feature extraction method for effective speech emotion classification. The DNN is comprised of a dense layer and a dropout layer. The prominent role of feature extraction in emotion recognition from speech signals motivated the authors to first adopt principal component analysis (PCA) and later utilized DNN for extracting the Mel spectrogram image. The spectrogram image served as input to the pre-trained CNN model (VGG16) before the classification of emotional states. RAVDESS and TESS speech datasets were used during the experiment and the proposed model was able to classify emotion into 8 different categories, with an overall accuracy of 81.94% and 97.15% respectively. The result shows that the proposed model outperforms other studies.

In Singh and Prasad. [181], a CNN model for SER that is gender-dependent was proposed. The author updated the RAVDESS speech dataset by capturing two different emotional intensities (normal and strong) on six emotions (sad, fear, furious, pleased, disgusted, and calm). They included emotional elements like Chromogram, Mel-Spectrogram, MFCC, and Spectral contrast in their investigation, and an in-depth performance comparison revealed a relative improvement over the baseline system. Muhammad et al. [144] a well-executed preprocessing strategy and a mixed model made up of LSTM and CNN for emotion recognition from human speech. A high pass and Savitzky Golay Filter were employed to obtain a noise-free audio signal after about 2800 audio files were retrieved from the TESS database. Their model successfully classifies emotions with an accuracy of 97.5%. However, since only one speech dataset was utilized, it is impossible to determine whether this study is generalizable. Increasing feature representation to enhance speech emotion classification propelled Nasir et al. [151] to come up with a novel deep learning framework with a fusion of two significant features(temporal and spectral) for improving emotion recognition from the speech signal. They utilized a layered-based CNN transformer in their fusion techniques, which resulted in low computational complexity and reduction in feature map. Their model was experimented on two datasets comprised of EMODB and IEMOCAP, and they achieved a state-of-the-art 94.2% and 81.1% accuracy on both datasets respectively. Ayush et al. [93] proposed an ensemble deep learning techniques of CNN, attention mechanism, LSTM and Vision Transformer for speech emotion classification. They achieved a promising result of 85.3% accuracy of classification on EMODB dataset. However, their method is computationally expensive and only one dataset used, is not sufficient for generalizing the result.

Researchers have expressed grave concern over the variety of human speech, which makes it challenging to develop a single, standardized method for uncovering concealed emotions. By integrating a multilingual emotional dataset with the development of a more comprehensive and successful deep learning model for categorizing human emotions, Waleed [15] made an effort to address this research challenge. The model was made using a two-step technique. The first step in the classification process was the extraction of features. The

Table 3: Summary of state-of-the-art deep learning techniques: **A** – Angry, **H** - Happy, **S** - Sad, **D** - Disgust, **N** - Neutral, **F** - Fear, **B** – Boredom, **Sr** - Surprise, **J** - Joy

Publication	Proposed Techniques	Dataset	Emotions	Accuracy Reported
[227]	CNN-RF	CASIA	A, H, J, S	75.57%
[120]	ELM-CA (fisher criterion)	CASIA	A, F, H, N, S, Sr	89.6%
[65]	DNN-CNN-LSTM	IEMOCAP	A, H, N, S	60.4%
[223]	CNN-Attention	IEMOCAP	A, H, N, S	70.4%
[159]	WMFCCVQ-RBFN	EMO-DB, SAVEE	A, B, D, F, H, N, S	93.6%, 91.82%
[45]	3D-Attention-CRNN	EMO-DB, IEMOCAP	A, B, D, F, H, N, S	82.82%, 64.74%
[68]	3D-CNN	RML, SAVEE, eNTERFACE'05	A, B, D, F, H, N, S	81.05%, - , -
[170]	CNN-LSTM	AEFW(AudioVisual)	A, F, H, D, Sr, S, N	60.59%
[98]	EMD-TKEO-RNN	EMO-DB	A, B, D, F, H, N, S	91.16%
[225]	1D, 2DCNN-LSTM	EMO-DB, IEMOCAP	A, B, D, F, H, N, S	95.33%, 89.16%
[89]	DNN-SVM	IEMOCAP	A, H, N, S	75.57%
[215]	CNN-DNN-RNN	IEMOCAP	A, H, N, S	58.3%
[228]	TDNN-MFCC	IEMOCAP	A, H, N, S	71.7%
[23]	CNN	IEMOCAP, EMO-DB	A, B, D, F, H, N, S	77.01%, 92.0%
[146]	CNN-RBF-BiLSTM	RAVDESS, EMO-DB, IEMOCAP	A, B, D, F, H, N, S, Sr	77.02%, 85.57%, 72.25%
[74]	RNN-MLA(Multi-level attention)	IEMOCAP, MELD, CMU-MOSEI	A, B, D, F, J, N, S, Sr	60.71%, 63.26%, 99.19%
[17]	CNN-AWGN	RAVDESS	A, D, F, H, N, S, Sr	80.46%
[226]	CNN-SADRN	IEMOCAP	A, H, N, S	73.10%
[28]	3D-LSTMwA	RAVDESS, RML, SAVEE	A, D, F, H, N, S, Sr	86.0%
[218]	DNN	EMODB, IEMOCAP	A, B, D, F, H, N, S	87%, 86%, 68.5%
[158]	DNN-ResNet	IEMOCAP	A, H, N, S	66.02%
[195]	CNN-RNN-GRU	IEMOCAP	A, H, N, S	97.4%
[168]	2D-CNN	RAVDESS	A, D, F, H, N, S, Sr	98%
[4]	DNN-PCA	RAVDESS, TESS	A, B, D, F, H, N, S, Sr	81.94%, 97.15%
[48]	2D CNN+LSTM	EMODB	A, B, D, F, H, N, S, Sr	71.28%
[144]	2D CNN+LSTM	TESS	A, D, F, H, N, S, Sr	97.5%
[181]	2D CNN+MFCC	RAVDESS	A,B, D, F, H, N, S, Sr	72%
[151]	DeepCNN+Transfomer	EMODB, IEMOCAP	A,B, D, F, H, N, S,	94%, 81.1%
[93]	Vision Transfomer(ViT)	EMODB	A,B, D, F, H, N, S,	85.36%

well-known MFC coefficients, RMSE, and ZCR coefficients were retrieved as features to be input into two proposed models for classification: 1D CNN+LSTM and attention+2D CNN architecture. The results showed that the model achieved accuracy of 96.72%, 97.13%, 96.72%, and 88.39%, respectively, on the EMO-DB, SAVEE, ANAD, and BAVED datasets.

The summary of these proposed studies is presented in table 3.

However, in this work, a critical study of selected articles on SEC using a deep learning model revealed that the DCNN requires millions of datasets for its training which unfortunately are not available in SEC domain. As a result of the limited dataset, results of emotion classification based on deep learning models cannot be totally insulated from overfitting. The combination of LSTM for learning spectral features and DCNN has proven to yield increased performance in many SEC systems, but the requirement of high computing resources and complexity of computation is also a drawback.

9 Discussion

Speech emotion classification can be termed as a precise indication of the emotional state of an individual, from the speech utterances. As shown in Figure. 10, the present techniques for speech emotion classification fall under four specific categories which are: conventional, Neural network, deep transfer learning and hybrid methods. Any one of these can be employed for emotion classification from speech but the choice must be strictly guided by useful parameters and relevant features extraction of interest from the speech signal. For deep learning, it is obvious from several research which has been carried out, that no single feature learning approach is the most suited for speech emotion classification, considering a lot of trade-offs. But, a possible combination of different approaches has yielded a good performance, against single convolutional neural network model.

The major limitation to SEC has been the non-availability of a sufficient (large) speech emotion dataset, with which a deep learning model can be trained. Although, deep transfer learning seems to overcome this challenge, by employing an existing pre-trained model, however, majority of these model were trained using the dataset (ImageNet) that are not speech based. Of course, several speech emotion corpora have been made available, but a larger percentage of them are synthetic(acted) in nature and this also poses another challenge to the result of the speech emotion classification model that has been experimented so far. Real live speech emotion corpus from call centres or television shows would have produced the best performance, but they are difficult to access because of license issues. However, the need for standardization of available speech datasets is paramount in order to attain the utmost height in affective computing.

One of the most essential stages in the SEC system is classification, which depends on the classifier's ability to correctly comprehend the output from one layer of the model to the other. There are many difficulties with classifiers, such as the deep learning classifier. Max-pooling makes DCNN much slower, which increases the training process duration. To handle the larger datasets, conventional classifiers like GMM, Random Forest, Decision Tree, and SVM require more time in feature mapping.

The challenge of exactly describing the meaning of emotions is one of the fundamental flaws. Emotions are frequently muddled and difficult to understand. The collection of speech emotion corpora shows that there is no consensus over what constitutes an emotion. However, if we take into account how people and computers interact on a daily basis, we might realize that emotions are spontaneous. As opposed to being more stereotypical traits, these deviations could be buried and hardly distinguishable.

Furthermore, it has been observed as well that emotion from speech utterances is not static in most real live scenarios and most of the existing speech dataset has not been able to capture the continuous nature of emotion from speech utterances. Although deep learning is giving SEC domain a wider view and acceptability beyond the psychological corridor, the proportion of speech features that contributes mostly to the efficient classification of emotion is yet untapped. It is also pertinent to develop speech-based feature selection techniques that can discriminatively identify emotionally rich features to enhance classification. The present framework in SEC has failed to recognize this pertinent issue.
Component wisely, cultural background, accent

and speech emotion are inseparable, although HCI is not peculiar to a particular race. The SEC domain must as a matter of urgency takes a closer look at how deep learning can improve emotion classification on a cross-corpus, language and cultural basis. The best-suited deep learning techniques that can address this issue in terms of local and global speech feature extraction and learning need utmost priority. Also, an explainable and interpretable deep learning model will address many biases in language, as against the black box approach of deep learning

Additionally, most deep learning models have pre-defined input size. These sizes in most cases do not fit naturally with speech features representation (MFCC and log-mel spectrogram) as extracted from raw speech utterances. This backdrop has necessitated scaling and resizing of speech features to match the deep learning input specification before feeding them into the neural network convolutional layers. However, the impact of this disparity has often been overlooked, and it possesses an innate tendency to reduce the performance of the deep learning model and classification accuracy for speech emotion classification.

10 Challenges and Future Directions

Biased models result from the lack of diversity in languages, cultural backgrounds, and emotional expressions found in existing emotion datasets. Creating more representative and diversified datasets that span a variety of linguistic and cultural variations is one way to address this issue. Effective collaboration with various communities to guarantee inclusion when creating datasets will also be of great benefit. It can also be advantageous to apply transfer learning from pre-trained models on various datasets.

Since emotions vary depending on the context, the same auditory characteristics may convey a different emotion under different circumstances. Emotional expressions with ambiguities make proper classification difficult. This problem can be circumvented by including contextual information in the model, such as dialogue history or situational background. Investigate multi-modal strategies to capture more cues by fusing audio with visual data. Emotional expressions can be better understood when using context-aware models.

In applications such as human-computer interaction, real-time processing of speech signals for emotion recognition is essential, but deep learning models with heavy parameter tuning can be computationally costly, resulting in latency. To lessen the computational complexity, model quantization techniques can be applied, and lightweight architectures, or optimize model architectures for efficiency can be of immense benefit too. To improve real-time performance, a closer look into hardware acceleration possibilities like edge computing solutions or specialized processors like GPUs or TPUs can be of great advantage.

Subjective emotion annotation causes diversity in datasets that have been labelled. Furthermore, the generalization of a model may be impacted by individual variations in how they express and perceive emotions. Increasing annotation consistency by providing annotators with training and explicit rules can mitigate this challenge. Additionally, exploring personalized models that accommodate many modalities of expression and the utilization of transfer learning techniques to make use of insights from a variety of datasets and extrapolate to new domains can render future solutions to the problem of subjectivity.

11 Conclusion

In this study, we have carried out a comprehensive review of state-of-the-art methodologies and techniques for speech emotion classification. Various algorithms, deep learning architectures and models used for speech emotion classification have been presented. A critical survey of different speech emotion databases that are readily available for accurate classification has been studied as well. Feature extraction and selection play a key role in speech emotion classification, different feature extraction techniques that have yielded an improved performance without misclassification have been explored, including mechanisms for speech signal noise impairment removal.

The performance of the deep learning based models which include: ResNet, EfficientNet, AlexNet, GoogleNet, DenseNet and VGGNet, on

the speech databases with the use of standardized evaluation metrics has been reviewed. Some of these deep learning models require fine-tuning and sometimes freezing of layers before they can yield appreciable results in speech emotion classification. It was observed from most of the reviewed works that the standard evaluation metrics like accuracy, specificity, and confusion matrix have shown unambiguous results and therefore, we recommended them for more performance evaluation in enhancing speech emotion classification from the speech signal.

From several speech databases surveyed in this paper, it is observed that the ones with short utterances require lesser speech processing and analysis in terms of widowing, framing and feature extraction with deep learning model. It is less time-consuming. Therefore, we hope to carry out an experiment using deep learning techniques in combination with a dimensionality reduction and feature selection algorithm on a number of them such as TESS, EMOVO, ENTERFACE, MSPODCAST, etc, to improve the classification of speech emotion.

As critical analysis of available methods and techniques gives a better opportunity for the best choice in system design and implementation, just as we have reviewed many publications on SEC, this paper can serve as a good precept in selecting the best performing deep learning approach for speech emotion classification with an improved state-of-the-art result. After discussing the aforementioned observation, we can draw the conclusion that deep learning is shaping the future of speech emotion classification in the right direction, even with a meagre speech dataset. However, more speech features that are rich in emotional content need to be examined in order to enhance deep learning model performance in SEC. The challenges highlighted in this paper demand holistic consideration in future research, if SEC and affective computing, will witness tremendous improvement in the next decade.

Declarations

Ethics approval and consent to participate

Yes, consent is granted.

Consent for publication

Yes, consent is granted for publication.

Availability of data and materials

The datasets generated and/or analyzed during the current study are open-access and publicly available.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

Both authors, Samson Adebisi Akinpelu1 and Serestina Viriri, contributed equally to the manuscript.

Acknowledgements

Not applicable.

References

[1] Abbaschian, B., Sierra-Sosa, D., and Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21:74–80.

[2] Abhishek, K. A. G. (2022). Speech emotion recognition by using feature selection and extraction. In *22nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), IEEE, 978-1-6654-9710.*

[3] Agbo-Ajala, O. J. and Viriri, S. (2021). Deep learning approach for facial age classification: a survey of the state-of-the-art. *Artificial Intelligence Review. Springer*, 54(1).

[4] Aggarwal, A., Srivastava, A., Agarwal, A.and Chahal, N., Singh, D., Alnuaim, A., Alhadlaq, A., and Lee, H. (2022). Two-way feature extraction for speech emotion recognition using deep learning. *Artificial Intelligence Review. Springer*, pages 1–11.

[5] ahia Cherif, R., Moussaoui, A., Frahta, N., and Berrimi, M. (2021). Effective speech emotion recognition using deep learning approaches for algerian dialect. In *In:2021 International Conference of Women in Data Science at Taif University, WiDSTaif*, volume 1-6.

[6] Ahmed, M. R., Salekul Islam, A., Islam, M., and Shatabda, S. (2023). An ensemble 1d-cnn-lstm-gru model with data augmentation for speech emotion recognition. *Expert Systems with Applications*, 218:119633.

[7] Ajibola Alim, S. and Khair Alang Rashid, N. (2018). Some commonly used speech feature extraction algorithms. from natural to artificial intelligence. *Algorithms and Applications*, 23(1386):3–20.

[8] Akcay, M. and Oguz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. In *In Speech Communication*, pages 56–76.

[9] Akinpelu, S. and Viriri, S. (2023a). A robust deep transfer learning model for accurate speech emotion classification. In *In International Symposium on Visual Computing*.

[10] Akinpelu, S. and Viriri, S. (2023b). Speech emotion classification using attention based network and regularized feature selection. *Scientific Reports*, 13(1):11990.

[11] Akinpelu, S., Viriri, S., and Adegun, A. (2023). Lightweight deep learning framework for speech emotion recognition. *IEEE Access*, 11:77086–77098.

[12] Al Dujaili, M. J., Ebrahimi-Moghadam, A., and Fatlawi, A. (2021). Speech emotion recognition based on svm and knn classifications fusion. *Int. J. Electr. Comput. Eng (IJECE)*, 11:1259–1264.

[13] Alexei, B., Yuhao, Z., Abdelrahman, M., and Michael, A. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

[14] Alexei, B., Yuhao, Z., Abdelrahman, M., and Michael, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hiddenunits. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

[15] Alsabhan, W. (2023). Human–computer interaction with a real-time speech emotion recognition with ensembling techniques 1d convolution neural network and attention. *Sensors*, 23:1386.

[16] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaria, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. (2021). Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8,1.

[17] An, X. D. and Ruan, Z. (2023). Speech emotion recognition algorithm based on deep learning algorithm fusion of temporal and spatial features. In *Journal of Physics: Conference Series*, volume 1861(1).

[18] Ancilin, J. and Milton, A. (2021). Improved speech emotion recognition with mel frequency magnitude coefficient. *Applied Acoustics*, 179.

[19] Andayani, F., Theng, L. B., Tsun, M., and Chua, C. (2022a). Hybrid lstm-transformer model for emotion recognition from speech audio files. *Access*, 10.

[20] Andayani, F.and Theng, L. B., Tsun, M., and Chua, C. (2022b). Recognition of emotion in speech- related audio files with lstm-transformer. In *in: 2022 5th International Conference on Computing and Informatics (ICCI)*, page 087–091.

[21] Ando, A., Mori, T., Kobashikawa, S., and Toda, T. (2021). Speech emotion recognition based on listener-dependent emotion perception models. *APSIPA Transactions on Signal and Information Processing*, 10.

[22] Andry, C., Irene Anindaputri, I., and A., E. W. (2023). Exploring deep learning algorithm to model emotions recognition from

speech. In *In 7th International Conference on Computer Science and Computational Intelligence 2022, Procedia Computer Science*, pages 706–713.

[23] Anvarjon, T. and Mustaqeem Kwon, S. (2020). Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. *Sensors (Switzerland)*, 18:1–16.

[24] Apaydin, H., Feizi, H., Sattari, M., Colak, M., Shamshirband, S., and Chau, K. (2020). Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting. *Water (Switzerland)*, 12:1–18.

[25] Arias, J. P., Busso, C., and Becerra, N. (2014). Shape-based modeling of the fundamental frequency contour for emotion detection in speech. *Computer Speech and Language*, 28:278–294.

[26] Arun, B., Changhan, W., Andros, T., Kushal, L., Qiantong, X., Naman, G., Kritika, S., Patrick, P., Yatharth, S., Juan, P., Alexei, B., Alexis, C., and Michael, A. (2021). Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv:2111.09296*.

[27] Asifullah, K., Anabia, S., Umme, Z., and Aqsa Saeed, Q. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53:5455–5516.

[28] Atila, O. and Sengur, A. (2021). Attention guided 3d cnn–lstm model for accurate speech based emotion recognition. *Applied Acoustics*, 182.

[29] Baabbad, I., Althubiti, T.and Alharbi, A., and Alfarsi, K.and Rasheed, S. (2021). A short review of classification algorithms accuracy for data prediction in data mining applications. *Journal of Data Analysis and Information Processing*, 09:162–174.

[30] Bakhshi, A., Harimi, A., and Chalup, S. (2022). Cytex: Transforming speech to textured images for speech emotion recognition. *Speech Communication*, 139:62–75.

[31] Bakir, C. and Yuzkat, M. (2018). Speech emotion classification and recognition with different methods for turkish language. *Balkan Journal of Electrical and Computer Engineering*, page 54–60.

[32] Balcar, K. (2009). Trends in studying emotions. re-constructing emotional spaces. http://www.pvsps.cz/data/2017/05/30/12/1\protect\@normalcr\relax-trends-in-emotion-research.pdf.

[33] Bhangale, K. B. and Kothandaraman, M. (2023). Speech emotion recognition using the novel pemonet (parallel emotion network). *Applied Acoustics*, 212:109613.

[34] Bhaskar, J.and Sruthi, K. and Nedungadi, P. (2015). Hybrid approach for emotion classification of audio conversation based on text and speech mining. *Procedia Computer Science*, 46:635–643.

[35] Braunschweiler, N., Doddipatla, R., Keizer, S., and Stoyanchev, S. (2022). Factors in emotion recognition with deep learning models using speech and text on multiple corpora. *IEEE Signal Processing Letters*, 29:722–726.

[36] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of german emotional speech (emodb). *INTERSPEECH*, pages 1517–1520.

[37] Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S., and Narayanan, S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42:335–359.

[38] Busso, C. and Narayanan, S. (2008). Recording audio-visual emotional databases from actors: a closer look. In *in Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC), Marrakech, Morocco*.

[39] Byun, S. and Lee, S. (2021). A study on speech emotion recognition system with

effective acoustic features using deep learning algorithms. *Applied Sciences(Switzerland)*, 11(4):1–15.

[40] Cao, H., Cooper, D., Keutmann, M., Gur, R., Nenkova, A., and Verma, R. (2014). Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. *IEEE Transactions on Affective Computing*, 5:377–390.

[41] Carneiro, H., Weber, C., and Wermter, S. (2023). Whose emotion matters? speaking activity localisation without prior knowledge. *Neurocomputing*, 545:126271.

[42] Chai, J., Zeng, H., Li, A., and Ngai, E. (2014). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100–134.

[43] Chakroborty, S.; Roy, A. and Saha, G. (2006). Fusion of a complementary feature set with mfcc for improved closed set text-independent speaker identification. In *In Proceedings of the IEEE International Conference on Industrial Technology*.

[44] Chen, L., Wang, K., Li, M., Wu, M., Pedrycz, W., and Hirota, K. (2022). K-means clustering based kernel canonical correlation analysis for multimodal emotion recognition in human-robot interaction. *IEEE Transactions on Industrial Electronics*.

[45] Chen, M., He, X., Yang, J., and H., Z. (2018). 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Process. Lett.*, 25(10):1440–1444.

[46] Chen, R., Dewi, C., Huang, S., and Craka, R. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7(1):1440–1444.

[47] Cheng, X. and Duan, Q. (2012). Speech emotion recognition using gaussian mixture model. In *In Proceedings of the 2012 International Conference on Computer Application and System Modeling (ICCASM), Taiyuan, China, 27–29 July*, page 1222–1225.

[48] Chimthankar, P. (2021). Speech emotion recognition using deep learning.

[49] Costantini, G., Iadarola, I., Paoloni, A., and Todisco, M. (2014). Emovo corpus: An italian emotional speech database. In *Proceedings of the 9th International Conference on Language Resources and Evaluation LREC*, pages 3501–3504.

[50] Costantini, G., Parada-Cabaleiro, E., Casali, D., and Cesarini, V. (2023). The emotion probe: On the universality of cross-linguistic and cross-gender speech emotion recognition via machine learning. *Sensors*, 22(7).

[51] Cristina, L., David, G., Zoraida C., Ricardo, K., Juan, M., and Fernado, F. (2022). A proposal for multimodal emotion recognition using aural transformer on ravdess. *Sensors(Switzerland)*.

[52] D., V. and Mukhopadhyay, D. (2016). Age driven automatic speech emotion recognition system. In *In: Proc. Int. Conf. Comput. Commun. Autom. (ICCCA)*, volume 1005–1010.

[53] Daneshfar, F. and Jamshidi, M. B. (2023). An octonion-based nonlinear echo state network for speech emotion recognition in metaverse. *Neural Networks*, 163:108–121.

[54] Daneshfar, F., Kabudian, S., and Neekabadi, A. (2020). Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and gaussian elliptical basis function network classifier. *Applied Acoustics*, 166.

[55] Ding, R., Li, X., Nie, L., Li, J., Si, X., Chu, D., Liu, G., and Zhan, D. (2018). Empirical study and improvement on deep transfer learning for human activity recognition. *Sensors*, 19(1).

[56] Dong, G. and Pun, C.M.and Zhang, Z. (2022). Temporal relation inference network for multimodal speech emotion recognition. In *In IEEE Transactions on Circuits and Systems for*

Video Technology.

[57] Dupuis, K. and Kathleen Pichora-Fuller, M. (2011). Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set. *Canadian Acoustics - Acoustique Canadienne*, 39.

[58] Engberg, I. S. and Hansen, A. V. (2007). Documentation of the danish emotional speech database. Tech. rep., Center for Person Kommunilation, Denmark.

[59] Farooq, M., Hussain, F., Baloch, N., Raja, F., Yu, H., and Zikria, Y. (2020). Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. *Sensors (Switzerland)*, 20:1–18.

[60] Gan, C., Wang, K., Zhu, Q., Xiang, Y., Jain, D. K., and García, S. (2023). Speech emotion recognition via multiple fusion under spatial–temporal parallel network. *Neurocomputing*, 555.

[61] Gao, Y. (2019). Speech-based emotion recognition. https://libraetd.lib.virginia. edu/-downloads/2f75r8498?filename=1 Gao Ye 2019 MS.pdf.

[62] Gao, Y., Li, B., Wang, N., and Zhu, T. (2017). Speech emotion recognition using local and global features. In *Int. Conf. Brain Informatics*, page 3–13.

[63] Gokilavani, M., Katakam, H. Basheer, S., and Srinivas, P. (2022). Ravdness, crema-d, tess based algorithm for emotion recognition using speech. In *in: 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, volume 20.

[64] Gosztolya, G. (2019). Posterior-thresholding feature extraction for paralinguistic speech classification. *Knowledge-Based Systems*, 20:104943.

[65] Gu, Y., Chen, S., and Marsic, I. (2018). Deep multimodal learning for emotion recognition in spoken language. In *In - Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, volume 2018-April, pages 5079–5083.

[66] Guo, L., Wang, L., Dang, J., Chng, E., and Nakagawa, S. (2022). Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition. *Speech Communication*, 136:118–127.

[67] Haider, F., Pollak, S., Albert, P., and Luz, S. (2021). Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods. *Computer Speech and Language*, 65:101–119.

[68] Hajarolasvadi, N. and Demirel, H. (2019). 3d cnn-based speech emotion recognition using k-means clustering and spectrograms. *Entropy*, 21(5).

[69] Haq, S. and Jackson, P. (2091). Multimodal emotion recognition. *Machine Audition: Principles, Algorithms and Systems*.

[70] Harati, S., Crowell, A., Mayberg, H., and Nemati, S. (2018). Depression severity classification from speech emotion. In *In Annual International Conference of the IEEE Engineering in Medicine and Biology Society-IEEE Engineering in Medicine and Biology Society*, pages 5763–5766.

[71] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 770–778.

[72] Heredia, J., Lopes-Silva, E., Cardinale, Y., Diaz-Amado, J., Dongo, I., and Graterol, A. (2022). Adaptive multimodal emotion detection architecture for social robots. *IEEE Access*, 10:20727–20744.

[73] Hisako, O. and Akira, I. (2018). Noise cancellation method for speech signal by using an extension type ukf. *IEEE, Signal Processing: Algorithms, Architectures, Arrangements, and Applications*.

[74] Ho, N., Yang, H., Kim, S., and Lee, G. (2023). Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access*, 8:61672–61686.

[75] Hou, M., Li, J., and Lu, G. (2020). A supervised non-negative matrix factorization model for speech emotion recognition. *Speech Communication*, pages 13–20.

[76] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. (2017a). Densely connected convolutional networks. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition CVPR 2017*, pages 2261–2269.

[77] Huang, Y., Cheng, Y., Bapna, A., Firat, O.and Chen, M., Chen, D., Lee, H. J., Ngiam, J., Le, Q. V., Wu, Y., and Chen, Z. (2019). Gpipe: Efficient training of giant neural networks using pipeline parallelism. *Advances in Neural Information Processing Systems*, 32:1–11.

[78] Huang, Z., Dong, M., and Mao, Q.and Zhan, Y. (2017b). Densely connected convolutional networks. In *Proceedings of the 22nd ACM international conference on Multimedia*, page 801–804.

[79] Iman, H., Arabnia, R., and Branchinst, R. (2017). Pathways to artificial general intelligence: A brief overview of developments and ethical issues via artificial intelligence, machine learning, deep learning, and data science. In *22nd Int'l Conf on Artificial Intelligence: Las Vegas, Nevada, USA*, page 801–804.

[80] Iman, M., Rasheed, K., and Arabnia, H. (2022). A review of deep transfer learning and recent advancements. http://arXiv:2201.09679.

[81] Imani, M. and Montazer, G. A. (2019). A survey of emotion recognition methods with emphasis on e-learning environments. *In Journal of Network and Computer Applications*, 147.

[82] Inger, S., Engberg, V., Hansen, A., and Paul, D. (1997). Design, recording and verification of danish emotional speech database. In *In EUROSPEECH '97, 5th European Conference on Speech Communication and Technology,Rhodes, Greece September 22-25*.

[83] Iqbal, A. and Barua, K. (2019). A real-time emotion recognition from speech using gradient boosting. In *In: Proc. Int. Conf. Electrical, Computer and Communication Engineering*, pages 1–5.

[84] Issa, D., Ehsan, A., and Myoung, L. (2022). Performance evaluation of machine learning and neural network-based algorithms for predicting segment availability in aiot-based smart parking. *MDPI Network*, 2:225–238.

[85] Issa, D., Fatih Demirci, M., and Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59:101894.

[86] Jain, M., Narayan, S., Balaji, K., Bharath, K., Bhowmick, A., Karthik, R., and Muthu, R. K. (2013). Speech emotion recognition using support vector machine. http://arXiv:2002.07590.

[87] Jane, O., Serestina, V., and Adekanmi, A. (2023). Long short-term memory recurrent neural network for automatic speech recognition. *IEEE Access*, 10:30069–30079.

[88] Javier, L. Manuel, G. (2023). An ongoing review of speech emotion recognition. *Neurocomputing*, 528:1–11.

[89] Jiang, W., Wang, Z., Jin, J. S., Han, X., and Li, C. (2020). De-capsnet: A diverse enhanced capsule network with disperse dynamic routing. *Sensors (Switzerland)*, 19(12):1–15.

[90] Jing, E., Liu, Y., Chai, Y., Sun, J., Samtani, S., Jiang, Y., and Qian, Y. (2023). A deep interpretable representation learning method for speech emotion recognition. *Information Processing and Management*, 60(8):103501.

[91] Joshi, V., Ghongade, R., Joshi, A., and Kulkarni, R. (2022). Deep bilstm neural network model for emotion detection using cross-dataset approach. *Biomedical Signal Processing and Control*, 73.

[92] Jothimani, S. and Premalatha, K. (2022). Mff-saug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network. *Speech Communication*, 162:112512.

[93] K., A., M., A., Srinath, K., H., S., Jyothish, L., and Vinayakumar, R. (2023). Speech emotion recognition using cnn-lstm and vision transformer. In *In book: Innovations in Bio-Inspired Computing and Applications*.

[94] Kadiri, S., Gangamohan, P., Gangashetty, S., Alku, P., and Yegnanarayana, B. (2020). Excitation features of speech for emotion recognition using neutral speech as reference. *Circuits, Systems, and Signal Processing*, 39:4459–4481.

[95] Kallipolitis, A., Revelos, K., and Maglogiannis, I. (2021). Ensembling efficientnets for the classification and interpretation of histopathology images. *Algorithms*, 14(10).

[96] Kaur, J. and Kumar, A. (2020). Databases, features and classification techniques for speech emotion recognition. *Journal of Innovative Technology and Exploring Engineering*, 9:185–190.

[97] Kaya, H. and Karpov, A. A. (2018). Efficient and effective strategies for cross-corpus acoustic emotion recognition. *Speech Communication*, 275:1028–1034.

[98] Kerkeni, L., Serrestou, Y., Raoof, K., Mbarki, M., Mahjoub, M., and Cleder, C. (2021). Automatic speech emotion recognition using an optimal combination of features based on emd-tkeo. *Speech Communication*, 114:22–35.

[99] Khalil, R. A., Jonesa, E., Babar, M.I. Jan, T., Zafar, M. H., and Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345.

[100] Koolagudi, S. and Rao, K. (2012). Emotion recognition from speech using source, system, and prosodic features. *Journal of Speech Technology*, 15:265–289.

[101] Kotti, M. and Kotropoulos, C. (2008). Gender classification in two emotional speech databases. In *in: Proc. 19th Int. Conf. on Pattern Recognition*, page 1–4.

[102] Krishna, K. V., Sainath, N., and Posonia, A. M. (2022). Speech emotion recognition using machine learning. In *in: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, page 1014–1018.

[103] Krizhevsky, B., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. communications of the acm). *Applied Acoustics*, 60:84–90.

[104] Kumar, D., Nahidul, I., Rayhan, A., Salekul, I., Swakkhar, S., and Muzahidul, I. (2022a). Banglaser: A speech emotion recognition dataset for the bangla language. *Data in Brief*, 42:108091.

[105] Kumar, S., Shrimal, A., Akhtar, M. S., and Chakraborty, T. (2022b). Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems*, 240:108112.

[106] Kwabena Patrick, M., Felix Adekoya, A., Abra Mighty, A., and Edward, B. (2019). Capsule networks–a survey. *Journal of King Saud University–Computer and Information Sciences*, 34:1295–1310.

[107] Langari, S., Marvi, H., and Zahedi, M. (2020). Efficient speech emotion recognition using modified feature extraction. informatics in medicine unlocked. *Journal of King Saud University–Computer and Information Sciences*, 20:100424.

[108] Lanjewar, R. B., Mathurkar, S., and Patel, N. (2020). Implementation and comparison of speech emotion recognition system using gaussian mixture model (gmm) and k- nearest neighbor (k-nn) techniques. *Phys. Rev. E.*, 49:50–57.

[109] Latif, S., Qayyum, A., Usman, M., and Qadir, J. (2018). Cross lingual speech emotion recognition: Urdu vs. western languages. In *Proceedings - 2018 International Conference*

on Frontiers of Information Technology, FIT, pages 88–93.

[110] Latif, S., Shahid, A., and Qadir, J. (2023). Generative emotional ai for speech emotion recognition: The case for synthetic emotional speech augmentation. *Applied Acoustics*, 210:109425.

[111] Lerner, J., Li, Y.and Valdesole, P., and Kassam, K. (2014). Emotion and decision making. *Annual Review of Psychology*, June:1–45.

[112] Li, R.and Zhao, J. and Jin, Q. (2021). Speech emotion recognition via multi-level cross-modal distillation. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 606–610.

[113] Li, X. and Akagi, M. (2019). Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model. *Speech Communication*, 110:1–12.

[114] Li, Y., Sun, H., Feng, S., Zhang, Q., Han, S., and Du, W. (2021). Capsule-lpi: a lncrna-protein interaction predicting tool based on a capsule network. *BMC Bioinformatics*, 22:1–19.

[115] Lian, Z., Liu, B., and Tao, J. (2021). Decn: Dialogical emotion correction network for conversational emotion recognition. *Neurocomputing*, 454:483–495.

[116] Lieskovska, E., Jakubec, M., Jarina, R., and M., C. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics(Switzerland)*, 10(10).

[117] Lin, M., Chen, Q., and Yan, S. (2014). Network in network. In *2nd International Conference on Learning Representations, ICLR 2014-Conference Track Proceedings*, pages 1–10.

[118] Lin, Y. L. and Wei, G. (2005). Speech emotion recognition based on hmm and svm. In *Proc. Fourth IEEE Int. Conf. on Machine Learning and Cybernetics*, page 4898–901.

[119] Liu, L.-Y., Liu, W.-Z., Zhou, J., Deng, H.-Y., and Feng, L. (2022). Atda: Attentional temporal dynamic activation for speech emotion recognition. *Knowledge-Based Systems*, 243:108472.

[120] Liu, Z., Wu, M., Cao, W., Mao, J., Xu, J., and Tan, G. (2018). Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*, 273:271–280.

[121] Liu, Z.-T., Rehman, A., Wu, M., Cao, W.-H., and Hao, M. (2021). Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence. *Information Sciences*, 563:309–325.

[122] Liu, Z.-T., Wu, B.-H., Han, M.-T., Cao, W.-H., and Wu, M. (2023). Speech emotion recognition based on meta-transfer learning with domain adaption. *Applied Soft Computing*, 110766.

[123] Livingstone, S. and Russo, F. (2018). The ryerson audio-visual database of emotional speech and song (ravdess). *In PLoS ONE*.

[124] Lotfian, R. and Busso, C. (2016). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *Journal of IEEE Transactions on Affective Computing*.

[125] Lotfian, R. and Busso, C. (2018). Curriculum learning for speech emotion recognition from crowdsourced labels. http://arxiv.org/abs/1805.10339.

[126] Lotfidereshgi, R. and Gournay, P. (2017). Biologically inspired speech emotion recognition. In *In: Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, page 5135–5139.

[127] Ma, F., Li, Y., Ni, S., Huang, S., and Zhang, L. (2022). Data augmentation for audio–visual emotion recognition with an efficient multimodal conditional gan. *Applied Sciences*, 12(1).

[128] Mahla, A. and Peter, X. L. (2021). Deep learning for face image synthesis and semantic manipulations: a review and future perspectives. *Artificial Intelligence Review*, 53:5847–5880.

[129] Maithri, M., Raghavendra, U., Gudigar, A., Samanth, J., Prabal, D., Murugappan, M., Chakole, Y., and Acharya, U. (2022). Automated emotion recognition: Current trends and future perspectives. *Computer Methods and Programs in Biomedicine*, 215.

[130] Manohar, K. and Logashanmugam, E. (2022). Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm. *Knowledge-Based Systems*, 246:215–227.

[131] Mansour, S., Mahdi, B., and Davood, G. (2013). Modular neural-svm scheme for speech emotion recognition using anova feature selection method. *Neural Computing and Applications*, 23(1):215–227.

[132] Mao, X., Chen, L., and Fu, L. (2009). Multi-level speech emotion recognition based on hmm and ann. In *In Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, CA, USA, 31 March–2 April*, page 225–229.

[133] Mapelli, V. (2011). Spanish emotional speech synthesis database, european language resources association. http://metashare.ilsp.gr:8082/repository/browse/emotional-speech-synthesis-database.

[134] Marc, K., Hasan, F., Heshmat, A., and Cliston, C. (2014). Noise removal in speech processing using spectral subtraction. *Journal of Signal and Information Processing*.

[135] Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). The enterface 05 audio-visual emotion database. In *ICDEW 2006 Proceedings of the 22nd International Conference on Data Engineering Workshops*, pages 1–9.

[136] Mary Little Flower, T. and Jaya, T. (2022). Speech emotion recognition using ramanujan fourier transform. *Applied Acoustics*, 201:109133.

[137] Matin, R. and Valles, D. (2020). A speech emotion recognition solution-based on support vector machine for children with autism spectrum disorder to help identify human emotions. In *In: 2020 Intermountain Engineering, Technology and Computing (IETC)*.

[138] Meng, H., Yan, T., Yuan, F., and Wei, H. (2019). Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE Access*, 7:125868–12588.

[139] Middya, A. I., Nag, B., and Roy, S. (2022). Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowledge-Based Systems*, 244:108580.

[140] Middya, A. I., Nag, B., and Roy, S. (2023). Deep learning based multimodal emotion recognition using model-level fusion of audio-visual modalities. *Knowledge Based Systems*, 244.

[141] Ming, L., Angeliki, M., Daniel, B., and N., S. (2019). Speaker states recognition using latent factor analysis based eigenchannel factor vector modeling. In *In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*.

[142] Mishra, S. P., Warule, P., and Deb, S. (2022). Variational mode decomposition based acoustic and entropy features for speech emotion recognition. *Applied Acoustics*, 212:109578.

[143] Mocanu, B. and Tapu, R. (2022). Emotion recognition from raw speech signals using 2d cnn with deep metric learning. In *In: 2022 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–5.

[144] Mohammad, A., Mohammad, S. U., Mayeen, U., Nissren, T., and S., A. (2023). The efficacy of deep learning-based mixed model for speech emotion recognition. *Computers, Materials & Continua*.

[145] Mohammadrezaei, P., Aminan, M., Soltanian, M., and Borna, K. (2023). Improving cnn-based solutions for emotion recognition using evolutionary algorithms. *Results in Applied Mathematics*, 18:100360.

[146] Mustaqeem Sajjad, M. and Kwon, S. (2021). Clustering-based speech emotion recognition by incorporating learned features and deep bilstm.

IEEE Access, 8:79861–79875.

[147] Naderi, N. and Nasersharif, B. (2023). Cross corpus speech emotion recognition using transfer learning and attention-based fusion of wav2vec2 and prosody features. *Knowledge-Based Systems*, 277:110814.

[148] Nagase, R., Fukumori, T., and Yamashita, Y. (2022). Speech emotion recognition using label smoothing based on neutral and anger characteristics. In *In: 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech)*, page 626–627.

[149] Nakisa, B. (2019). Emotion classification using advanced machine learning techniques applied to wearable physiological signals data. https://eprints.qut.edu.au/129875/9/Bahareh%20Nakisa%20Thesis.pdf.

[150] Nam, Y. and Lee, C. (2021). Cascaded convolutional neural network architecture for speech emotion recognition in noisy conditions. *Sensors*, 21:1–17.

[151] Nasir, S., Jiechao, G., Rizwana, I., Ahmad, A., Tayyab, R., Yudong, Z., and Seifedine, K. (2021). Caai transactions on intelligence technology. *Jowh Wiley*, 8:401–417.

[152] Neyshabur, B., Sedghi, H., and Zhang, C. (2020). What is being transferred in transfer learning? Adv.NeuralInf.Process.Syst.,arXiv:2008.11687v2.

[153] Oh, S. and Kim, D. (2022). Comparative analysis of emotion classification based on facial expression and physiological signals using deep learning. *Applied Sciences (Switzerland)*, 12(3).

[154] Ooi, C. S., Seng, K. P., Ang, L. M., and Chew, L. (2014). A new approach of audio emotion recognition. *Expert Systems with Applications*, 41:5858–5869.

[155] Ortega, J., Cardinal, P., Koerich, A., and Jun, L. G. (2019). Emotion recognition using fusion of audio and video features. http://arxiv.org/abs/1906.10623v1.

[156] Ouyang, M., Das, R., Yang, J., and Li, H. (2021). Capsule network based end-to-end system for detection of replay attacks. In *In: 2021 12th International Symposium on Chinese Spoken Language Processing, ISCSLP*.

[157] OZER, I. (2021). Pseudo-colored rate map representation for speech emotion recognition. *Biomedical Signal Processing and Control*, 66:102502.

[158] Padi, S., Sadjadi, S., Sriram, R., and Manocha, D. (2021). Improved speech emotion recognition using transfer learning and spectrogram augmentation. In *ICMI '21: Proceedings of the 2021 International Conference on Multimodal Interaction*, page 45–652.

[159] Palo, H. and Mohanty, M. (2018). Wavelet based feature combination for recognition of emotions. *Ain Shams Engineering Journal*, 9:1799–1806.

[160] Pan, S. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359.

[161] Pandey, S. K., Shekhawat, H., and Prasanna, S. (2018). Attention gated tensor neural network architectures for speech emotion recognition. *Biomedical Signal Processing and Control*, 71.

[162] Pham, N. T., Dang, D. N. M., Nguyen, N. D., Nguyen, T. T., Nguyen, H., Manavalan, B., Lim, C. P., and Nguyen, S. D. (2023). Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition. *Expert Systems with Applications*, 230:120608.

[163] Pinto, M., Polignano, M., Lops, P., and Semeraro, G. (2020). Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients. In *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), Bari, Italy*, pages 1–5.

[164] Popova, A., Rassadin, A., and A., P. (2018). Emotion recognition in sound. *Advances in*

Neural Computation, Machine Learning, and Cognitive Research, page 117–124.

[165] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

[166] Prasanth, S., Roshni Thanka, M., Bijolin Edwin, E., and Nagaraj, V. (2021). Speech emotion recognition based on machine learning tactics and algorithms. *Materials Today: Proceedings*.

[167] Praseetha, V. M. and Vadivel, S. (2018). Deep learning models for speech emotion recognition. *J. Computer Science*, 14 (11):1577–1587.

[168] Puri, T., Soni, M., Dhiman, G., Khalaf, O. I., alazzam, M., and Khan, I. R. (2022). Detection of emotion of speech for ravdess audio using hybrid convolution neural network. *Journal of Healthcare Engineering*, 8472947.

[169] Ramakrishna Thirumuru, Krishna Gurugubelli, A. K. V. (2022). Novel feature representation using single frequency filtering and nonlinear energy operator for speech emotion recognition. *Digital Signal Processing*, 120, 103293.

[170] Ren, M., Nie, W., Liu, A., and Su, Y. (2019). Multi-modal correlated network for emotion recognition in speech. *Digital Signal Processing*, 3:150–155.

[171] Renjith, S. and Manju, K. (2017). Speech-based emotion recognition in tamil and telugu using lpcc and hurst parameters-a comparative study using knn and ann classifiers. In *Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies ICCPCT*.

[172] Ruhul, K., Edward, J., Mohammad, B. Tariqullah, J., Mohammad, Z., and Thamer, A. (2016). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345.

[173] S, A. and Viriri, S. (2022). A robust feature selection-based speech emotion classification using deep transfer learning. *Appl. Sci.*, 12, 8265.

[174] Sabour, S., Frosst, N., and Hinton, G. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems*, pages 3857–3867.

[175] Satya, P. Y., Subiya, Z. Annu, M., and Vibhash, Y. (2022). Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (rnn). *Archives of Computational Methods in Engineering*, 1753–1770:1753–1770.

[176] Shahin, I., Alomari, O. A., Nassif, A. B., Afyouni, I., Hashem, I. A., and Elnagar, A. (2023). An efficient feature selection method for arabic and english speech emotion recognition using grey wolf optimizer. *Applied Acoustics*, 205:109279.

[177] Shahin, I., Hindawi, N., Nassif, A., Alhudhaif, A., and Polat, K. (2022). Novel dual-channel long short-term memory compressed capsule networks for emotion recognition. *Expert Systems with Applications*, 188.

[178] Shixin, P., Kai, C., Tian, T., and Jingying, C. (2022). An autoencoder-based feature level fusion for speech emotion recognition. *Digital Communications and Networks*, 18.

[179] Shou, Y., Meng, T., Ai, W., Yang, S., and Li, K. (2022). Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis. *Neurocomputing*, 501:629–639.

[180] Simonyan, K. and Zisserman (2015). Very deep convolutional networks for large–scale image recognition. In *In: 3rd International Conference on Learning Representations, ICLR*, volume 1-14.

[181] Singh, V. and Prasad, S. (2022). Speech emotion recognition system using gender dependent convolution neural network. *Procedia Computer Science*, 218:2533–2540.

[182] Singh, Y. B. and Goel, S. (2023). A systematic literature review of speech emotion recognition approaches. *Neurocomputing*, 492:245–263.

[183] Sinith, M., Aswathi, E., Deepa, T., Shameema, C., and Rajan, S. (2015). Emotion recognition from audio signals using support vector machine. In *In: 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, volume 139–144.

[184] Slimi, A., Hamroun, M., Zrigui, M., and Nicolas, H. (2020). Emotion recognition from speech using spectrograms and shallow neural networks. In *In:ACM Int. Conf. Advances in Mobile Computing and Multimedia*, volume 298–301.

[185] Sowmya, G., Naresh, K., Sri, J., Sai, K., and Indira, D. (2022). Speech2emotion: Intensifying emotion detection using mlp through ravdess datase. In *In: 2022 International Conference on Electronics and Renewable Systems (ICEARS*, volume 1–3.

[186] Staudemeyer, R. and Morris, E. (2019). Understanding lstm: a tutorial into long-short-term memory recurrent neural networks. arxiv.org/abs/1909.09586.

[187] Sunitha-Ram, C. and Ponnusamy, R. (2014). An effective automatic speech emotion recognition for tamil language using support vector machine. In *In: 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, volume 19–23.

[188] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2017). Going deeper with convolutions. proceedings of the ieee computer society conference on computer vision and pattern recognition. In *In:Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1–9.

[189] Sánchez-Gutiérrez, M. E. and González-Pérez, P. P. (2020). Discriminative neural network pruning in a multiclass environment: A case study in spoken emotion recognition. *Speech Communication*, 120:20–30.

[190] Talo, M., Baloglu, B., YıldırımU, O., and Acharya, R. (2015). Application of deep transfer learning for automated brain abnormality classification using mr images. *Cognitive Systems Research*, 492:176–188.

[191] Tan, C.and Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2019). A survey on deep transfer learning. arxiv.org/abs/1808.01974.

[192] Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *In:36th International Conference on Machine Learning, ICML*, volume 10691-10700.

[193] Tapu, B. M. R. and Zaharia, T. (2023). Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. *Image and Vision Computing*, 133:104676.

[194] Thirumuru, R., Gurugubelli, K., and Vuppala, A. K. (2022). Novel feature representation using single frequency filtering and nonlinear energy operator for speech emotion recognition. *Digital Signal Processing*, 120:176–188.

[195] Van, L., Le Dao, T., Le Xuan, T., and Castelli, E. (2022). Emotional speech recognition using deep neural networks. *Sensors*, 22(4).

[196] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 120:5999–6009.

[197] Ververidis, D. and Kotropoulos, C. (2004). Automatic speech classification to five emotional states based on gender information. In *In: 12th IEEE European Signal Processing Conf.*, volume 341–344.

[198] Vryzas, N., Kotsakis, R., Liatsou, A., Dimoulas, C., and Kalliris, G. (2018). Speech emotion recognition for performance interaction. *AES: Journal of the Audio Engineering Society*, 66(6):457–467.

[199] Vryzas, N., Vrysis, L., Kotsakis, R., and Dimoulas, C. (2021). A web crowdsourcing framework for transfer learning and personalized speech emotion recognition. *Machine Learning with Applications*, 6:100132.

[200] Wang, C., Ren, Y., Zhang, N., Cui, F., and Luo, S. (2022). Speech emotion recognition based on multi-feature and multi-lingual fusion. *Multimedia Tools and Applications*, 81(4):4897–4907.

[201] Wang, J., Liu, Q., Xie, H., Yang, Z., and Zhou, H. (2021a). Boosted efficientnet: Detection of lymph node metastases in breast cancer using convolutional neural networks. *Cancers (Basel)*, 13:1–14.

[202] Wang, K., Su, G., Liu, L., and Wang, S. (2020a). Wavelet packet analysis for speaker independent emotion recognition. *Neurocomputing*, 398:257–264.

[203] Wang, X., Chen, X., and Cao, C. (2020b). Human emotion recognition by optimally fusing facial expression and speech feature. *Signal Processing: Image Communication*, 84:115831.

[204] Wang, X., Mao, Q., Zhan, Y., Simos, E., and Psihoyios, G. (2008). Speech emotion feature selection method based on contribution analysis algorithm of neural network. In *AIP Conference Proceedings*, page 336–339.

[205] Wang, Y., Boumadane, A., and Heba, A. (2021b). A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. arxiv.org/abs/2111.02735.

[206] Wang, Z., Chen, J., and Hoi, S. (2021c). Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3365–3387.

[207] Wani, T., Gunawan, T., Qadri, S. A. A., and Kartiwi, M. Ambikairajah, A. (2021). A comprehensive review of speech emotion recognition systems. *IEEE Access*, 9:47795–47814.

[208] Weiss, K., Khoshgoftaar, T., and Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*.

[209] Wen, G., Liao, H., Li, H., Wen, P., Zhang, T., Gao, S., and Wang, B. (2022). Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition. *Knowledge-Based Systems*, 254:109589.

[210] Wen, J., Jiang, D., Tu, G., Liu, C., and Cambria, E. (2023). Dynamic interactive multiview memory network for emotion recognition in conversation. *Information Fusion*, 91:123–133.

[211] Wu, S. (2021). Expression recognition method using improved vgg16 network model in robot interaction. *Journal of Robotics*, 398.

[212] Wu, S., Zhong, S., and Liu, Y. (2018). Deep residual learning for image steganalysis. *Multimedia Tools and Applications*, 77:10437–10453.

[213] Xie, J., Zhu, M., and Hu, K. (2023a). Fusion-based speech emotion classification using two- stage feature selection. *Speech Communication*, 66(6):102955.

[214] Xie, J., Zhu, M., and Hu, K. (2023b). Fusion-based speech emotion classification using two-stage feature selection. *Speech Communication*, 152:102955.

[215] Yao, Z., Wang, Z., Liu, W., Liu, Y., and Pan, J. (11-19). Speech emotion recognition using fusion of three multi-task learning-based classifiers: Hsf-dnn, ms-cnn and lld-rnn. *Speech Communication*, 120.

[216] Yi, L. and W., M. M. (2022). Improving speech emotion recognition with adversarial data augmentation network. *IEEE Transactions on Neural Networks and Learning*, 33 (1):172–184.

[217] Zamil, A., Hasan, S., Baki, S., Adam, J., and Zaman, I. (2019). Emotion detection from speech signals using voting mechanism on classified frames. In *In: 2019 International Conference on Robotics, Electrical and Signal Processing Technique, Bangladesh*, volume 281–285.

[218] Zhang, H., Gou, R., Shang, J., Shen, F., Wu, Y., and Dai, G. (2021a). Pre-trained deep convolution neural network model with attention for speech emotion recognition. *Frontiers in Physiology*, 12.

[219] Zhang, H., Li, Z., Zhao, H., Li, Z., and Zhang, Y. (2022). Attentive octave convolutional capsule network for medical image classification. *Mathematics*.

[220] Zhang, J., Meng, C., Xu, C., Ma, J., and Su, W. (2021b). Deep transfer learning method based on automatic domain alignment and moment matching. *Mathematics*, 10(2531).

[221] Zhang, Q., An, N., Wang, K., Ren, F., and Li, L. (2013). Speech emotion recognition using combination of features. In *In: 2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP)*, volume 23–528.

[222] Zhang, S., Yang, Y., Chen Chen, R. L., Tao, X., Guo, W., Xu, Y., and Zhao, X. (2023a). A systematic literature review of speech emotion recognition approaches. *Neurocomputing*, 85:105052.

[223] Zhang, Y., Du, J., Wang, Z., and Zhang, J. (2018). Attention based fully convolutional network for speech emotion recognition. In *In: Proc. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, volume 1771–1775.

[224] Zhang, Y., Wang, J., Liu, Y., Rong, L., Zheng, Q., Song, D., Tiwari, P., and Qin, J. (2023b). A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations. *Information Fusion*, 93:282–301.

[225] Zhao, J., Mao, X., and Chen, L. (2021a). Speech emotion recognition using deep 1d, 2d cnn lstm networks. *Biomedical Signal Processing and Control*, 2019:312–323.

[226] Zhao, Z., Li, Q., Zhang, Z., Cummins, N., Wang, H., Tao, J., and Schuller, B. (2021b). Combining a parallel 2d cnn with a self-attention dilated residual network for ctc–based discrete speech emotion recognition. *Neural Networks*, 141:52–60.

[227] Zheng, L., Li, Q., Ban, H., and Liu, S. (2018). Speech emotion recognition based on convolution neural network combined with random forest. In *Proceedings of the 30th Chinese Control and Decision Conference, CCDC*, volume 4143-4147.

[228] Zhou, S. and Beigi, H. (2020). A transfer learning method for speech emotion recognition from automatic speech recognition. arxiv.org/abs/2008.02863.

[229] Zhou, Y., Liang, X., Gu, Y., Yin, Y., and Yao, L. (2021). Multi-classifier interactive learning for ambiguous speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:695–705.

[230] Zhu, L., Chen, L., Zhao, D., Zhou, J., and Zhang, W. (2017). Emotion recognition from chinese speech for smart affective services using a combination of svm and dbn. *Sensors(Switzerland)*, 17.

[231] Zhu-Zhou, F., Gil-Pita, R., Garcia-Gomez, J., and Rosa-Zurera, M. (2021). Robust multi-scenario speech-based emotion recognition system. *Sensors*, 22 (6).

Chapter 3

Feature Selection and Attention Mechanisms in Deep Learning for Speech Emotion Classification

Deep learning approaches to improve Speech Emotion Classification are discussed in this chapter.

3.1 Introduction

The published works in this chapter utilized a deep transfer learning and attention network architecture for speech emotion classification. It summarises the speech emotion classification model that uses transfer learning to achieve better performance, looks into the role of robust feature selection in deep learning models for SEC, and researches how attention-based networks may be used to focus on important emotional cues in speech signals. The speech feature extraction and selections from speech signals, the transfer learning model, and the dataset utilized are presented.

3.2 A Robust Deep Transfer Learning Model for Accurate Speech Emotion Classification

3.2.1 Brief Review

This chapter presents the paper entitled A Robust Deep Transfer Learning Model for Accurate Speech Emotion Classification. In this paper, a deep learning technique that is based on the transfer of knowledge from source to target domain has been adapted for the classification of speech emotion. The paper utilized a deep

convolutional neural network that is based on VGG-Net architecture and incorporated a principal component analysis for feature-down sampling. The top layer of the model architecture is replaced with a classifier for classification of seven different emotions, after feature extraction. An efficient deep learning model with batch normalization and Multilayer perceptron classifier is developed, which overcomes the challenge of a limited annotated speech emotion dataset.

Paper status: Published in Springer, LNCS 13599, 2022.

A Robust Deep Transfer Learning Model for Accurate Speech Emotion Classification

Samson Akinpelu and Serestina Viriri( )

School of Mathematics, Statistics and Computer Science,
University of KwaZulu-Natal, Durban, South Africa
222068579@stu.ukzn.ac.za, viriris@ukzn.ac.za

Abstract. The significant role of emotion in human daily interaction cannot be over-emphasized, however, the pressing demand for a cutting-edge and highly efficient model for the classification of speech emotion in effective computing has remained a challenging task. Researchers have proposed several approaches for speech emotion classification (SEC) in recent times, but the lingering challenge of the insufficient dataset, which has been limiting the performances of these approaches, is still of major concern. Therefore, this work proposes a deep transfer learning model, a technique that has been yielding tremendous and state-of-the-art results in computer vision, for SEC. Our approach used a pre-trained and optimized model of Visual Geometry Group (VGGNet) convolutional neural network architecture with appropriate fine-tuning for optimal performance. The speech signal is converted to a mel-Spectrogram image suitable for deep learning model input (224×244 x 3) using filterbanks and Fast Fourier transform (FFT) on the speech samples. Multi-layer perceptron (MLP) algorithm is adopted as a classifier after feature extraction is carried out by the deep learning model. Speech pre-processing was carried out on Toronto English Speech Set (TESS) speech emotional corpus used for the study to prevent the low performance of our proposed model. The result of our experiment after evaluation using the TESS dataset shows an improved result in SEC with an accuracy rate of 96.1% and specificity of 97.4%.

Keywords: Deep learning · Speech emotion · Classification · Deep convolutional neural network

1 Introduction

Speech emotion classification (SEC) entails classifying emotion from emotional content inherent in the speech signal. Obviously, from all human modes of communication, speech occupies a significant position, because it carries several non-linguistic information that is useful and vital to human-computer interaction (HCI). This has resulted in the change of focus by researchers in human-computer interaction, especially in the emotion recognition domain, towards exploring speech signals for efficient and accurate classification of emotion. Apart from

G. Bebis et al. (Eds.): ISVC 2022, LNCS 13599, pp. 419–430, 2022.
https://doi.org/10.1007/978-3-031-20716-7_33

rich para-linguistic information that resides in speech signal, it is less cumbersome to acquire compared to other non-speech signals [1]. As a 1D (dimensional) function of frequency and time, the speech signal is produced at a varying interval (though in a parallel form) through a combination of different human sound production organs. Precisely, the mouth, lips, tongue, and vocal cord among others are majorly involved in the articulation of speech sounds [2].

Emotion is crucial in everyday human interactions. It is a determining factor in decision-making and expression of interest [3]. The totality of human comportment revolves around his emotion per moment. This is the reason why emotion classification from speech signal has recently attracted an influx of researchers. Besides, the application of emotion classification in adaptive e-learning, psychological disorder treatment, self-driving vehicles, and a host of others contributes to its popularity across the research community [4,5]. Therefore, making a machine to identify key emotional features for accurate classification of emotion without much complexity, coupled with limited dataset availability and considering human language accents from multi-cultural backgrounds remains a lingering challenge in SEC [6]. Several approaches, from traditional to neural networks and convolutional neural networks (CNN) have been adopted by many published works in SEC, but the expected performance has not reached a satisfactory level yet. Traditional machine learning algorithms [7] like support vector machine (SVM), Random Forest (RF) and K-Nearest Neighbour (KNN) have been applied for SEC, but their limitations lie in the inability to handle large speech samples with complex features [8]. They can easily identify primitive features such as chroma, pitch, and formant from speech signals [9]. Contrariwise, CNN [10] architectures have proven to be highly efficient in extraction of feature in computer vision and deep learning related task. Literature on SEC is littered with several layered approaches of CNN which have been used by scientific scholars. However, how to manage an increasing number of parameters that exist through the convolutional process and training for optimal performance without a loss of distinctive features from the speech signal and yet minimize the computational cost is still a subject of concern.

Consequently, we proposed a deep transfer learning approach that is based on deep convolutional neural network (DCNN) for speech emotion classification. We used a pre-trained model, already trained on ImageNet; with moderate set of parameters for extraction of spectral and relevant features that carries emotional information for accurate classification of emotion. These features are extracted from a mel-spectrogram image generated from the audio signal. After this, we utilized the feature dimensionality reduction approach of principal component analysis (PCA), but retained the distinctive feature for SEC. To reduce high parameters and computational cost that is peculiar to many CNN-based models, we introduced a multi-layer perceptron (MLP) classifier at the topmost layer of our model to replace the fully connected layers. The performance of our proposed model is evaluated on an open-source speech emotion corpus known as TESS (Toronto English Speech Set) dataset. The speech samples contained in the original TESS corpus requires little class data balancing. We used an

up-sampling approach to mitigate the class imbalance of the dataset. The accuracy report shows a state-of-the-art improved performance compared to other studies which have been carried out.

The contributions of this study are:

1. To re-establish the possibility of using pre-trained DCNN model for classification of speech emotion and efficiently utilize deep transfer learning without training from scratch.
2. A PCA algorithm to reduce dimensionality while retaining emotional features from mel-spectrogram to avoid misclassification of emotion.
3. The proposed model (DCNN-MLPwPCA) achieved a state-of-the-art result over existing method for SEC.

The remaining section of this work is arranged in this order: Sect. 2, a review of several methods adopted for SEC, is carried out. The detail information of our proposed methodology is highlighted in Sect. 3. In Sect. 4, we presented the result of our experiment over the benchmark speech dataset used. Finally, in Sect. 5, the conclusion and future work is presented.

2 Literature Review and Related Work

The result-oriented performance and recent upsurge in the application of deep learning in computer vision, have motivated several research works to be carried out on speech emotion classification while adopting a state-of-the-art approach. In fact, many transfer learning models have been used in the last decade. Latif et al. [11] adopted transfer learning techniques for multi-corpus speech emotion classification. Deep belief network (DBN) was used in their experiment on five different speech emotional corpora. the DBN was based on a contiguous Restricted Boltzmann Machine (RBM). Their investigation also revealed the effect of cross-language in SEC, especially for training and validation. In Farooq et al. [12], the effect of the feature selection algorithm on the transfer learning model for SEC was investigated. A pre-trained DCNN (AlexNet) model was used for feature extraction before feature selection method was applied, after the fourth convolutional layer. Several classifiers like SVM, KNN, RF and MLP were utilized on four speech emotions corpora and the accuracy result surpassed handcrafted approach for SEC. Lech et al. [13] proposed a transfer learning of deep CNN for expanding the speech emotion dataset with the utmost aim of increasing the accuracy of the classification of emotion. From their methodology, knowledge transfer from one domain (source) to another domain (target) can achieve good result with minimal dataset and lesser complexity. This is peculiar to the speech emotion classification domain, in which a large dataset is scarce.

A distance loss transfer learning techniques using siamese neural network was presented by Feng et al. [14] for automatic classification of speech emotion. After fine-tuning of the model, an experiment was carried using eNTERFACE,

RAVDESS and CREMA-D dataset respectively and the result outperform state-of-the-art model. Atsavasirilert et al. [15] applied transfer learning technique of an optimized AlexNet pre-trained model on log-mel-spectrogram for SEC. The first two convolutional layers of their model were fine-tuned using the non-square kernel approach. The parameters of their model were reduced by 272 times after optimization, and an 87.16% unweighted accuracy rate was recorded on EMO-DB dataset. Motivated by the performance of deep transfer learning model in SEC without training cost, Padi et al. [16] proposed a pre-trained Residual network (ResNet34) and spectrogram augmentation for accurate classification of speech emotion. They prevented overfitting of their model through generation of additional training samples.

A bi-directional feature extraction with mel-spectrogram was used in transfer learning of Visual Geometry Group network (VGGNet) pre-trained model for speech emotion classification in Aggarwal et al. [17]. The deep learning approach utilized dense and dropout layer for significant accuracy over RAVDESS dataset. Dimensionality of their proposed model was reduced using principal component analysis. An attention based DCNN with LSTM was proposed by Zhang et al. [18] for SEC. Multi-channel mel-spectrogram was generated from raw audio samples and fed into the DNN for segment-level features. Two fully connected layers were used for classification. The effectiveness of their architecture was evaluated on two datasets (IEMOCAP and EMO-DB) and an improved result was recorded. A fusion of audio and video features using pre-trained CNN techniques was used in Ortega et al. [19]. Their model combined several temporal and spatial features for classification of speech emotion with higher accuracy on RECOLA dataset. Some convolutional layers were frozen to minimize complexity, while the resultant output layer was replaced by a neuron with linear activation function. Vatcharaphrueksadee et al. [20] combined CNN with pre-trained VGG-16 for emotion classification. Their model was optimized before they could achieve an improved result. However, despite the optimization, the complexity of their model was high with over 30 million trainable parameters. Lastly, frantic effort toward increasing availability of large speech emotion dataset has also leverage on deep transfer learning model for implementing new Amharic language speech emotion corpus as presented by Retta et al. [21]. Three models comprising ResNet50, Alex-Net and LSTM were compared using Cepstral coefficient and mel-spectrogram features as input from speech signal.

Some proposed methods above involve huge number of trainable parameters which tends to increase complexity, memory consumption and other far-reaching effect in terms of their performance. However, our proposed method is structured in terms of economising memory consumption, reducing complexity as trainable parameters have drastically reduced to the barest minimum and dimensionality reduction mechanism is also applied on features. This will eliminate overfitting as well. The top layer of our model is replaced with a classifier, thereby reducing the number of convolutional layers while mitigating misclassification of emotion from speech signal.

3 Methods and Techniques

3.1 DCNN for Speech Emotion Classification

Deep convolutional neural networks from VGGNet have been adopted in this work. The learning process through its weight for image classification problem whereby it had been trained is transferred into emotion classification. The description of the proposed method is depicted in Fig. 1 below.

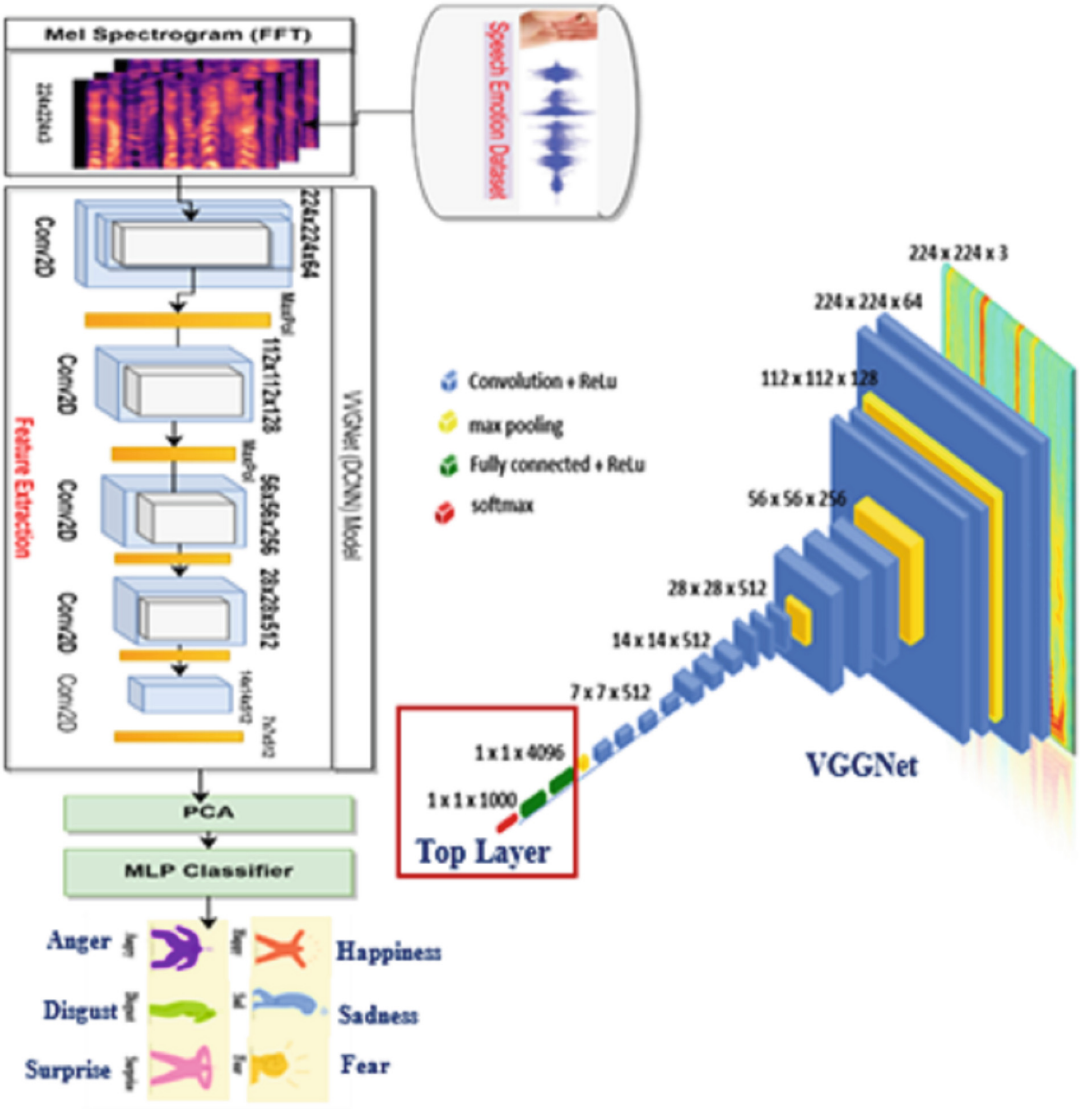


Fig. 1. Propose model framework.

Fine-tuning of the architecture for improved performance is carried out. Following an end-to-end approach, the model operates by learning and extracting features from mel-spectrogram image fed into it from speech signal. The supervised end-to-end top layer (FCN) is replaced with an efficient classifier to produce

emotion-wise prediction. Both feedforward and backpropagation networks of the DCNN are used for the extraction of features from the input image simultaneously.

The input to the DCNN model of VGGNet is generated from speech signal converted to mel-spectrogram through filterbank and FFT. Mel-spectrogram has proven to be useful in many speech classification problems because of the rich information it retains from the original audio wave (speech) signal, even after conversion. It shows a time-frequency spectrum [22] visualization of speech signal. The filterbank performs decomposition of raw signal into several components.In processing the speech signal, 16kHz sampling rate, 512 FFT and filterbanks of 40 were utilized. The input layer, five convolutional layers, pooling layers, and a fully connected (FC) layer as the top layer make up the VGGNet [23] framework of our proposed model. As shown in Fig. 1 of our proposed model, the model has been fine-tuned by replacing the FC layer with a classifier for efficient classification of emotion. VGGNet has been pre-trained on ImageNet, a million database of images, and it requires less time in adapting it for a new domain like speech emotion classification. It has the ability to learn discriminating features [24] through randomly initialized layers, which has given it an edge in feature extraction over other deep transfer learning models. It is such an architecture with a sizeable number of filters (3×3). The next subsection gives brief the details of our proposed model.

3.2 Extraction of Features

The level of accuracy of a model is determined by the feature extraction approach employed in speech emotion classification. In deep learning models for speech emotion classification, feature extraction from speech signals plays an important role. This is done by translating the speech signal (mel-spectrogram) into a quantitative representation that can then be processed and analyzed at a lower data rate. It is carried out in our proposed model by utilizing VGGNet, a pretrained DCNN. The input mel-spectrogram is resized ($224 \times 224 \times 3$) to conform to the requirement of the model. The stacked 2D convolutional (Conv1 - Conv5) layers comprised of 64, 128, 256 and 512 filters size. Each layer of the convolution is followed by a max-pooling layer. The first convolutional layer consists of $224 \times 224 \times 64$ kernel size with 4 pixels stride. There are 5 max-pooling layers altogether, with a pooling size of 3×3 each. The max-pooling layer utilized is for down sampling the output from each convolutional layer to highlight the most available feature map in the patch and produce unitary output. The last convolutional layer (Conv5) contained a huge feature extracted from the mel-spectrogram. The training process is boosted through the adopted Reactivation Linear Unit (ReLU) activation function, after each convolutional layer's output. Usually, there are 3 fully connected layers, after the last convolutional layer of a typical VGGNet. However, in our fine-tuned model, the FC layer is nor longer needed as it has been replaced by a dimensionality reduction mechanism and a classifier to enhance the performance of our model. This is because of the poor performance of FC layer in classifying spectrogram image pixels.

The extracted feature maps from the stacked convolutional layer are followed by feature dimensionality reduction techniques; thereafter, the output is passed to a classification algorithm.

3.3 Principal Component Analysis (PCA)

From the standpoint of feature relevance and dimensionality reduction, PCA is an analytical technique that turns multiple feature indicators into a few exhaustive indicators [25]. To achieve the goal of reduction, it allows the original complex feature set to be replaced by many integrated factors that as closely as possible reflecting the information contained in the original feature set, while ignoring the irrelevant ones. Despite reducing dimensionality, PCA retains sufficient information about features that are necessary for simplifying classification and increasing the learning rate of an algorithm. PCA lowers the storage size and computational cost. PCA is employed in this work and is computed from feature matrix $\boldsymbol{Z}$ with $\boldsymbol{Ip}$ and $\boldsymbol{N}$ metric matrix define in $\mathbb{R}^p$ and $\mathbb{R}^n$ using equations (1–4) below:

$$N = (1/n)I_n \tag{1}$$

$$PCA(Z) = XS^{-1} \tag{2}$$

$$Z^T N S_u = \lambda u; u^T u = 1 \tag{3}$$

$$u = \frac{1}{\sqrt{\lambda}} Z^T N^{\frac{1}{2}} v \tag{4}$$

where X represent the matrix of n rows and p columns, u is a unit vector, S represent the diagonal matrix of standard deviation, I represents Identity matrix and λ represent the eigenvalue from an eigenvector.

3.4 Multilayer Perceptron (MLP) Classifier

MLP is a machine learning classification algorithm that utilizes backpropagation learning techniques to update its weight in an iterative manner. It is a basic sort of feed-forward neural network that comprises an input layer, a number of hidden layers, and an output layer. In our proposed model, the number of input neurons is determined by a number of features; an output of feature dimensionality reduction PCA. The average number of emotional labels in the speech corpus determines the number of hidden layers. The output of MLP represents the class label of emotion in the corpus, through a sigmoid activation function used as depicted in Eq. 5. Only one hidden layer with 200 neurons exists in the speech corpus (TESS) used in this work.

3.5 Random Forest Classifier (RF)

As an ensemble learning classifier, Random Forest found its application in classification and regression problems. It operates by forming a large number of decision trees in the training of the data, and the result is a mean indicator of

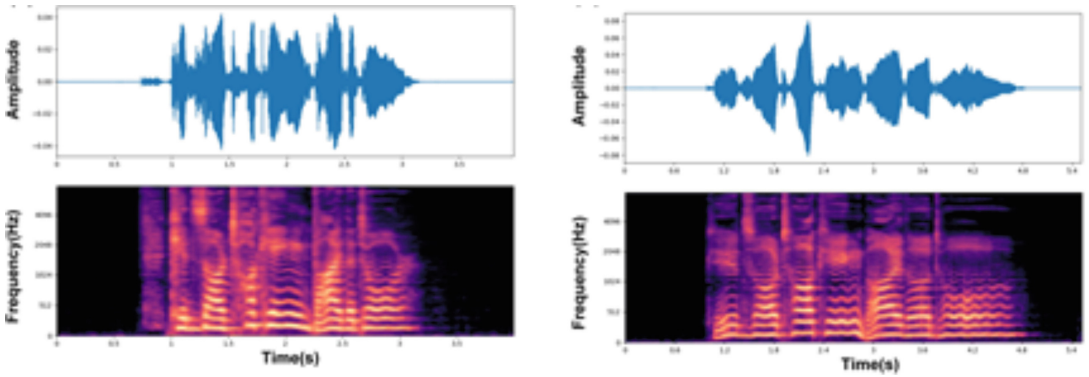


Fig. 2. Speech waveform and mel-spectrogram of emotional utterance

all the individual trees as well as the output of the class. With a replacement, an RF randomly samples each tree from the database, producing unique trees known as bagging. A random selection of features set for each tree is used in the RF classifier node splitting.

$$\sigma(Z_n) = \frac{1}{1 + e^{-Z_n}} \tag{5}$$

σ represents the sigmoid function, Z_n represents the sigmoid input weight, and e is the Euler's number (2.71828).

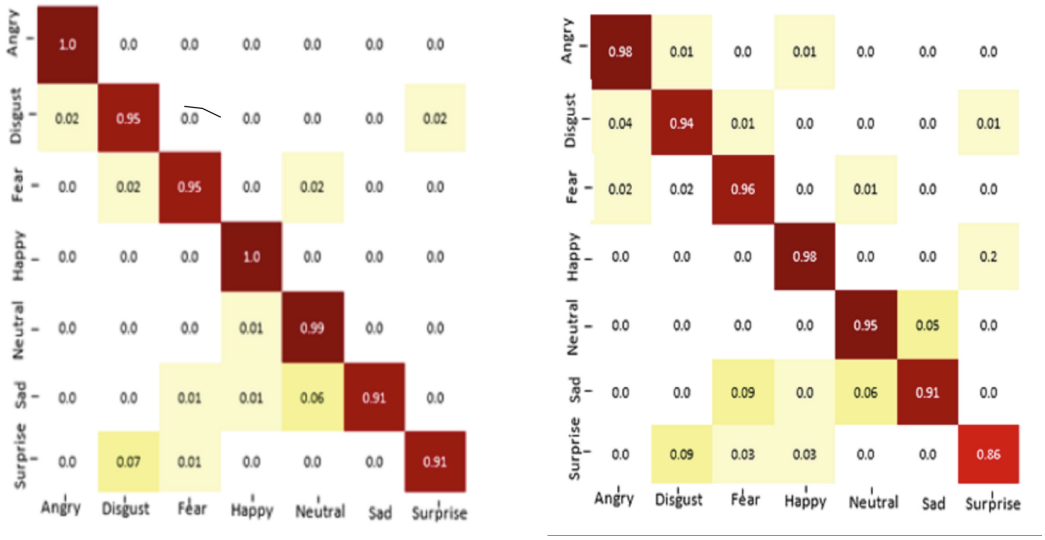
4 Experimental Results and Analysis

4.1 Datasets

The speech corpus (dataset) used in this work is from one of the publicly available speech emotion database known as Toronto English Speech Set (TESS) [26]. TESS is a standard synthetic speech emotion dataset that is simple to use for the classification of emotions because of its minimized reverberation effect. A total of 2,800 files of emotional utterances are contained in the TESS dataset. The audio sample had 2.5s of average time period and the wave signal is shown in Fig. 2.

4.2 Results and Discussion

The result obtained from our experiment on the TESS dataset achieved a classification accuracy of 96.1% for 7 different emotions, as shown by the confusion matrix in Fig. 3 below. A confusion matrix is employed in this work, to give a vivid picture of the classification rate for each emotional label. For the "angry" and "happy" emotional class, a 100% accuracy was achieved, which indicate the highest accuracy with our model as against other emotional classes. The rest of the emotional classes had an accuracy of classification not less than 90%, which



(a) DCNN-PCA-MLP (b) DCNN-PCA-RF

Fig. 3. Confusion matrix of emotion classification

further indicate the high performance of our proposed model compared to other state-of-the-art techniques.

Beside the confusion matrix, Return Operating Characteristics (ROC) curve for a multi-class task like speech emotion classification was also utilized to prove the efficiency and performance of our model. As shown in Fig. 4, the black dotted line denotes the threshold and all emotions classified were above the threshold. ROC for classes 0 to 6 represents each emotional class (angry, disgust, fear, happy, neutral, sad, and surprise). The Area under Curve (AUC) for each emotional class is not less than 0.98, an indication that performed better. AUC score ranges from 0 to 1, so the closer the value to 1, the better the performance.

4.3 Performance Evaluation

The accuracy, specificity, and unweighted average recall (UAR) of our model are presented in Table 1. While accuracy measures the ratio of true emotion classified to all emotional classes in the dataset, specificity measures the ratio of actual emotion to predicted emotion. The evaluation result in this table showed the outcome of our work when a separate experiment using another classification algorithm (Random Forest) was conducted. The performance of our proposed method was compared with other recent studies as indicated in Table 2 and our approach outperformed all of them in speech emotion classification.

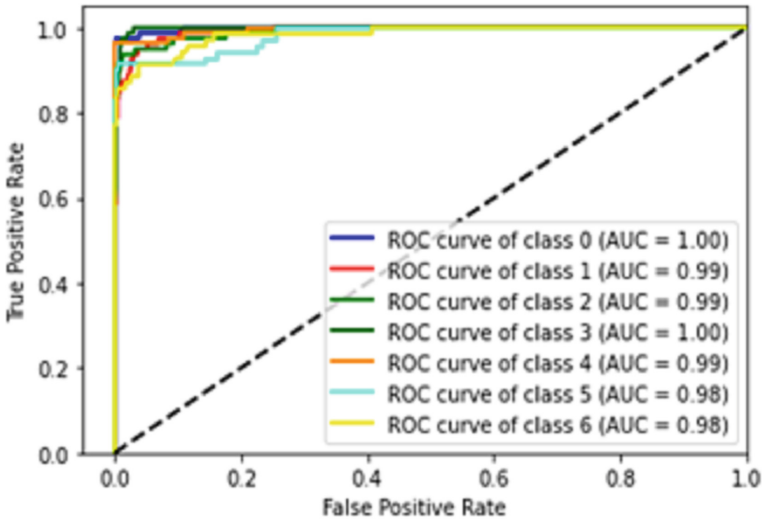


Fig. 4. Model Accuracy curve

Table 1. Performance evaluation

Experiment	Specificity	Accuracy	UAR
DCNN-PCA-MLP	**97.4%**	**96.1%**	**98.7%**
DCNN-PCA-RF	96.1%	94.0%	97.4%

Table 2. Performance comparison

Publication	Techniques	Dataset	Reported accuracy
Praseetha et al.(2018) [27]	DNN-GRU	TESS	89.96%
Venkataramanan et al.(2019) [28]	DNN-LSTM	TESS	70.00%
Krishnan et al.(2021) [29]	IMF-SVM, KNN	TESS	93.30%
Blumentals et al.(2022) [30]	LSTM-FCNN	TESS	86.02%
Proposed Model	**DCNN-PCA-MLP**	**TESS**	**96.10%**

5 Conclusion

In this work, the classification of speech emotion was performed using a deep transfer learning approach. The proposed DCNN-PCA-MLP model outperforms the recently adopted model on the same dataset with a noticeable gap. We achieved an accuracy of 96.1% and 97.4% specificity with the MLP classifier. The reason for this is because of the robustness of the adopted and fine-tuned VGGNet model. The dimensionality of feature reduction through PCA contributed to the performance of the proposed model and reduced the misclassification of emotion and elimination of highly correlated features. We also reaffirm that the application of deep transfer learning through a pre-trained DCNN for speech emotion classifica-

tion promises high efficiency in affective computing, even with the challenge of the limited dataset. However, the future direction of this work can be in using other pre-trained DCNN models to get an improved result.

References

1. Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., Montero, J., Fernández-Martínez, F.: A Proposal for Multimodal Emotion Recognition Using Aural transformer on RAVDESS. Appl. Sci. MDPI **12**, 327 (2022). https://doi.org/10.3390/app12010327
2. Firoozabadi, A., et al.: A multi-channel speech enhancement method based on subband affine projection algorithm in combination with proposed circular nested microphone array. Appl. Sci. MDPI **10**(3955), 455–464 (2021)
3. Leem, S., Fulford, D., Onnela, J., Gard, D., BussoAuthor, C.: separation of emotional and reconstruction embeddings on ladder network to improve speech emotion recognition robustness in noisy conditions. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 1, pp. 516–520 (2021). https://doi.org/10.21437/Interspeech.
4. Imani, M., Montazer, G.: A survey of emotion recognition methods with emphasis on e-learning environments. J. Netw. Comput. Appl. 147. Academic Press, (2019). https://doi.org/10.1016/j.jnca.2019.102423
5. Lieskovská, E., Jakubec, M., Jarina, R., Chmulik, M., Olave, M.: A review on speech emotion recognition using deep learning and attention mechanism. Electronics (Switzerland) MDPI **10**(10), 455–464 (2021). https://doi.org/10.10.3390/electronics10101163
6. Saad, F., Mahmud, H., Shaheen, M., Hasan, M., Farastu, P., Kabir, M.: is speech emotion recognition language-independent? analysis of english and bangla languages using language-independent vocal features, pp 1–9 (2021). http://arxiv.org/abs/2111.10776
7. Padmavathi, K., et al.: Transfer learning techniques for medical image analysis: a review. Biocybern. Biomed. Eng. **42**(1), 79–107 (2022). https://doi.org/10.1016/j.bbe.2021.11.004
8. Akçay, M., Oğuz, K.: Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers.Elsevier B.V. vol. 116, pp. 56–76 (2020). https://doi.org/10.1016/j.specom.2019.12.001
9. El Ayadi, M., Kamel, M., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. **44**(3), 572–587 (2011). https://doi.org/10.1016/j.patcog.2010.09.020
10. Kwon, S.: A CNN-Assisted Enhanced Audio Signal Processing. Sensors **20**(1), 183(2020)
11. Latif, S., Rana, R., Younis, S, Qadir, J., Epps, J.: Cross corpus speech emotion classification - an effective transfer learning technique (2018)
12. Farooq, M., Hussain, F., Baloch, N., Raja, F., Yu, H., Bin-Zikria Y. : Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. Sensors (Switzerland) **20**(21), 1–18 (2020). https://doi.org/10.3390/s20216008
13. Lech, M., Stolar, M., Best, C., Bolia, R.: Real-time speech emotion recognition using a pre-trained image classification network: effects of bandwidth reduction and companding. Front. Comput. Sci. **2**, 1–14 (2020). https://doi.org/10.3389/fcomp.2020.00014

14. Feng, K., Chaspari, T.: A siamese neural network with modified distance loss for transfer learning in speech emotion recognition. Sensors (2020). arXiv:2111.10776
15. Kamin, A., et al.: A light-weight deep convolutional neural network for speech emotion recognition using mel-spectrograms. In: Proceedings of 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP) (2019)
16. Padi, S., Sadjadi, S., Sriram, R., Manocha, D.: Improved speech emotion recognition using transfer learning and spectrogram augmentation. Sensors, pp. 645–652 (2021). https://doi.org/10.1145/3462244.3481003
17. Aggarwal, A., et al.: Two-way feature extraction for speech emotion recognition using deep learning, Sensors (Switzerland), **22**, 237 (2022). https://doi.org/10.3390/s22062378
18. Zhang, H., Gou, R., Shang, J., Shen, F., Wu, Y., Dai, G.: Pre-trained deep convolution neural network model with attention for speech emotion recognition. Front. Physiol. **12**, 1–13 (2021). https://doi.org/10.3389/fphys.2021.643202
19. Ortega, J., Cardinal, P., Koerich, A., Jun, L.: Emotion recognition using fusion of audio and video features. (2019). arXiv:1906.10623v1
20. Vatcharaphrueksadee, A., Viboonpanich, R.: VGG-16 and optimized CNN for emotion classification, 16(2), 10–15, (2020). https://ph01.tci-thaijo.org/index.php/IT-Journal/article/download/243769/165748/848686
21. Retta, E., Almekhlafi, E., Sutcliffe, R., Mhamed, M., Ali, H., Feng J. : Amharic speech emotion dataset and classification benchmark. (2022). arxiv:abs/2201.02710
22. Parra-Gallego, L., Orozco-Arroyave, J.: Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments. Digit. Signal Process. A Rev. J. **120**, 1–18 (2022). https://doi.org/10.1016/j.dsp.2021.103286
23. Alzubaidi, L., et al.: Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J. Big Data **8**(1), 1–74 (2021). https://doi.org/10.1186/s40537-021-00444-8
24. Pusarla, A., Singh, B., Tripathi, C.: Learning DenseNet features from EEG based spectrograms for subject independent emotion recognition. Biomed. Signal Process. Control, **74**(1), 103485 (2022). https://doi.org/10.1016/j.bspc.2022.103485
25. Jia, W., Sun, M., Lian, J., Hou, S.: Feature dimensionality reduction: a review. Complex Intell. Syst. (2022). https://doi.org/10.1007/s40747-021-00637-x
26. Pichora-Fuller, M., Kate, K.D.: Toronto emotional speech set (TESS), scholars portal dataverse, V1 (2020). https://doi.org/10.5683/SP2/E8H2MF
27. Praseetha, V., Vadivel, S.: Deep learning models for speech emotion recognition. J. Comput. Sci. **14**(11), 1577–1587 (2018). https://doi.org/10.3844/jcssp.2018.1577.1587
28. Krishnan, P., Joseph, A., Rajangam, V.: Emotion classification from speech signal based on empirical mode decomposition and non-linear features. Complex Intell. Syst. **7**(4), 1119–1934 (2021). https://doi.org/10.1007/s40747-021-00295-z
29. Venkataramanan, K., Rajamohan, H. : Emotion recognition from speech. audio and speech processing. pp 1–14 (2019). https://doi.org/10.48550/arXiv.1912.10458 arXiv:1912.10458
30. Blumentals, E., Salimbajevs, A., : Emotion recognition in real-world support call center data for latvian language. Jt. Proc. ACM IUI Work. Helsinki, (Finland) (2022).http://ceur-ws.org/Vol-3124/paper23.pdf

3.3 Robust Feature Selection-Based Speech Emotion Classification Using Deep Transfer Learning

3.3.1 Brief Review

This section is an extension to Paper 4 which presents the paper titled Robust Feature Selection-Based Speech Emotion Classification Using Deep Transfer Learning. It addressed the issue associated with speech utterance with over-blotted features and irregularities in spectral characteristics. In the paper, a methodology that is based on an efficient emotional feature selection technique is utilized in combination with a deep learning model. The paper made use of a convolutional neural network (CNN) that has already been trained to efficiently extract features from mel-spectrograms that are taken from speech signals. Effective feature extraction is achieved while minimizing computational costs by freezing a large portion of the CNN layers during training. The paper utilizes the Neighborhood Component Analysis (NCA) feature selection algorithm to minimize feature dimensionality and prevent misclassification. The problem of over-blotted features brought on by anomalies in spectral properties and irregularities in speech features is overcome in Chapter 7. At the top layer of the model, the paper used Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) classifiers to classify emotions after feature selection.

Paper status: Published in MDPI, Applied Sciences.

applied sciences

MDPI

Article

Robust Feature Selection-Based Speech Emotion Classification Using Deep Transfer Learning

Samson Akinpelu † **and Serestina Viriri** *,†

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban 4041, South Africa
* Correspondence: viriris@ukzn.ac.za
† These authors contributed equally to this work.

Abstract: Speech Emotion Classification (SEC) relies heavily on the quality of feature extraction and selection from the speech signal. Improvement on this to enhance the classification of emotion had attracted significant attention from researchers. Many primitives and algorithmic solutions for efficient SEC with minimum cost have been proposed; however, the accuracy and performance of these methods have not yet attained a satisfactory point. In this work, we proposed a novel deep transfer learning approach with distinctive emotional rich feature selection techniques for speech emotion classification. We adopt mel-spectrogram extracted from speech signal as the input to our deep convolutional neural network for efficient feature extraction. We froze 19 layers of our pretrained convolutional neural network from re-training to increase efficiency and minimize computational cost. One flattened layer and two dense layers were used. A ReLu activation function was used at the last layer of our feature extraction segment. To prevent misclassification and reduce feature dimensionality, we employed the Neighborhood Component Analysis (NCA) feature selection algorithm for picking out the most relevant features before the actual classification of emotion. Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) classifiers were utilized at the topmost layer of our model. Two popular datasets for speech emotion classification tasks were used, which are: Berling Emotional Speech Database (EMO-DB), and Toronto English Speech Set (TESS), and a combination of EMO-DB with TESS was used in our experiment. We obtained a state-of-the-art result with an accuracy rate of 94.3%, 100% specificity on EMO-DB, and 97.2%, 99.80% on TESS datasets, respectively. The performance of our proposed method outperformed some recent work in SEC after assessment on the three datasets.

Keywords: feature selection; speech emotion; classification; deep convolutional neural network; transfer learning

Citation: Akinpelu, S.; Viriri, S. Robust Feature Selection-Based Speech Emotion Classification Using Deep Transfer Learning. *Appl. Sci.* **2022**, *12*, 8265. https://doi.org/10.3390/app12168265

Received: 28 July 2022
Accepted: 12 August 2022
Published: 18 August 2022

1. Introduction

SEC occupies an important position in affective computing and plays a cogent role in human relationships. Psychologically, there is no better way: a relationship can neither be initiated nor sustained without adequate communication. However, communication takes place in many ways (verbal and non-verbal), but auditory speech communication has been the hub for all other forms. It is the nucleus to which all other means of human interaction revolve [1]. One major factor that determines human decisions is emotion [2,3]. Whether on an inter-personal or intra-personal basis, emotion influences why human beings act or react the way they act, informs their judgment, and even sometimes molds their opinions. Classification of emotion through human speech, therefore, is the act of recognizing emotional content from speech [4] as depicted in Figure 1. The growth of human–computer interaction in the last decade cannot be unconnected to SEC [5,6]. Apart from wider application in major fields and sub-fields, SEC is reshaping the nearby future next revolution of human–machine interaction as it is attracting the vast majority

of researchers worldwide. Mental disorders and depression-related diseases have been drawing great benefit from SEC coupled with the acceleration of patient treatment [7].

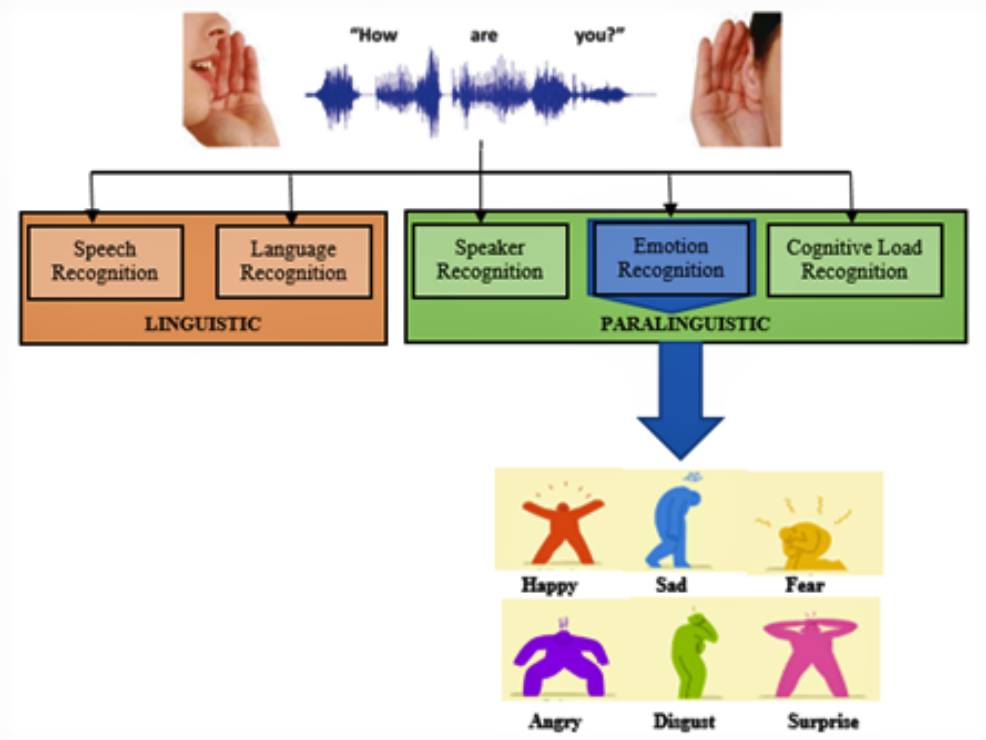


Figure 1. Emotion Classification Framework.

However, the curiosity of why speech emotion classification has been experiencing high processing cost and increased memory consumption with less accuracy has remained a great question among researchers today [8]. Presently, one lingering challenge identified so far is the lack of a sufficient annotated speech dataset, but then, several models have been proposed to mitigate this challenge and improve accuracy, as was obtainable in [9] using Surrey Audio-Visual Expressed Emotion (SAVEE), Ryerson Multimedia Laboratory (RML), and eNTERFACE'05 as the speech corpus. A 3D convolutional neural network (CNN) model with fully connected layers was developed for SEC. They attained a state-of-the-art performance, but the huge features [10] extracted without a corresponding feature selection method applied before actual classification tends to increase computational [11] complexity and reduce the accuracy rate.

By way of improving on the previous study, we proposed a novel and robust deep transfer learning approach for speech emotion classification with feature selection techniques. Transfer learning leverages on knowledge acquired from the source task and implements it on the target task [12]. It reduces computational complexity as it requires no training before its usage. Not all features from speech signals carry emotional content, and this is the reason why many models' failed to attain an optimal result. There is a need for a mechanism to subject extracted speech features [13,14] (spatial) to selection techniques for effectively identifying discriminating features that can increase accuracy while reducing computational cost.

Main Contributions

The specific contributions of this work are:

1. Applying an optimized deep transfer learning for feature extraction using the pre-trained model on ImageNet.

2. Adopting an efficient feature selection technique for selecting a discriminative feature for optimal classification of emotion, because not all features from a speech signal carry paralinguistic information relevant to accurate emotion classification.
3. Reducing misclassification of emotion and computational cost, while achieving state-of-the-art and improving accuracy on three speech datasets with two different classifiers.
4. Systematic performance evaluation and critical comparison of the model with other methods, which indicates a high efficiency of deep transfer learning with feature selection for SEC.

The remaining part of this work is arranged as follows. Related work with some useful background concepts is presented in Section 2. Section 3 deals with the proposed methodology of the work. In Section 4, the experimental flow with exact results achieved was presented. Section 5 presents the discussion of the obtained results and comparison with recent works that have been published. A conclusion with recommendation for future work is given in Section 6.

2. Literature Review and Related Works

From a psychological point of view, Refs. [15,16], human emotions are categorized into six major groups, which fall under positive and negative emotion, respectively. However, this can also be sub-divided into valence or arousal. Valence emotion determines the extent to which emotion is either positive or negative, while arousal indicates the level of intensity. The description of these basic enotions are shown in Table 1. These emotions goes through the process of vectorization in speech data processing (encoding) before emotional features extraction and classification can be performed.

Table 1. Basic Emotion Description.

Emotion	Description	Expression	Class
Happiness	A state of pleasantness characterized by joyful mood	Upbeat	Positive
Sadness	A transitional emotion state usually characterized by grieving felling	Leathery and dampened	Negative
Fear	Plays a vital role in survival	Rapid heratbeat	Positive/Negative
Disgust	Results from unpleasant taste or smell	Retching/nauseating	Negative
Anger	A state of hot temper, hostility and aggression	Yelling	Negative
Surprise	Usually occurs when the unexpected happens	Screaming and hilarious	Positive/negative

For more than half a decade, several deep transfer learning techniques rooted in deep convolutional neural network (DCNN) have been applied in speech recognition with the utmost focus on emotion classification from speech. Transfer learning minimized computational cost while maintaining an accurate extraction of features from speech signals [17,18]. Despite several outstanding records of successful application of convolutional neural network (CNN), the SEC domain is still facing the intense challenge of optimal performance with minimal complexity. As the literature is saturated with many of the applications of deep transfer learning techniques for an improved performance in speech emotion classification, this work has reviewed some published works.

Obviously, the recent advancement in Artificial Intelligence and affective computing has conspicuously made the deep transfer learning model very popular in major fields, such as image segmentation, visual objects recognition systems, facial emotion recognition

systems, age and gender estimation, and speech recognition systems, which is due to its extraordinary performance in prediction and low error rate.

Vryzas et al. [19] applied a pretrained DCNN-based transfer learning for speech emotion classification. A multi-user speech dataset was used in training their model, while fine tuning took place on the speaker-dependent speech emotion dataset. The first 11 layers of the model that served in feature learning were frozen from being trained according to the structural setup of the model with utmost the aim of fine-tuning the model on the target's dataset input. Mustaqeem and Kwon, Ref. [20] investigated the possibility of adopting a double stream approach of CNN and DCNN for extracting spectral and spatial features for an improved performance in SEC. Aside from the fusion of dual features techniques, considerable effort was made by applying a feature selection algorithm for the effective classification of emotion. Mustaqeem and Kwon Ref. [21] proposed an attention-based convolutional neural network with a fully connected layer trained in an end-to-end manner for the effective classification of speech emotion. An optimized DNN was proposed by Aouani and Ayed, Ref. [22] for real-time speech emotion classification. The optimization was carried out using stochastic gradient descent.

Learning the most salient features from a speech spectrogram image through a light-weight CNN model for an improved performance in SEC was presented by Anvarjon et al. [23]. The model was also trained in an end-to-end mode, with an enhanced pooling strategy and fewer parameters. A deep transfer learning model that is based on AlexNet with correlation-based feature selection techniques was presented by Farooq et al. [24] in order to investigate the overall impact of the feature selection algorithm on speech emotion classification. Haider et al. [25] evaluated three different feature selection techniques, namely, Relief feature Infinite Latent Feature Selection, and a generalized Fisher score for speech emotion recognition, along with the recently proposed automatic feature selection method. They employed three datasets from three different languages to affirm that higher accuracy can only be achieved through a reduced feature set.

Zhang et al. [26] adopted transfer learning using a pretrained model on ImageNet (AlexNet) to extract segment-level features and an attention mechanism for learning emotionally rich features. Their model also combine Bi-directional Long Short-Term Memory (BiLSTM) for effective speech emotion classification using two different datasets (EMO-DB and IEMOCAP). Feng and Chaspari Ref. [27] proposed a transfer learning technique with distance loss for speech emotion classification. They fine-tuned a siamese nueral network in order to achieve a state-of-the-art result. An experiment was carried out on the eNTERFACE, RAVDESS, and CREMA-D datasets, respectively. Spectrogram features generated from speech signals were augmented in Padi et al. [28] and fed as input into a transfer learning model (ResNet) with a significant pooling layer (statistics) for speech emotion classification. They prevented overfitting and improved classification accuracy in spectrogram augmentation techniques.

Joshi et al. [29] proposed a deep neural network BiLSTM for emotion classification using a cross dataset that was rooted in EEG (electroencephalography). Out of the three datasets (DEAP, SEED, IDEA) used in their work, two were for training, while one (IDEA) was for testing the performance of their proposed method. Four classifiers (SVM, KNN, MLP and Deep RNN) were trained with four speech features (Hjorth parameters, DE, LF-DE and PSD), and the results of their experiments indicated that DNN parameters optimization can positively enhance the performance of the emotion classification system. In addition, the deep BiLSTM model improved emotion recognition from EGG through the extraction of discriminative features and temporal dependencies of speech samples. However, the datasets utilized were insufficient for a typical Deep BiLSTM network model learning, which eventually hampered the accuracy (59%) recorded. The most popular datasets used within the speech emotion domain consist of synthetic data because of the difficulty and security issues associated with real-world speech-based datasets. However, Blumentals and Salimbajevs [30] carried out a speech emotion classification task using a real-world dataset (from phone calls). In their work, a deep learning network made

up of two LSTM layers and fully connected layers with an Adam optimizer for back propagation was utilized. Their model was first tested on three popular datasets which included IEMOCAP, RAVDESS and TESS with accuracy of 60.4% and 86.02% before the real-world data obtained from a call center was used. The real dataset was based on Lavitian Language, and a model accuracy of 52.48% was recorded. It very evident from this work that the result is lower than state-of-the-art; moreover, the testing accuracy on the real-world dataset used was not stated.

Really, it is very cumbersome to build a standardized speech corpus for emotion recognition, and this is a major difficulty confronting speech-based emotion classification tasks. Some of these deep learning methods adopted so far require extensive training, which in most cases increases computational cost. However, the accuracy of speech emotion classification must need to be improved upon, as proposed in our study. Additionally, determining which set of features carries emotional information through the feature selection algorithm can lessen misclassification and increase model performance. In Table 2, a summary of the reviewed literature is given.

Table 2. Summary of Related Works.

Author	Techniques	Dataset	Accuracy/ Unweighted Average Recall (UAR)	Number Emotion
Joshi et al. [29]	DNN-BiLSTM	EEG	58.44%	-
Blumetals and Salimbajos [30]	DNN-LSTM	TESS, RAVDESS, IEMOCAP	86.02%	7, 4
Padi et al. [28]	ResNet	IEMOCAP	66.02%	4
Feng and Chaspari [27]	SiameseNet	RAVDESS	32.8% (UAR)	7
Zhang et al. [26]	DCNN-BiLSTMwA	EMODB, IEMOCAP	87.86%, 68.5%	7, 4
Haider et al. [25]	AFS-SVM	EMODB, EMOVO, SAVEE	76.9%, 41.0%, 42.4% (UAR)	6
Farooq et al. [24]	DCNN-CFS	EMODB, SAVEE, IEMOCAP, RAVDESS	95.10%, 82.10%, 83.30%, 81.30%	7, 4
Anvarjon et al. [23]	CNN	IEMOCAP, EMODB	77.01%, 92.02%	7, 4
Aouani and Ayed [22]	DNN	EMODB	96.97%	7
Mustaqeem and Kwon [21]	DNN-INCA	EMODB, SAVEE, RAVDESS	95.0%, 82.0%, 85.0%	7
Vryzas et al. [19]	CNN-VGGNet	Personalized(Web Crowd Sourcing)	69.9%	-

3. Methods and Techniques

A unique approach of DNN-NCA-MLP (Deep transfer learning with Neighborhood Component Analysis and Multi-Layer Perceptron) for speech emotion classification is introduced. At first, our proposed method carried out preprocessing on a raw speech signal to generate suitable input (Mel-spectrogram) for the model. The structural framework of the proposed model is shown in Figure 2 below. The input image generated from the audio signal is fed into the pretrained DCNN model for robust utterance-level [31] features extraction. Thereafter, a feature selection technique is applied to carefully determine relevant and pertinent features for the accurate classification of emotion with the chosen classifier at the top layer of the model.

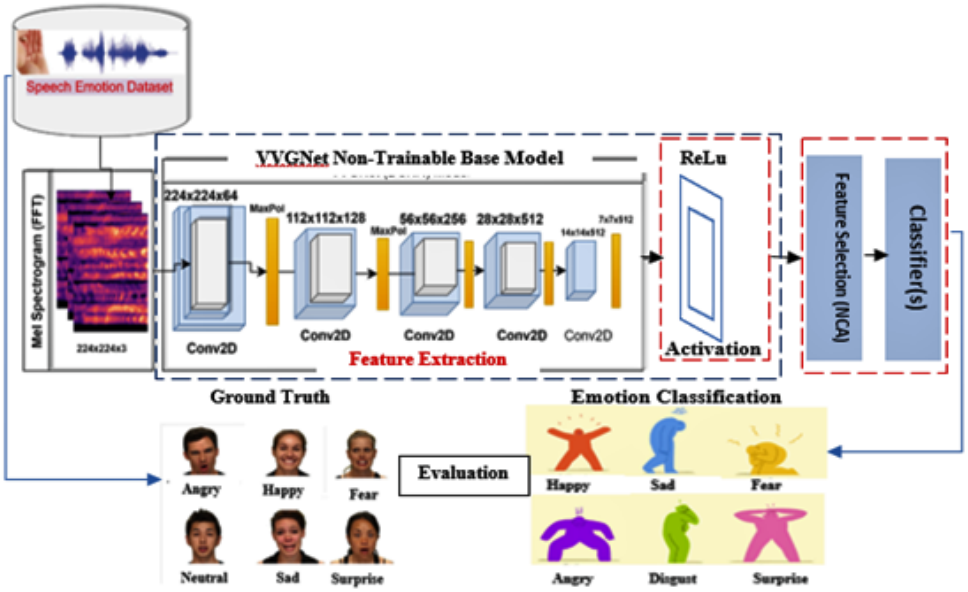


Figure 2. Structure of the Proposed Model.

3.1. Speech Data Pre-Processing

Any successful SEC model begins with speech database preparation. Most speech emotion corpus are influenced by speakers' age, gender and cultural background, among others. It then becomes necessary to diligently carry out adequate preparation of the dataset. Unlike other image processing sub-fields, applying transfer learning to a speech-based task requires substantive data preprocessing in order to make the input image suitable for the deep learning network. Specifically, speech signals carry information such as the sampling ratio, frequency and speed that must be enhanced before their usage for training and the corresponding feature map.

The required input to our proposed model is log mel-spectrogram. This is extracted from the speech signal. Mel-spectrogram has proven to be efficient and rich in emotional content for speech emotion classification. In generating the mel-spectrogram, a series of operations such as pre-emphasis, framing and windowing must need to be performed. The speech signal is first subjected to pre-emphasis for amplification of the frequency. Because of the continuous nature of the speech signal, framing is performed to disintegrate the signal into a fixed-length size, after which a hamming window of 20 ms size function is applied on the frame with 10 ms shift, as computed in Equation (1)

$$w(n) = 0.54 - 0.46cos\left(\frac{2\pi n}{M-1}\right) 0 \leq n \leq M-1 \quad (1)$$

where M represents the window size for the hamming function $w(n)$.

The mel-filter bank used in this work is 40, while the sample rate is 16 kHz and FFT (Fast Fourier Transform) is 512. A speech signal (x) of length (N) can be converted into a signal in the frequency domain (X) using the FFT technique as computed in Equation (2). The size of FFT corresponds to the size of the last convolution layer of our pretrained model. FFT is adopted [32] in generating the complete mel-spectrogram of the speech sample. As part of preprocessing, the mel-spectrogram generated using FFT is resized to the input specification (size) of our proposed model, which is 224 × 224 × 3. This represents the size (height × width) and the number of channels, respectively.

$$X(t) = \Sigma_{r=1}^{N-1}\left(x[r]W\cos\frac{n\pi x}{L} + b_n\sin\frac{n\pi x}{L}\right) \quad (2)$$

where $x(r)$ represents a signal, W represents an N-point window function and $h = 0, 1, 2, \ldots, N-1$.

3.2. Feature Extraction with Deep Convolutional Neural Network

This section gives the detail performance of our proposed model with respect to feature learning from the original input mel-spectrogram. An utterance level feature is extracted using VGGNet (Visual Graphics Group Network) as the base model for the deep transfer learning model. As shown in Figure 2, the model consists of five, 2D convolutional layers in series of $224 \times 224 \times 3$, $112 \times 112 \times 128$, $56 \times 56 \times 256$, $28 \times 28 \times 512$ and $14 \times 14 \times 512$ kernel size. Altogether, four pooling layers were used, with the last pooling layer having a size of $7 \times 7 \times 512$. A dropout layer and two dense layers with ReLu (Reactivation Linear Unit) were used as the topmost level of the feature extraction segment. The dropout layer is a hyperparameter that can be used to prevent overfitting [33], and ReLu is the activation function. The pretrained model also carries an internal linear activation unit. The model is properly fine-tuned in order to extract all features (global and spatial) before the selection of high-level features in the next section. We utilized sparse-categorical cross-entropy and an Adams optimization algorithm in compiling our model.

The number of trainable parameters for our model is obtained from the fully connected layers (dense layer), because our base model does not need training. The dense layer consists of a total number of neurons from the current layer (nc) and previous layer (np). The product of these two mini parameters plus one (bias term) gives the value for the dense layer, which represents the total number of trainable parameters for our model. For instance, the single dense layer of our model (Desnse-1), $nc = 7, np = 512$; therefore, total parameters = $7 \times (512 + 1) = 3591$.

3.3. Neighborhood Component Analysis (NCA) Feature Selection

After the extraction of features with the DCNN segment of our proposed model, it is necessary to select relevant features in order to prevent misclassification and increase the accuracy of the entire model. Feature selection [34] occupies a significant position in this work, as it determines the level of performance. Many models other than the proposed model could have yielded an optimal result in speech emotion classification, but the absence of an appropriate feature selection algorithm incapacitated those models. It is very essential to know that huge features from speech data can be extracted, but not all these features are relevant [35] for emotion classification. The more irrelevant the features in the classifier, the greater the likelihood of misclassification and reduction in the accuracy and efficiency of the SEC model. However, even though feature selection is important, a careful choice of which feature selection techniques perform best and under which condition is another dicey question. The choice must be made depending on the model architectural framework. In this work, NCA, a feature selection method based on distance, has been chosen for selecting discriminative features and its performance in maximizing classification accuracy. The normalized equation for computing the NCA feature selection for a multi-class SEC problem is given in Equation (3).

$$S = \{(x_i, y_i), i = 1, 2, \ldots, n\} \tag{3}$$

where n is the number of observations, $x_i \in \mathbb{R}^p$ denotes the feature vectors, $y_i \in \{1, 2, \ldots, c\}$ denotes class labels of emotional features, and c is the number of emotional class labels. To simplify, the goal is to provide relevant features (from the feature extraction segment of the model) for the classifier f: $\mathbb{R}^p \rightarrow \{1, 2, \ldots, c\}$, which takes a feature vector and makes classification $f(x)$ for the ground truth (label) y of x. The NCA algorithmic [36] procedure is defined in Algorithm 1.

Algorithm 1: NCA Feature Selection Procedure

1: **procedure** NCAFS (T, α, σ, λ, η) ▷ T: set of training,α: initial step length, σ: kernel width, λ: parameter for regularization, η: small positive integer constant;
2: Initialization: $w^0 = (1, 1, \ldots, 1), \epsilon^0 = -\infty, t = 0$
3: **while** $[\epsilon^t - \epsilon^{t-1}] < \eta$
4: for $i = 1, \ldots, N$ do
5: Compute p_{i_j} and p_i using $\mathbf{w}^t$ with respect to (**2**) and (**3**)
6: for $l = 1, \ldots, d$ do
7: $\Delta l = 2\Big(\frac{1}{\sigma}\Sigma_i\Big(p_i\Sigma_i \neq jp_{ij}|x_{il} - x_{jl}| - \Sigma_j y_{ij}p_{ij}|x_{il} - y_{il}|\Big) - \lambda\Big)w_l^t$
8: t = t + 1
9: $w^t = w^{t-1} + \alpha\Delta$
10: $\epsilon^t = \xi(w^{t-1})$
11: **if** $\epsilon^t > \epsilon^{t-1}$ **then**
12: $\alpha = 1.01\alpha$
13: **else**
14: $\alpha = 0.4\alpha$
15: **wend**
16: $w = w^t$
17: **return**

3.4. Classifiers

The classifiers [4] employed in this work are Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) . For a multi-class problem such as SEC, the choice of an efficient classifier after feature selection is crucial. The classifier is found at the top-most layer of our proposed model. The fully connected layer and softmax function in the normal pretrained model have been replaced with two different classifiers for an improvement performance of speech emotion classification.

3.4.1. Multi-Layer Perceptron

The Multi-Layer Perceptron classifier is a feedforward neural network model that has three nodes of input, output and hidden layers [37]. Depending on the implementation, MLP can take one or more hidden layers, but in this work, one hidden layer is used for all the datasets (TESS, EMO-DB and TESS-EMO-DB). The input layers consist of a certain number of neurons which represent the total number of features from the feature selector. TESS has 2280 neurons and EMO-DB has 240 neurons. The output layer represents the actual number of emotional categories as present in the ground label of speech samples. The sigmoid activation function is used for our MLP classifier in determining the output of the neuron. After each iteration, the weights of the connections are continuously adjusted as part of the MLP's training process.

3.4.2. Support Vector Machine

SVM is a sort of classifier that builds a hyperplane and optimizes the margin between the two classes in order to achieve classification purpose [38]. It has been proven to be efficient in many classification tasks. In this work, SVM is utilized as a second classifier too, and the result obtained is compared with our own MLP.

4. Experimental Result and Analysis

4.1. Emotion Datasets

We tested our proposed model on three datasets, which are TESS, EMO-DB and TESS-EMO-DB (hybridized) consisting of two languages (English and German).

4.1.1. TESS

Many researchers have used the Toronto Emotional Speech Set, a publicly accessible speech emotion database, to classify emotions. The speech samples were captured in 2010 at Northwestern University Auditory Laboratory [39]. Two actresses were requested to recite a few of the 200 words during the impromptu event, and their voices were recorded. During the scene, seven different emotions were recorded (happy, angry, fear, disgust, pleasant, surprise, sad and neutral). There are 2800 files in all that show human emotion.

4.1.2. EMO-DB

Seven acted emotional states are represented in the EMO-DB speech dataset: anger, disgust, boredom, joy, sadness, neutral, and fear. Ten actors who are natural German speakers contributed 535 emotional utterances in the German language. It was initiated by the ICS, Technical University, Berlin, German [40]. Five of them are actresses who are women, while the other five are actors who are men. The speech files are 3 s long on average, and they were initially collected at a sampling rate of 48 kHz before being resampled to 16 kHz.

4.1.3. TESS-EMO-DB

The combination of the two datasets, a formed and hybridized dataset, in our model yielded different results. Unlike the seven emotions in the two separate speech corpus, six emotions were used for classification.

4.2. Experiment Setup

In this work, the experiment was conducted on a Python 3.9 environment with an Intel Core i7 processor, 8 GB of RAM, and a 64-bit operating system. Additionally, other third-party libraries and deep learning tools (such as Tensorflow, Numpy, and audio processing) were used. Since our model's input layer needs to be $224 \times 224 \times 3$, preprocessing of the audio sample was initially necessary. The speech signal had to be converted to a log-mel spectrogram and resized to fit the model's specifications. The mel spectrogram feature was extracted from the original audio data using the FFT method. The dataset is then divided into a training and testing set at a ratio of 80:20. The data for the test and training sets were both standardized to pixel values between 0 and 1. The basic implementation parameters used are shown in Table 3.

Table 3. Implementation Parameters.

Parameter	Value
Optimizer	Adam
Loss Function	Sparse Categorical Cross-Entropy
Activation Function	Softmax
Environment	CPU
Output Classes	7
Learning Rate	0.001
Maximum Epochs	50
MLP Hidden Neurons	20
SVM Kernel Function	Linear

5. Result and Discussion

The result obtained in this work proved the efficiency of our proposed model. In the first stage of our experiment, we carried out training of the additional (topmost) layers of the feature extraction segment of our proposed model, which comprises a dropout layer, two dense layers, and an activation function, since all layers of our based transfer learning network are non-trainable. We obtained a significant accuracy and loss curve (94%, 14%) after training and validation was carried out with the dataset as indicated in Figure 3.

A normalized confusion matrix which depicts a holistic insight of various results obtained from our experiments with different classifiers is shown in Figures 3–5, respectively. Looking at the confusion matrix, for the TESS dataset with the MLP classifier, 100% accuracy was recorded for angry, disgust, fear and happy emotion, while the surprise emotion yielded the lowest accuracy of 89%, and others were above 90%. With the SVM classifier, the minimum accuracy increased to 90%.

On the EMO-DB dataset, our model achieved 94%, 89%, 98%, 94%, 91%, 98% and 99% accuracy in classification for angry, boredom, disgust, fear, happy, neutral and sad emotion using the MLP classifier. When SVM was introduced, there was a slight improvement in recognition accuracy, as the three emotions (disgust, neutral and sad) recorded 99% accuracy while the happy emotion recorded the lowest accuracy of 87%. As with the combined dataset of TEES-EMO-DB, the highest accuracy of 99% was recorded for the sad, neutral and disgust emotions with the MLP classifier, while the happy emotion recorded the lowest accuracy of 87%, as depicted in Figure 6. For the SVM classifier, only the sad emotion produced the highest accuracy of 99%, while others were 90% and above.

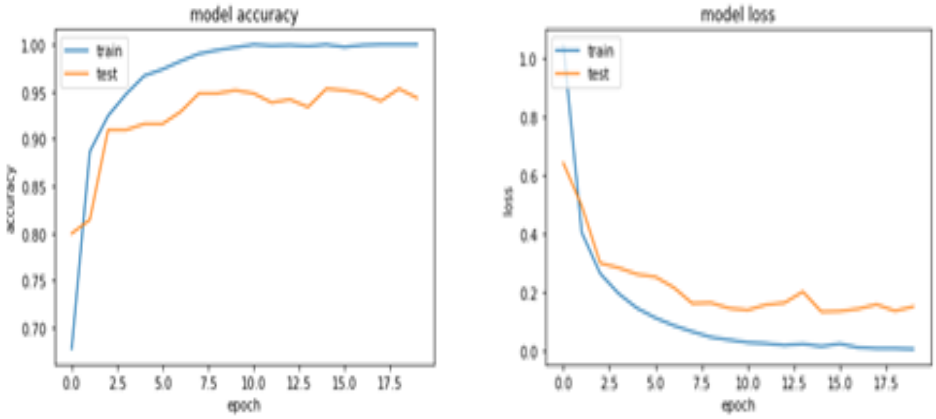
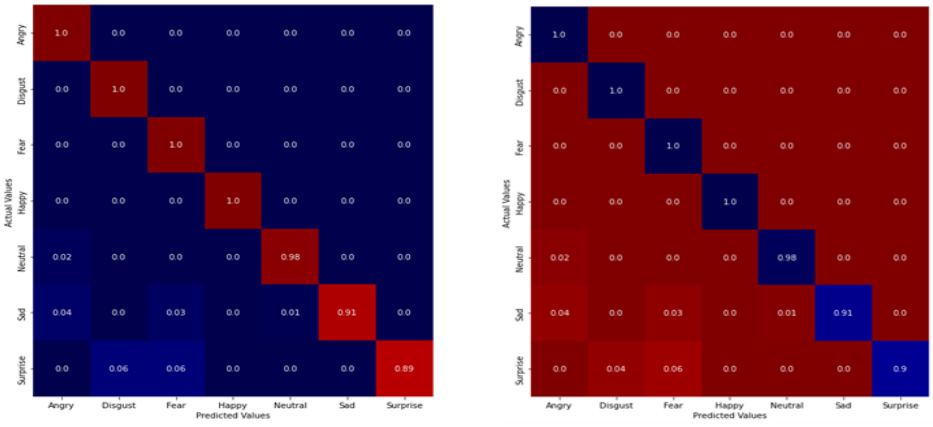


Figure 3. Proposed model's accuracy and loss performance curves for both training and validation.

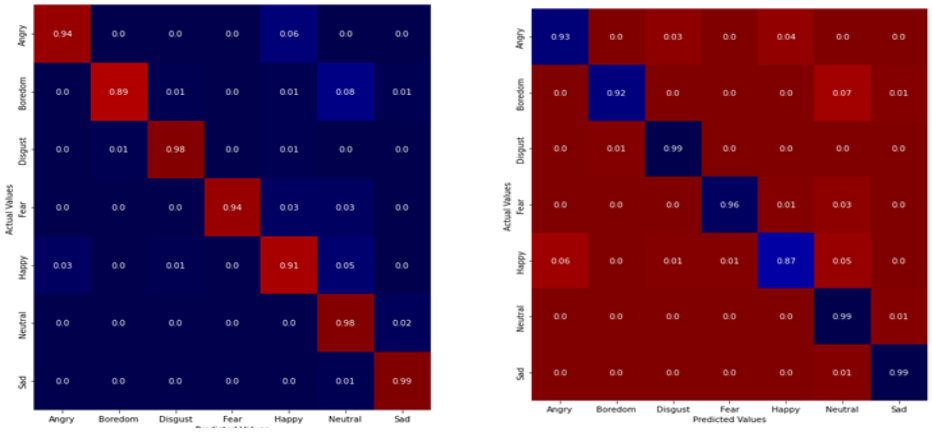


(**a**) (**b**)

Figure 4. Confusion matrix of emotion classification on TESS dataset. (**a**) DCNN-NCA-MLP. (**b**) CNN-NCA-SVM.

As we analyze further, the classification report comprising accuracy, F1-score, specificity and sensitivity evaluation metrics was employed in order to fully grasp how efficient our proposed method performed in its discriminatory ability to learn features for speech emotion classification. As shown in Table 4, we achieved the highest accuracy of 97.2% on the TESS dataset with the SVM classifier, while the lowest accuracy of 94.3% was recorded

from the EMO-DB dataset with the MLP classifier. The highest specificity and sensitivity of 100% was recorded with experiments on the two datasets for both classifiers, while 98.8% specificity was recorded on the combined dataset.

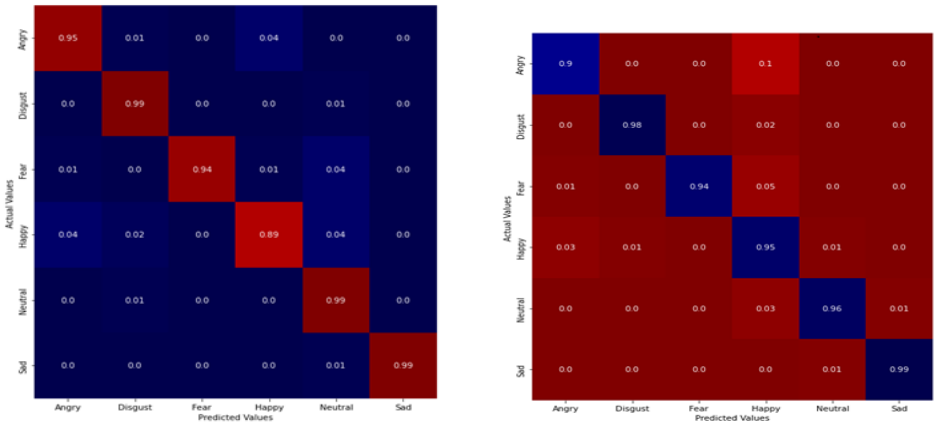


(**a**) (**b**)

Figure 5. Confusion matrix of emotion classification on the EMO-DB dataset. (**a**) DCNN-NCA-MLP. (**b**) CNN-NCA-SVM.

Table 4. Performance Evaluation.

MLP/SVM	TESS	EMO-DB	TESS-EMODB
F1-score (%)	97.20, 97.68	94.40, 94.40	95.70, 94.90
Sensitivity (%)	99.04, 100	100, 100	99.01, 98.95
Specificity (%)	98.90, 99.80	100, 100	100, 98.85



(**a**) (**b**)

Figure 6. Confusion matrix of emotion classification on TESS-EMODB dataset. (**a**) DCNN-NCA-MLP. (**b**) CNN-NCA-SVM.

As with F1-score, we achieved a state-of-the-art result of 94.4%, 95.7% and 97%, respectively after evaluation on the three datasets. A return of operating curve (ROC) was adopted to present the graphical outlook of classification results of six emotions from the hybrid dataset. The ROC analysis result for six emotions on the combined dataset is illustrated in Figure 7. On the curve, each class represents an emotion: angry, sad, disgust, happy, fear and neutral with class 0 to 5, respectively. From the curve, the highest classification performance of Average Area Under the Curve (AUC) of 98% was achieved on classes 3 and 5 (neutral and disgust), while class 4 (fear) yielded the lowest AUC of 50%.

However, none of the classes gave a classification outcome that is below the threshold (black dotted line). A chart indicating performance classification of emotion is also illustrated in Figure 8. The bars in represent experiments with MLP and SVM classifiers. From the chart, with SVM, five emotions comprising disgust, fear, happy, neutral and sad yielded accuracy results above 90%, while five emotions (sad, neutal, fear, disgust and angry) also yielded 90% and above accuracy.

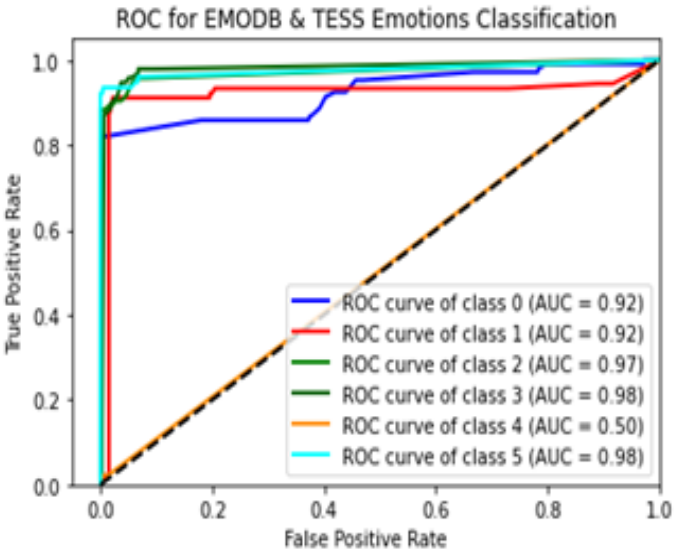


Figure 7. ROC curve of classification.

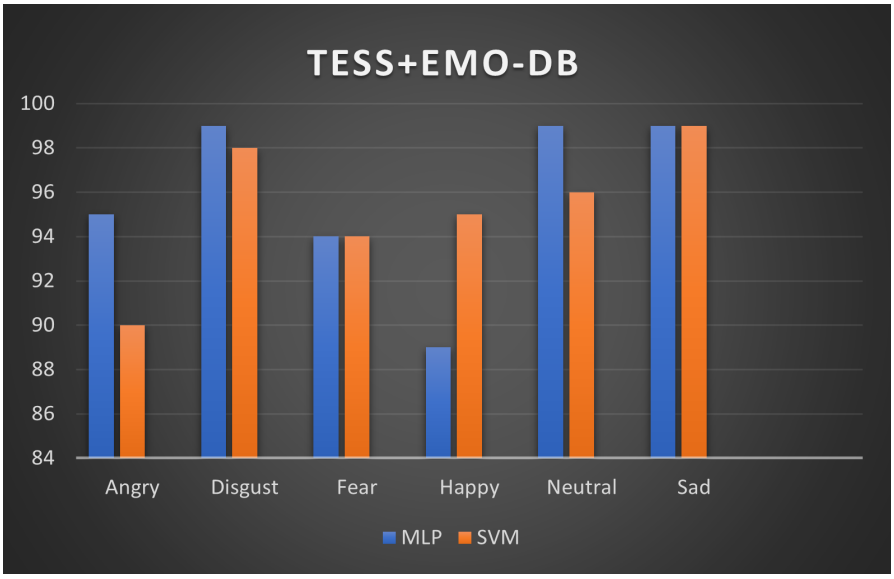


Figure 8. Performance chart on TESS-EMODB with the two classifiers.

5.1. Performance Evaluation

The accuracy, specificity, and unweighted average recall (UAR) of our model are presented in Table 4. While accuracy measures the ratio of true emotion classified to all emotional classes in the dataset, specificity measures the ratio of actual emotion to predicted emotion. The evaluation result in this table showed the outcome of our work when a separate experiment using another classification algorithm (Random Forest) was conducted. The performance of our proposed method was compared with other studies in

recent times, as indicated in Table 5, and our approach outperformed all of them in speech emotion classification. Additionally, a drastic reduction in complexity (at feature extraction and selection level) has been achieved, while the accuracy of emotion classification has been increased in our model as against what is obtained in the literature [2,38–42].

Table 5. Performance Evaluation.

Dataset	Reference	Methods	Reported Accuracy
TESS	(2018) [41]	DNN-GRU	89.96%
"	(2019) [42]	DNN-LSTM	70.00%
"	(2021) [2]	IMF-SVM, KNN	93.30%
"	(2022) [30]	LSTM-FCNN	86.02%
"	**Proposed**	**DCNN-NCA-MLP**	**96.10**%
EMO-DB	(2019) [43]	DCNN	80.79%
"	(2020) [44]	CNN-BiLSTM	85.50%
"	(2020) [22]	CNN	89.02%
"	(2021) [31]	DCNN-LSTM-Attention	87.86%
"	(2021) [20]	CNN-Attention	93.00%
"	(2021) [45]	LSTM-CNN	93.34%
"	**Proposed**	**DCNN-NCA-MLP**	**94.90**%

5.2. Performance Comparison

Our proposed model outperformed other existing methods when evaluated on the TESS and EMO-DB datasets, as shown in Table 3.

6. Conclusions and Future Work

In this work, a deep transfer learning approach with a feature selection algorithm that can select relevant features for speech emotion classification from a pool of extracted features through a deep convolutional neural network has been presented. It is obvious after a series of experiments have been carried out that the proposed method has improved accuracy and reduced the rate of misclassification as well as computational complexity using state-of-the-art emotional speech corpora. In achieving an optimum performance, the mel-spectrogram was extracted from raw speech signals through preprocessing and FFT. Thereafter, the mel-spectrogram was reshaped to match the requirement of the DCNN model for feature extraction. NCA feature selection was then applied in selecting the most salient and discriminating features that are relevant to emotion classification while reducing the workload of the classifier. With few hyperparameter tunning, our model achieved significant improvement in accuracy compared to existing methods as we carried out our experiment on the TESS, EMO-DB and combined (TESS+EMO-DB) dataset using two classifiers (MLP and SVM). Overall accuracies of 97.20%, 94.82% and 95.77% were obtained after we examined our method on TESS, EMO-DB and TESS+EMO-DB. The work has indicated that the proposed model is efficient for speech emotion classification. However, future experiments can be conducted using the cross-language speech corpus and testing with other classifiers to see their performance. We are also looking forward to adopting LSTM as an additional layer to our pretrained neural network and utilizing a multiple feature selection algorithm for an improved performance.

Author Contributions: Conceptualization, S.A. and S.V.; Methodology, S.A.; Software, S.A.; Validation, S.V.; Formal analysis, S.V.; Investigation, S.A.; Resources, S.V.; Data curation, S.A.; Writing—original draft preparation, S.A.; Writing—review and editing, S.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive any funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pusarla, A.N.; Singh, B.A.; Tripathi, C.S. Learning DenseNet features from EEG based spectrograms for subject independent emotion recognition. *Biomed. Signal Process. Control* **2022**, *12*, 74. [CrossRef]
2. Krishnan, P.; Joseph, A.; Rajangam, V. Emotion classification from speech signal based on empirical mode decomposition and non-linear features. *Complex Intell. Syst.* **2021**, *7*, 1919–1934. [CrossRef]
3. Jiang, W.; Wang, Z.; Jin, J.S.; Han, X.; Li, C. Speech emotion recognition with heterogeneous feature unification of deep neural network. *Electronics* **2019**, *19*, 2730. [CrossRef]
4. Lieskovská, E.; Jakubec, M.; Jarina, R.; Chmulík, M.; Olave, M. A review on speech emotion recognition using deep learning and attention mechanism. *Electronics* **2021**, *10*, 1163. [CrossRef]
5. Van, L.; Le Dao, T.; Le Xuan, T.; Castelli, E. Emotional Speech Recognition Using Deep Neural Networks. *Sensors* **2022**, *22*, 1414. [CrossRef]
6. Topic, A.; Russo, M. Emotion recognition based on EEG feature maps through deep learning network. *Eng. Sci. Technol. Int. J.* **2021**, *24*, 1442–1454. [CrossRef]
7. Moine, C.L.; Obin, N.; Roebel, A. Speaker attentive speech emotion recognition: Proceedings of the Annual Conference of the International Speech Communication Association. *Interspeech* **2021**, *1*, 506–510. [CrossRef]
8. Sattar, R.; Bussoauthor, C. Emotion Detection Problem: Current Status, Challenges and Future Trends Emotion Detection Problem. In *Shaping the Future of ICT: Trends in Information Technology, Communications Engineering, and Management: Global Proceedings Repository—American Research Foundation*; ICCIIDT: London, UK, 2020.
9. Hajarolasvadi, N.; Demirel, H. 3D CNN-based speech emotion recognition using k-means clustering and spectrograms. *Entropy* **2019**, *22*, 479. [CrossRef]
10. Wang, Y.; Boumadane, A.; Heba, A. A Fine-tuned Wav2vec 2.0/HuBERT Benchmark for Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding. *arXiv* **2021**, arXiv:2111.02735.
11. Luna-Jiménez, C.; Kleinlein, R.; Griol, D.; Callejas, Z.; Montero, J.; Fernández-Martínez, F. A Proposal for Multimodal Emotion Recognition Using Aural transformer on RAVDESS. *Appl. Sci.* **2022**, *12*, 327. [CrossRef]
12. Bashath, S.; Perera, N.; Tripathi, S.; Manjang, K.; Dehmer, M.; Streib, F.E. A data-centric review of deep transfer learning with applications to text data. *Inf. Sci.* **2022**, *585*, 498–528. [CrossRef]
13. Aggarwal, A.; Srivastava, A.; Agarwal, A.; Chahal, N.; Singh, D.; Alnuaim, A.A.; Alhadlaq, A.; Lee, H. Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning. *Sensors* **2022**, *22*, 2378. [CrossRef]
14. Badshah, A.M.; Rahim, N.; Ullah, N.; Ahmad, J.; Muhammad, K.; Lee, M.Y.; Kwon, S.; Baik, S.W. Deep features-based speech emotion recognition for smart affective services. *Multimed. Tools Appl.* **2019**, *78*, 5571–5589. [CrossRef]
15. Cowen, A.S.; Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl. Acad. Sci. USA* **2017**, *38*, E7900–E7909. [CrossRef] [PubMed]
16. Oaten, M.; Stevenson, R.J.; Case, T.I. Disgust as a disease-avoidance mechanism. *Psychol. Bull.* **2009**, *135*, 303–321. [CrossRef]
17. Elshaer, M.E.A.; Wisdom, S.; Mishra, T. Transfer Learning from Sound Representations for Anger Detection in Speech. *arXiv* **2019**, arXiv:1902.02120.
18. Nguyen, D.; Sridharan, S.; Nguyen, D.T.; Denman, S.; Tran, S.N.; Zeng, R.; Fookes, C. Joint Deep Cross-Domain Transfer Learning for Emotion Recognition. *arXiv* **2020**, arXiv:2003.11136.
19. Vryzas, N.; Vrysis, L.; Kotsakis, R.; Dimoulas, C. A web crowdsourcing framework for transfer learning and personalized Speech Emotion Recognition. *Mach. Learn. Appl.* **2021**, *6*, 100–132. [CrossRef]
20. Mustaqeem; Kwon, S. Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network. *Int. J. Intell. Syst.* **2021**, *36*, 5116–5135. [CrossRef]
21. Mustaqeem; Kwon, S. Att-Net: Enhanced emotion recognition system using lightweight self-attention module. *Appl. Soft Comput.* **2021**, *102*, 107101. [CrossRef]
22. Aouani, H.; Ayed, Y.B. Speech Emotion Recognition with deep learning. *Procedia Comput. Sci.* **2021**, *176*, 251–260. [CrossRef]
23. Anvarjon, T.; Mustaqeem; Kwon, S. Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. *Sensors* **2020**, *20*, 5212. [CrossRef] [PubMed]
24. Farooq, M.; Hussain, F.; Baloch, N.; Raja, F.; Yu, H.; Bin-Zikria, Y. Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. *Sensors* **2020**, *20*, 6008. [CrossRef]
25. Haider, F.; Pollak, S.; Albert, P.; Luz, S. Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods. *Comput. Speech Lang.* **2020**, *65*, 101–119. [CrossRef]
26. Zhang, H.; Gou, R.; Shang, J.; Shen, F.; Wu, Y.; Dai, G. Pre-trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition. *Front. Physiol.* **2021**, *12*, 643202. [CrossRef] [PubMed]
27. Feng, K.; Chaspari, T. A Siamese Neural Network with Modified Distance Loss For Transfer Learning in Speech Emotion Recognition. *arXiv* **2006**, arXiv:2006.03001.

28. Padi, S.; Sadjadi, S.O.; Sriram, R.D.; Manocha, D. Improved Speech Emotion Recognition using Transfer Learning and Spectrogram Augmentation. In Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), Montréal, QC, Canada, 18–22 October 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 645–652. [CrossRef]
29. Joshi, V.; Ghongade, R.; Joshi, A.; Kulkarni, R. Deep BiLSTM neural network model for emotion detection using cross-dataset approach. *Biomed. Signal Process. Control* **2022**, *73*, 103407. [CrossRef]
30. Blumentals, E.; Salimbajevs, A. Emotion Recognition in Real-World Support Call Center Data for Latvian Language. In Proceedings of the ACM IUI Workshops 2022, Helsinki, Finland, 22 March 2022.
31. Yao, Z.; Wang, Z.; Liu, W.; Liu, Y.; Pan, J. Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. *Speech Commun.* **2021**, *120*, 11–19. [CrossRef]
32. Atila, O.; Şengür, A. Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition. *Appl. Acoust.* **2021**, *182*, 108260. [CrossRef]
33. Uddin, M.Z.; Nilsson, E.G. Emotion recognition using speech and neural structured learning to facilitate edge intelligence. *Eng. Appl. Artif. Intell.* **2020**, *94*, 103775. [CrossRef]
34. Akçay, M.B.; Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **2020**, *166*, 56–76. [CrossRef]
35. Zhang, S.; Li, C. Research on Feature Fusion Speech Emotion Recognition Technology for Smart Teaching. *Hindawi Mob. Inf. Syst.* **2022**, *2022*, 7785929. [CrossRef]
36. Yang, W.; Wang, K.; Zuo, W. Neighborhood component feature selection for high-dimensional data. *J. Comput.* **2022**, *7*, 162–168. [CrossRef]
37. Ba'abbad, I.; Althubiti, T.; Alharbi, A.; Alfarsi, K.; Rasheed, S. A Short Review of Classification Algorithms Accuracy for Data Prediction in Data Mining Applications. *J. Data Anal. Inf. Process.* **2021**, *9*, 162–174. [CrossRef]
38. Wanni, T.; Gunawan, T.; Qadri, S.; Kartiwi, M.; Ambikairajah, E. A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access* **2021**, *9*, 47795–47814. [CrossRef]
39. Dupuis, K.; Kathleen Pichora-Fuller, M. Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set. *Can. Acoust.-Acoust. Can.* **2012**, *39*, 182–183.
40. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W. A database of German emotional speech. In Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2012; pp. 1517–1520. [CrossRef]
41. Praseetha, V.M.; Vadivel, S. Deep learning models for speech emotion recognition. *J. Comput. Sci.* **2021**, *14*, 1577–1587. [CrossRef]
42. Venkataramanan, K.; Rajamohan, H.R. Emotion Recognition from Speech. *Audio Speech Process.* **2021**, 1–14. [CrossRef]
43. Meng, H.; Yan, T.; Yuan, F.; We, H. Speech Emotion Recognition from 3D Log-Mel Spectrograms with Deep Learning Network. *IEEE Access* **2019**, *7*, 125868–125881. [CrossRef]
44. Mustaqeem; Kwon, S. Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. *IEEE Access* **2020**, *36*, 79861–79875. [CrossRef]
45. Yahia Cherif, R.; Moussaouni, A.; Frahta, N.; Berimi, M. Effective speech emotion recognition using deep learning approaches for Algerian dialect. In Proceedings of the International Conference of Women in Data Science at Taif University, WiDSTaif, Taif, Saudi Arabia, 30–31 March 2021; pp. 1–6. [CrossRef]

3.4 Speech Emotion Classification Using Attention-Based Network and Regularized Feature Selection

3.5 Brief Review

This paper entitled Speech Emotion Classification Using Attention Based Network and Regularized Feature Selection is an extension of this chapter's previous paper. The paper addresses the challenge of selecting the most discriminating and high-level emotionally relevant features for efficient speech-emotion classification. The paper utilized an attention-based network in conjunction with a pre-trained convolutional neural network (CNN) to extract global features from speech signals that are emotionally rich as well as local features. By concentrating on prominent features, the attention mechanism improves the model's capacity to identify pertinent emotional cues. To enhance feature selection and boost classification accuracy, regularized neighbourhood component analysis (RNCA) is incorporated at the third stage of the architecture. Prioritizing emotionally relevant features, this strategy helps select the most discriminative features while minimizing misclassification. Three key classifiers (SVM, MLP, and Random Forest) were employed at the top layer of the system to comprehensively assess the model on the speech emotion dataset for the classification of emotion. The result from this paper model (Attention-based DCNN+RNCA+RF) achieved a notable classification accuracy that outperforms state-of-the-art SEC approaches. The evaluation of the model showed that, when it came to classifying emotion from auditory speech, the attention mechanism and feature selection were consistent with human behavioural tendencies.

Paper status: Published in Springer, Scientific Reports.

scientific reports

Check for updates

OPEN

Speech emotion classification using attention based network and regularized feature selection

Samson Akinpelu[1,2] & Serestina Viriri[1,2]✉

Speech emotion classification (SEC) has gained the utmost height and occupied a conspicuous position within the research community in recent times. Its vital role in Human–Computer Interaction (HCI) and affective computing cannot be overemphasized. Many primitive algorithmic solutions and deep neural network (DNN) models have been proposed for efficient recognition of emotion from speech however, the suitability of these methods to accurately classify emotion from speech with multi-lingual background and other factors that impede efficient classification of emotion is still demanding critical consideration. This study proposed an attention-based network with a pre-trained convolutional neural network and regularized neighbourhood component analysis (RNCA) feature selection techniques for improved classification of speech emotion. The attention model has proven to be successful in many sequence-based and time-series tasks. An extensive experiment was carried out using three major classifiers (SVM, MLP and Random Forest) on a publicly available TESS (Toronto English Speech Sentence) dataset. The result of our proposed model (Attention-based DCNN+RNCA+RF) achieved 97.8% classification accuracy and yielded a 3.27% improved performance, which outperforms state-of-the-art SEC approaches. Our model evaluation revealed the consistency of attention mechanism and feature selection with human behavioural patterns in classifying emotion from auditory speech.

Human has various ways of exhibiting their emotion, which has placed them at the highest level of civilization among other creatures. These expressions can take the form of speech, facial, gestures, and other physiological modes. However, interaction and relationships among individuals are best sustained through communication from human speech. Human speech carries huge para-linguistic[1] content that can reveal the state of emotion, both in direct and indirect communication. Therefore, speech emotion classification has been occupying a key position in advancing affective computing and speech research domain. Besides, unlike other methods of recognizing emotion, speech emotion can be said to reveal 90% of the intent of the speaker without pretence, hence, the reason why it is sporadically attracting researchers within the last decade.

In SEC, the cultural and racial background may have a significant impact, but the ground truth remains that emotion is universal. Because of peculiarities associated with the speech emotion domain, efforts have been made by professionals to generate a standardized synthetic dataset (emotional corpus) that had been useful for conducting research on emotion classification[2]. Among these corpora are IEMOCAP (Interactive and Diadic Motion Capture), TESS (Toronto English Speech Set), RAVDESS (Rayson Visual Emotion Speech Set), EMOVO, etc and their performances concerning speech emotion classification has been yielding appreciable result, even when sometimes compared with real world dataset. These datasets came in different languages (English, Spanish, German, Chinese)[3]. Speech emotion classification has its application in customer support management, self-driving cars, psycho-medicine, e-learning, etc. Its importance in human-computer interaction cannot be overemphasized. Gordon[4] opined that affective behaviour may serve as a precursor to the emergence of mental health conditions like depression and cognitive decline and may aid in the development of therapeutic tools for automatically identifying and tracking the progress of diseases.

Classical techniques of classifying emotion in the past follows the extraction of primitives, acoustic features and low-level detectors (LLD), from raw speech[5]. These features (pitch, energy, etc) represents frame-level features and speech analysis on it, do generate another level of features (Utterance-level). Thereafter, the concatenation of these feature in vector form will be fed into a machine learning algorithm also referred to as classifiers in this context, for actual classification of emotion. Support Vector Machine (SVM), Gaussian Mixture Model (GMM),

[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban 4000, South Africa. [2]These authors contributed equally: Samson Akinpelu and Serestina Viriri. ✉email: viriris@ukzn.ac.za

Hidden Makov Model (HMM) and K-Nearest Neighbour (KNN) are popular classifiers[6–8]. Figure 1 shows a classical structure of emotion recognition.

Though these approaches have proven to be efficient in their capacity, however, they are bewildered with salient challenges that rendered them unsuitable in achieving state-of-the-art result for SEC.

The focus of this study is to enhance and improve performance of speech emotion classification through attention-based network and feature selection techniques. To the best of our knowledge, this is the first-time feature selection is to be fused with attention layer of high dimensional features extracted from deep convolutional neural network, for accurate emotion classification (Attention based DCNN+RNCA+RF). We utilized TESS dataset in this study as a standard speech emotion corpus which captures seven classes of emotions express by human. The main contributions of this study are:

1. To experiment the efficiency of attention mechanism and regularized feature selection (Regularized Neighbourhood Component Analysis) techniques for speech emotion classification. A pretrained transfer learning network is set up as the based model. The feature selection neutralizes additional parameter weight added by attention layer and thereby minimize complexity.
2. To propose an Attention-based DCNN+RNCA+RF. After exploratory and thorough experiment with three different classifiers, our model achieved 97.8% accuracy on TESS dataset.

The remainder of the article is arranged as follows. An overview of related works is presented in Section "Review of related works". The proposed technique and methods are described in Sect. "Methods and techniques". Results and discussions are given in Sect. "Experimental results and discussion", while Sect. "Conclusion" is the conclusion and future recommendation for further study.

Review of related works

The classification of emotion has its history traced to psychological submission[9,10] where human emotion are grouped into six main classes (Sadness, Happiness, Anger, Disgust, Surprise and Fear). However, affective computing cannot be based on this primitive divisions, as computers are not perceiving moods, but they are interpreting them as a set of sequence of technical parameters, that are captured from the audio decoding process. Therefore, speech emotion classification requires efficient learning of paralinguistic information that can mitigates misclassification of emotion. The machine learning classifiers were first explored for SEC before the application of convolutional neural network models. The shortcomings of conventional classification approaches have paved ways for Convolutional Neural Networks (CNNs)[11,12] and LSTM networks[13]. Occasionally, these two combined to form a robust model[14] which have been widely employed in sequence modelling and its associated domain. A feature selection-based CNN was utilized by Farooq et al.[15], for combating the artificial design influence which hampered accurate description of speakers' emotional condition. Hajarolasvadi & Demirel[16] proposed 3D-CNN for speech emotion classification based on overlapping frames segregation and MFCC features extraction. A 10-fold cross-validation parameter was used in their evaluation on three publicly available speech corpora, which were Ryerson Multimedia Laboratory (RML), Survey Audio-Visual Expressed Emotion (SAVEE) and eNTERFACE'05. The convolutional model achieved 81.05% accuracy on six emotions classes. Deep Belief Network (DBN) and SVM was proposed in Zhu et al.[17] for extracting acoustics features, MFCC and zero-crossing rate were employed before emotion classification. Wang et al.[18] combined Deep Neural Network and Extreme Learning machine (ELM) for speech emotion classification through the encoding of speech features (pitch and formants) and segmentation of audio feature vectors.

However, conventional CNN performs woefully in high-dimensional speech features extraction. This, and many more shortcomings, paved the way for the introduction of recurrent neural network (RNN) model. It was a

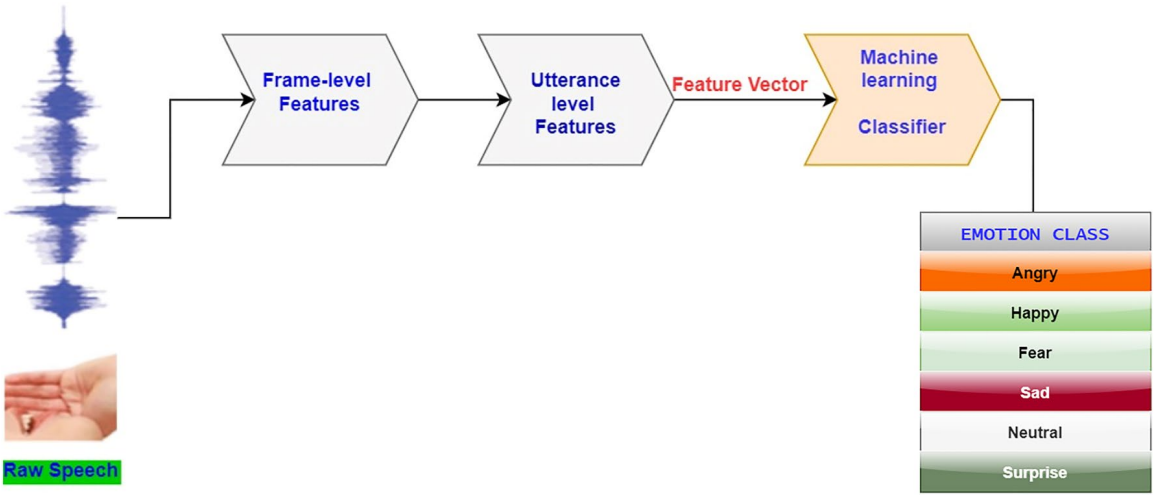


Figure 1. Conventional speech emotion classification system.

great milestone improvement over CNN in speech emotion classification, because it addresses the failure of CNN in time-series data extraction. RNN has a hidden layer in its structure that updates the output value with respect to time on constant basis[19]. Kerkeni et al.[20] proposed a RNN for speech emotion classification through analysis of speech signal using Teager-Kaiser Energy Operator (TKEO) combine with empirical mode decomposition (EMD). After extraction of speech cepstral features, SVM classifier was utilized for multi-classification of emotion. They achieved 91.16% on Berlin and Spanish based dataset. Nevertheless, RNN also suffers from dependency (long-term) and gradient descent problems. In some studies, CNN and RNN were combined to form a hybrid CRNN (Convolution Recurrent Neural Network) model to enhance speech emotion classification[21].

As RNN is not isolated from its own limitations and by way of proffering quick fix to the issues peculiar to it, Long-Short-Term-Memory (LSTM) was proposed by Hochreiter & Schmidhuber[22] and its combination with convolutional neural network has yielded a notable improvement. LSTM is a variant of RNN consisting of feedback connections for dependency learning in sequence prediction. A 1D and 2D CNN was combined with LSTM for SEC, which resulted in an appreciable accuracy of 82.4% with EMO-DB speech corpus by Zhao et al.[23]. Puri et al.[24], proposed a hybridized LSTM, CNN and DNN approach for speech emotion classification. MFCC and mel-spectrogram were fed into eight contiguous 2D convolutional sequential neural network layers of their model. RAVDESS dataset was used, but there was no accuracy of emotion recognition reported. Besides, their technique is expensive to train because of the huge convolutional layers adopted. LSTM has a key component called forget gate and research has proven that it has high probability of forgetting emotional feature, while it focuses on the most recent ones and this hampered its efficiency within SEC domain.

Recent advancement in deep learning coupled with incessant search for a way of improvement and addressing the age long challenges in SEC made Bahdanau et al.[25], to introduce attention network which is able to sieve out irrelevant information peculiar to speech data and concentrate on emotional rich information. Attention mechanism has been successfully adapted to other object recognition discipline with a notable improvement in models' performance. An attention-based network was adopted in the work of Qamhan et al.[26], where an accuracy of over 60% was achieved on IEMOCAP dataset. Attention models emulate the human way of focusing on important features for the recognition of an object.

Three-dimensional attention-based CRNN was used by Chen et al.[27] to choose discriminative features for speech emotion classification. Their proposed model's input layer accepted a Mel-spectrogram with delta-deltas. The employed delta-deltas reduced the intrusion of unimportant elements that can result in subpar classification performance, while keeping vital emotional data. Finally, a mechanism for attention that could take salient aspects into account was adopted. With an accuracy report of 82.82% on EMO-DB and 64.74% on the IEMOCAP speech dataset, their experiment's outcome was supported the efficacy of attention technique for emotion classification.

Zhao et al.[28]. utilized attention-based model comprises Bidirectional LSTM, a Fully Connected Networks (FCN) for learning spatio-temporal emotional features and machine learning classifier for speech emotion classification. In the same vain, the author in Du et al.[29]. utilized attention-based model and 1Dimensional CNN for SEC. Softmax activation function was used at the top layer of their model after feature extraction. A cross-modal SEC was carried out in Seo and Kim[30] using Visual Attention Convolutional Neural Network (VACNN) in partitioning the spectral feature from dataset. Combining speech dataset with text and video requires special techniques in extracting features for efficient prediction of emotion. In Zhang et al.[31], the author applied 5 attention heads mechanism for multimodal speech emotion classification. Their novel model achieved 75.6% on IEMOCAP dataset.

Zhang et al.[32] applied Deep convolutional Neural Network and attention-based network for emotion classification. In their method, a pre-trained DCNN was used as a based model in extracting segment-level features, before the introduction of Bidirectional LSTM for higher-level emotional features. Thereafter, an attention layer was introduced at the top layer of their model, with the utmost focus on features that are relevant to emotion recognition. Their model evaluation achieved UAR of 87.86% and 68.50% respectively on EMODB and IEMOCAP dataset. However, their experiment did not reflect the influence of speech enhancement carried out on raw speech. They augmented the speech corpus used through speed adjustment at varying time-step before the extraction of spectral features was fed into DCNN. Chen et al.[33] proposed self and global attention mechanism in determining the impact of the attention model on speech emotion classification. Their state-of-the-art approach achieved an accuracy of 85.43% on EMO-DB speech corpus. Their model was built using a sequential network, which requires more computing resources to train. In this paper, two pre-trainned DCNN model are used with attention model and regularized feature selection for SEC. More often than not, many researchers focused on the efficiency of attention mechanism as weight calculator in sequence representation Zhao et al.[34], however, our proposed model has revealed that the performance of attention-based network is increased when co-join with regularized feature selection for SEC. Nevertheless, this paper concludes with an opportunity for future research in the use of attention mechanism and feature selection to improve the accuracy of classification (Fig. 2).

Methods and techniques

A general description of the model proposed is given in this section. As a classification problem, speech emotion is categorized rather than dimensional representations[35]. It can be defined as follows, $D = (X, z)$, where X are input from the acoustic features and z is dimensional output equivalent to the emotion classification. Also, a function $D = f : X \rightarrow z$ representing emotional features is to be found before its classification.

This study proposed a unique framework for speech emotion classification using attention-based mechanism on pretrained DCNN with regularized feature selection (RNCA) algorithm, as shown in Fig. 3. There are four main phases in our model for speech emotion classification which includes, efficient pre-processing (pre-emphasis) of raw speech from TESS speech corpus, feature learning and extraction, feature selection and emotion classification. As noted in the literature[36] that the performance of any SEC model rests heavily on

107

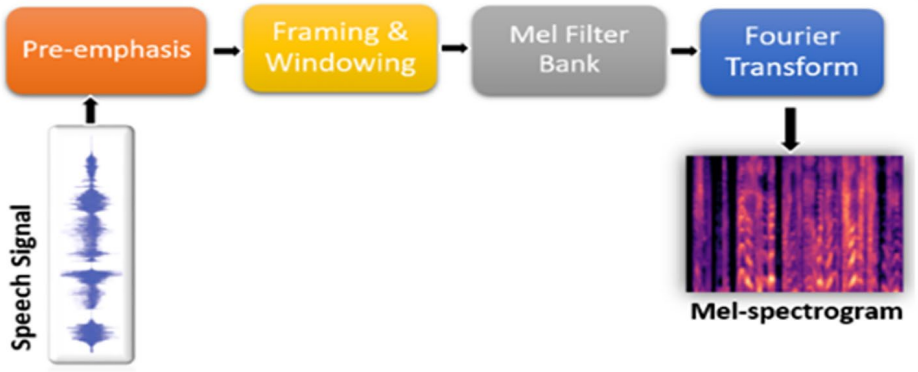


Figure 2. Structure of mel-spectrogram extraction.

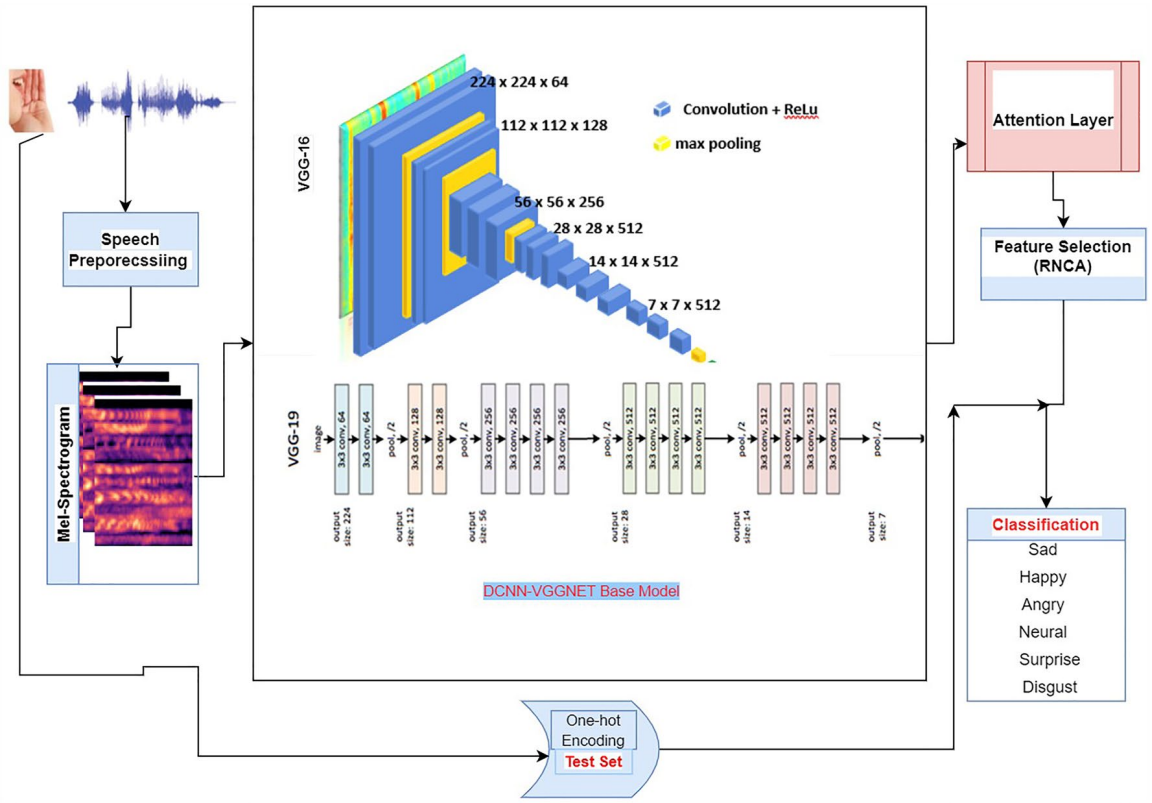


Figure 3. Proposed model architecture.

dataset pre-processing carried out. In this work, we extracted log mel-spectrogram with three channels (weight, height and input channel) from original speech database containing WAV files. Three channel mel-spectrogram usually comprises of the number of mel-filter banks (in terms of frequency dimension), frame number and the number of channel. The number of channels used for this paper is 3. Three different colours are used to indicate the magnitudes of the Short-Term Fourier Transform (STFT) in a three-channel mel-spectrogram. The low (below 500 Hz-blue), mid (between 500 Hz - 2 kHz-yellow), and high (above 2 kHz-red) frequency ranges of the audio signal are typically represented by the channels, which can offer a more intuitive visual form of the spectral content of the audio signal. The latter is used in this paper. Mel-spectrogram has been widely used[37,38] in speech-related task, and the reason is not far-fetched from the fact that it's representation involves time and frequency of speech signals.signal.

At the pre-emphasis stage, the amplification of speech signals (x) to high frequency[39] is performed through a pre-emphasis filter using Eq. (1), where $s(t)$ represent the speech audio signal before pre-emphasis. We utilized 64 mel-filter banks with 64 frames content window. To obtain the standard frame segment length, we processed $655ms(10ms \times 63 + 25)$ fragments, however, a frame segment over $250ms$ has been confirmed[40] to possess enough paralinguistic information rich enough for emotion classification. The speech signal framing adopted ensures the breaking down of the speech signal into segments of fixed-length. Because the length of human

speech varies, framing is required to maintain the size of the voice. The hamming window function of 25ms length and 10 ms shift was applied to frames as computed in Eq. (2), where S represents the size of the window $w(n)$. This is illustrated in Fig. 2.

$$y(t) = s(t) - \alpha s(t-1), 0.9 \leq \alpha \leq 1.0 \tag{1}$$

$$w(n) = 0.5 - 0.5cos[2\pi \frac{n}{S-1}]0 \leq n = S-1 \tag{2}$$

The FFT (Fast Fourier Transform) is applied to produce a three channel mel-spectrogram suited as input to our model from raw speech signal with a sample frequency rate of 16kHz. This mel-spectrogram can be represented as $M, M \in R^{K \times L \times C}$ where the total number of the filter bank is denoted[32] by K in terms of dimension of the frequency, L denotes the length of the segment and the number of channels is C.

Feature extraction. In this research study, two pre-trained DCNN model serve as our based model (VGG16 and VGG19). We experimented with both pre-trained network on our attention mechanism to establish which one yields better classification performance accuracy after feature selection. We leverage on the weight of these two networks being already trained on ImageNet. Therefore, the convolutional layers comprised of our based model are frozen from training. The input to our model is reshaped from the original $64 \times 64 \times 3$ to $224 \times 224 \times 3$, as the required input size to the base model of VGGNet.This is achieved using a built-in python library called OpenCV and a bilinear interpolation approach. The base model comprises five convolutional layers with ReLu (Reactivation Linear Unit) activation function for extracting segment-level features from the input mel-spectrogram. A drop-out layer is utilized to prevent overfitting. The output from based model feature extraction is also reshaped to make it suited for the attention layer in extracting high-level emotional features before it is fed into RNCA for eventual feature selection. This is carried out by The block diagram in Fig. 4 depicts the structure of DCNN phase of our model. The pooling layer adopted is max-pooling. This layer performs the function of aggregating the feature sample from the several convolutions of 2D convolutional layers and produces a unified output for the next layer. No fully connected layer was used in the base model.

Attention layer. Attention mechanism application in computer vision has contributed immensely to the task of image recognition[41]. It mimics the human mode of paying a closer look at what are relevant information that may contribute to their opinion or conclusion on what they see and hear.

In the speech emotion task, the role of the attention network cannot be overlooked, as it carefully concentrates the focus of the model on the frame segment with much emotional content. The attention mechanism lowers the training time[42] and ensures concentration on features with much emotional information, which can increase model performance. Silent and semi-silent frames are eliminated at the attention layer, as this has a tendency of impairing and distorting the model accuracy. In other words, attention gives insight into the behavioural performance of the deep learning model as it calculates weight from feature representation from the previous layer. The Eq. (3) and (4) indicate how the attention mechanism utilized in this work is computed. Given $X = (x_1, x_2 ... x_n)$ as the output of features from a convolutional layer.

$$\alpha_i = \frac{exp(\mu^T x_i)}{\sum_{j=1}^{J} exp(\mu^T x_i)} \tag{3}$$

$$Y = \sum_{j=1}^{I} \alpha_i x_i \tag{4}$$

where alpha α_i represents the weight of the attention network, μ and X are the output of feature representation from the attention layer. At first, the weight of the attention α_i is calculated, and it is obtained from Eq. (3) (softmax function) through the training process. Y is got from the weighted sum of X, as deeper features at the utterance level. The attention mechanism has proven to be of tremendous help in generating more distinctive features for SEC. The attention layer is responsible for dynamically highlighting and weighting various input feature components according to their applicability to the emotion recognition task. The power of the model to successfully learn and represent the attention weights depends on the number of neurons in the attention layer. We used 128 neurons, increasing our model's capacity for capturing fine-grained feature importance while minimizing complexity.

Regularized neighbourhood component analysis (RNCA) feature selection. The RNCA feature selection mechanism is a specific class of feature weighting approach that carries out its operation by learning

Figure 4. Convolutional layers block diagram.

feature weight and maximizing the leave-one-out (LOO) accuracy of classification over sample data[43]. The LOO provides an unbiased estimate of a deep learning model performance. RNCA works by assessing the vector weight w that corresponds to the feature vector x_i through the optimization of a classifier that is based on the nearest neighbour scheme. It has a mechanism for controlling complexity and preventing overfitting on the density estimation. RNCA adopts a framework of selecting a certain reference sample called x_j for the sample x_i from all emotion feature samples randomly. However, the probability of the selected feature (P_{ij}) to x_j rest heavily on the distance D_w that exists between two samples. This distance can be computed[44] as in Eq. (5) below:

$$D_w(x_i, x_j) = \sum_{m=-1}^{r} w^2{}_m |x_{im} - x_{jm}| \tag{5}$$

Where *mth* the feature's weight is denoted by w_m. A kernel function k established the relation P_{ij} and Dw on the condition that the smaller the Dw the larger the values of k. The likelihood P_{ij} and kernel function k can be computed for Eqs. 6 and 7 respectfully as below

$$P_{ij} = \frac{k(D_w(x_i, x_j)}{\sum_{j=1, j\neq i}^{n} k(D_w(x_i, x_j)} \tag{6}$$

$$k(z) = exp - \frac{z}{\sigma} \tag{7}$$

where the kernel width is represented by σ that influences the likelihood that a reference point selected will be x_j sample. Therefore, the likelihood of correctly classifying x_i can be computed from Eq. (8).

$$P_i = \sum_{j=1, j\neq i}^{n} P_{ij} Y_{ij} \tag{8}$$

Where y_{ij} can only indicate one if both y_i and y_j are equal to each other. The average LOO accuracy of classification is the sum of all P_i of all the samples divided by the total number of samples, as indicated in Eq. (9). This equation can be termed as the objective function that required maximization. Nevertheless, the objective function defined above is not insulated from overfitting, which calls for the introduction of a parameter λ termed regularizer to prevent overfitting. The modified objective function that represents RNCA can be defined as

$$Obj.(A) = \sum_{i=1}^{n} P_i - \lambda \sum_{m=1}^{r} w^2{}_m \tag{9}$$

The RNCA algorithm adopted in this work operates on the output from the attention layer of our model to aid feature selection, therefore, it is essential to evaluate generalization error (Eq. 10) to properly fine-tune the regularization parameter λ to obtain a minimized classification loss.

$$Err = \frac{1}{n} \sum_{i=1}^{N} I(k_i \neq t_i) \tag{10}$$

where the predicted label is represented by k_i and t_i denotes the real label of the feature sample. The RNCA feature selection technique is diagrammatically shown in Fig. 5.

Emotion classification. In this study, three primitive classifiers were utilized in carrying out the classification of emotion. The classifiers take their simplified input from the output of the feature selection layer of our model after feature extraction. The essence of employing three different classifiers is to ensure the

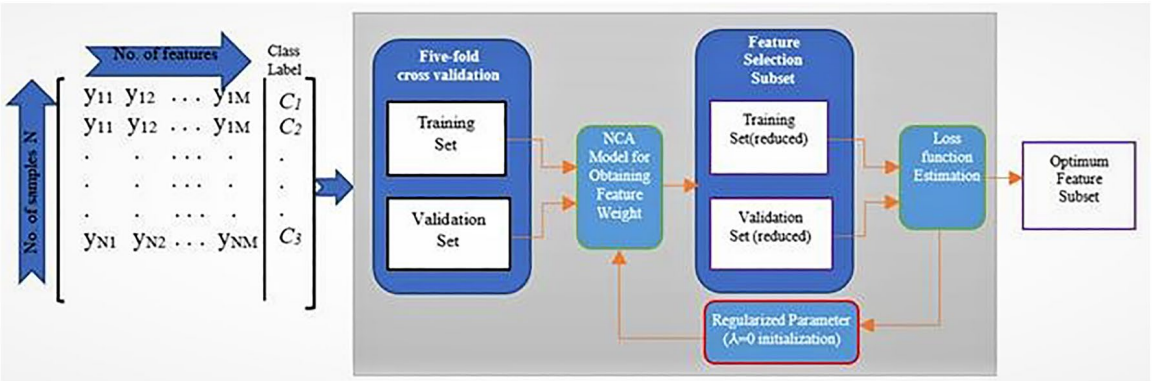


Figure 5. RNCA framework.

robustness of the entire model and aid the analysis of the result. Multi-layer perceptron (MLP) classifier is first introduced. As a feedforward network-based classifier[6], a set of suitable outputs are mapped from a set of input datasets by this feedforward artificial neural network model. An MLP is made up of several layers, each of which is completely connected to the one before it. Except for the nodes in the input layer, the nodes of the layers represent neurons with nonlinear activation functions.

Secondly, we also utilized a support vector machine (SVM). An SVM operates as a discriminative classifier, well-defined by dividing hyperplane. It fits into supervised and unsupervised machine-learning tasks. For instance, given a set of selected features (or data), the algorithm outputs an optimal hyperplane that classifies new samples. In two-dimensional space, this hyperplane is a line dividing a plane into two parts, wherein each class lay on either side[45]. Besides, SVM can effectively handle multiclass problems as it is obtainable with emotion classification. One distinguishable function of SVM is that it selects a hyper-plane with a large margin, reducing the likelihood of miss-classification and its low sensitivity to outliers.

Lastly, Random Forest (RF) was also employed as the third classifier. RF is a meta-estimator that employs averaging to increase classification accuracy and reduce overfitting after fitting several decision tree classifiers to different emotional feature subsamples. Random forest possesses an inbuilt mechanism for managing class imbalance, and this has given it an edge over other classifiers.

Experimental results and discussion

Dataset. In this study, we benchmarked our experiment on one of the publicly available datasets named Toronto English Speech Set (TESS). In 2010, at Northwestern University's Auditory Laboratory, TESS speech samples were recorded[46]. During the spontaneous event, two actresses were asked to recite a handful of the 200 words, and their voices were recorded, resulting in a complete collection of 2800 speech utterances. Seven different emotions comprise happy, angry, fear, disgust, pleasant, surprise, sad and neutral were observed in the scene.

Experimental configuration. In this study, the experiment was carried out using a 64-bit operating system, an Intel Core i7 processor, 8 GB of RAM, and a Python 3.9 environment. Deep learning software and additional third-party libraries (including Tensorflow, Numpy, and audio processing) were also utilized. The audio sample first needed to be pre-processed because the input layer for our model has to be in 224 x 224 x 3. To meet the requirements of the model, the voice signal has to be scaled and transformed into a log-mel spectrogram. The FFT technique was used to separate the mel-spectrogram feature from the original audio data. The dataset is then sectionalized into a training set and a testing set (80%:20%). Both the exam and practice sets' data were normalized to pixels.

Implementation parameters. In implementing our model and compilation of the network, we utilized the Adams optimizer with a learning rate set to 5e-5 notation. One-hot encoding technique was used in vectorizing the label. It ensures that the data point is binarized. We adopted sparse categorical cross entropy for the loss function. To actualize the objective of increasing accuracy, we initialize our model set up with 100 epochs and 16 batch size, however the result of our training after 25 epochs yielded optimum accuracy. We utilized a custom-early stopping mechanism to monitor (checkpoint) the loss and accuracy value to prevent overfitting, and the corresponding curve was obtained as well.

Experimental Results. The result of our experiment using attention-based networks and regularized feature selection with three classifiers are presented in this section. For the first experiment where the Vgg16 pre-trained network was utilized, the confusion matrix of emotion classification is shown in Figs. 6, 7, 8, 9. We observed that the attention network of our model achieved the highest accuracy (97.8%) of recognition with the RF classifier compared to the other classifiers (SVM:97.4% and MLP: 97.6%). From the figures, the emotional class of angry, disgust, fear and sad accuracy reach 100% with the attention-based network and RF, SVM and MLP. The Neural emotion class got 98% the highest accuracy of recognition with the RF classifier, while 94% best accuracy was obtained on surprise emotion from Figs. 6 and 7 respectively (Figs. 10 and 11). The performance evaluation chart in Fig. 12 shows other evaluation metrics (specificity, sensitivity, F1-score and unweighted average recall) used to establish the robustness of our model. The two experiments are captured on the chart.

In our second experiment, the pre-trained model used was Vgg19 before the attention layer was added. The result generated is shown in Figs. 9, 10, 11. Disgust emotion carries the highest classification accuracy of 100% from the three classifiers, while surprise emotion has the least classification accuracy of 93%. Neutral emotion differs in accuracy from the three classifiers, its optimum accuracy is at 99% with the SVM classifier. The overall model classification accuracy obtained from the second experiment is 97.5%. This is low compared to the previous experiment where vgg16 was used as the convolutional layer, however, the impact of the attention network for the extraction of emotionally related features combined with regularized feature selection has improved the classification accuracy of speech emotion.

Besides the accuracy obtained through the confusion matrix, the model loss and ROC (Return of Characteristics) curves in Figs. 13 and 14 further testify to the performance of our model in this paper. The loss value from the curve is relatively low, indicating that our model has prevented overfitting. The loss curve decreases over time as our model improved. Also, the loss curve shows a smooth convergence which is a further indication that our model prediction is accurate to an acceptable level. The low initial loss value with respect to the convergence point as shown confirmed the reduction in model complexity and training time. The ROC curve shows the seven categories of emotion as indicated in Table 1 below with the area under the curve (AUC) which demonstrates the performance average across all potential emotion classification thresholds. The diagonal dotted line is the

111

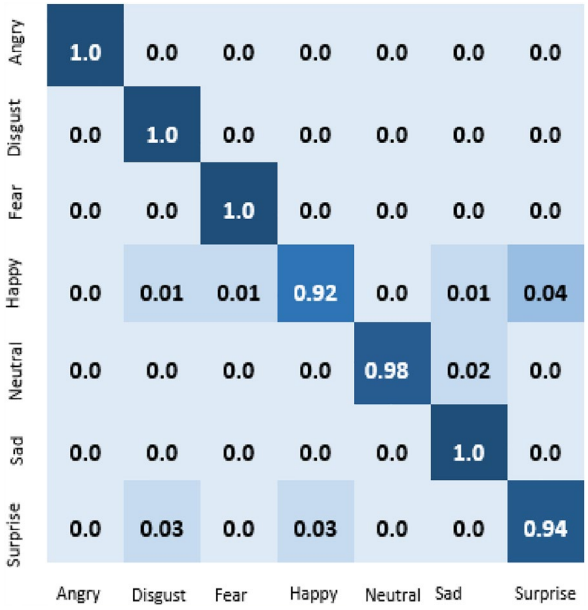


Figure 6. Attention-based Vgg16+RNCA+RF.

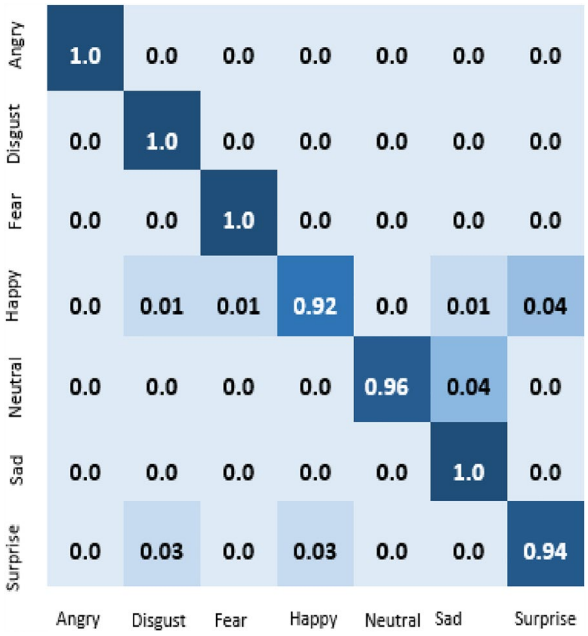


Figure 7. Attention-based Vgg16+RNCA+MLP.

threshold. The closeness of the curve to the top left-hand corner for the seven emotional classes indicates a high True Positive Rate (TPR) and low False Positive Rate (FPR). The least AUC score recorded is 0.98, an evidence of the good performance of our model on emotion classification.

In this work, the Mel-spectrogram was used to extract the input feature, producing a feature vector with a dimensionality of 40 (mel frequency bins). These features record important details about the speech signal's spectral composition and temporal dynamics. The feature space was high dimensional, so feature selection was used to lower the dimensionality and concentrate on the most useful features for the task. The feature selection algorithm assessed each feature's relevance based on how it contributed to the performance of emotion recognition, taking into account measures like mutual information and feature importance scores, thereby increasing the model's efficiency, lowering the amount of computing power required, and improving the interpretability of the learned representations. Our experiments' findings showed that feature selection significantly enhanced the speech emotion recognition model's performance, resulting in an increase in accuracy

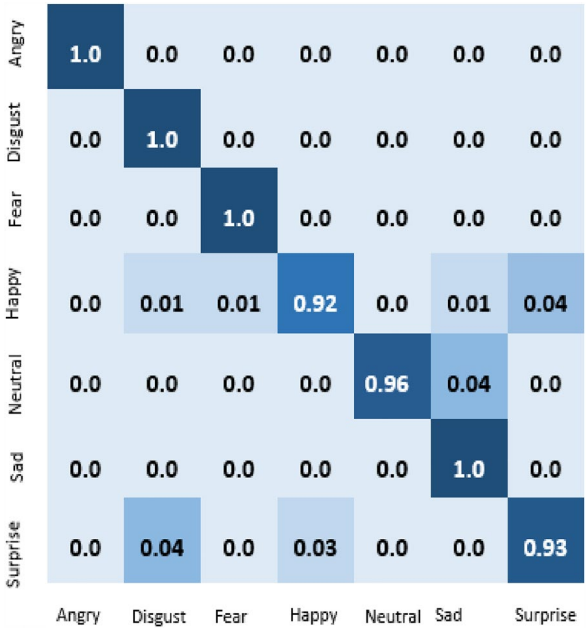


Figure 8. Attention-based Vgg16+RNCA+SVM.

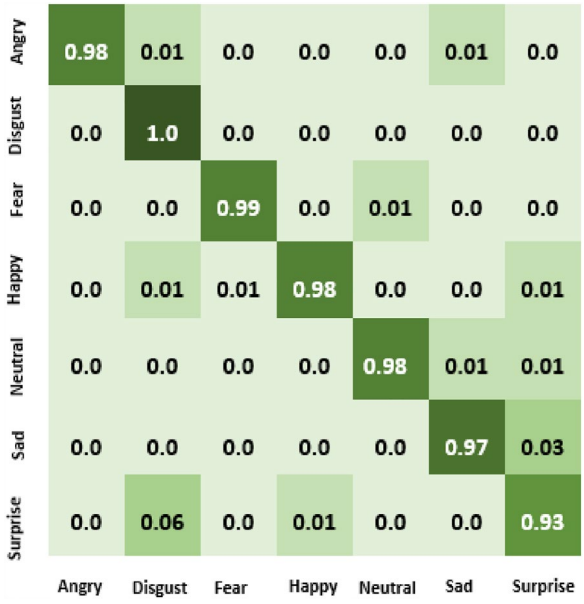


Figure 9. Attention-based Vgg19+RNCA+RF.

of 3.7%, underscoring the significance of feature selection in improving the model's discriminative power for emotion recognition tasks.

Performance Comparison. Additionally, our proposed model in this study was compared with other work carried out by others benchmarked on the same speech dataset, as indicated in Table 2. We also carried out a comparative analysis of our proposed model without the attention layer, RNCA, and with the attention mechanism and RNCA feature selection as shown in Table 3.

In terms of accuracy, reduction of complexity and prevention of overfitting, our method surpasses other methods[47–51] utilized for speech emotion classification or recognition.

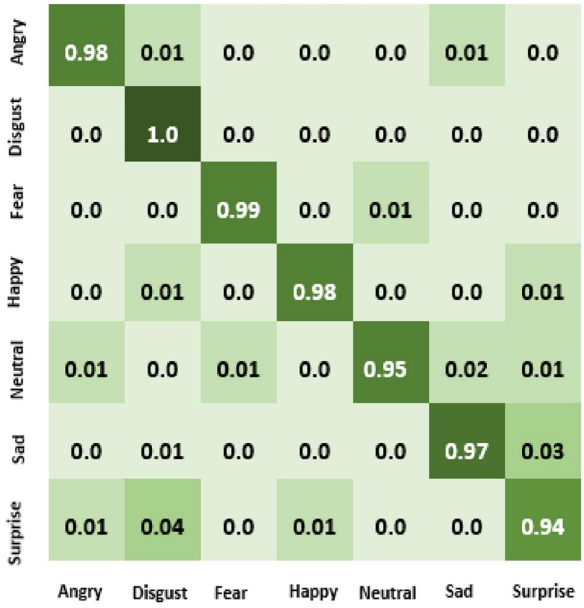


Figure 10. Attention-based Vgg19+RNCA+MLP.

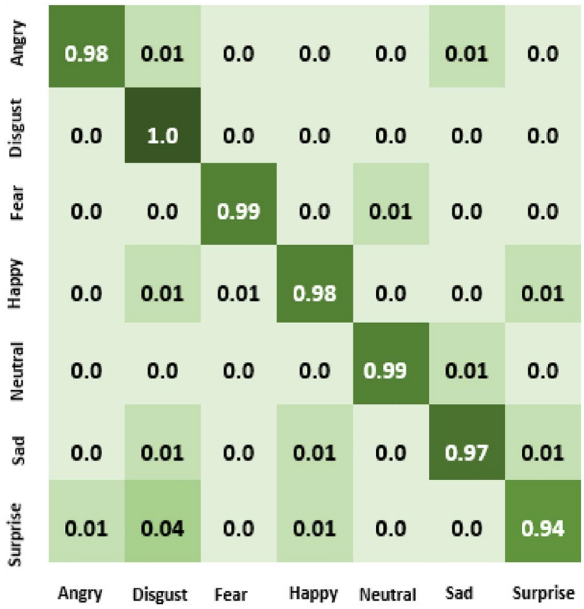


Figure 11. Attention-based Vgg19+RNCA+SVM.

Conclusion

In this study, we proposed a SEC system using an attention-based network and regularized feature selection. First and foremost, we extracted the mel-spectrogram from the TESS dataset used for this study. This was carried out, after extensive speech processing and analysis, to feed (input layer) our model with appropriate features for enhanced feature extraction in the subsequent layers. A pre-trained DCNN base model was adopted for our attention network to extract local features, while the attention layer deals with emotionally rich features (global features) which ultimately reduces misclassification to the barest minimum. The core principle of the attention network is to estimate feature weight. In our attempt to increase the efficiency of our model, a regularized feature selection is introduced after the attention layer to actualize optimum results. The feature selection aided the attention mechanism to focus more on salient features. Thereafter, three classifiers were fed with selected emotional features with RNCA, for the classification of emotion.

After a comparison of the result of our experiments, an attention-based DCNN+RNCA+RF model for speech emotion classification was proposed. The experimental result attained the optimum accuracy of 97.8% on the

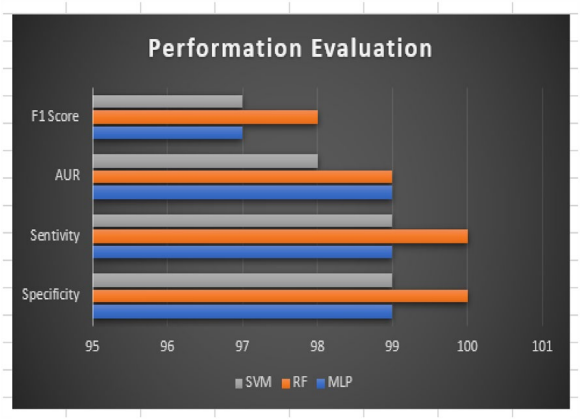


Figure 12. Performance chart with 4 metrics and 3 classifiers.

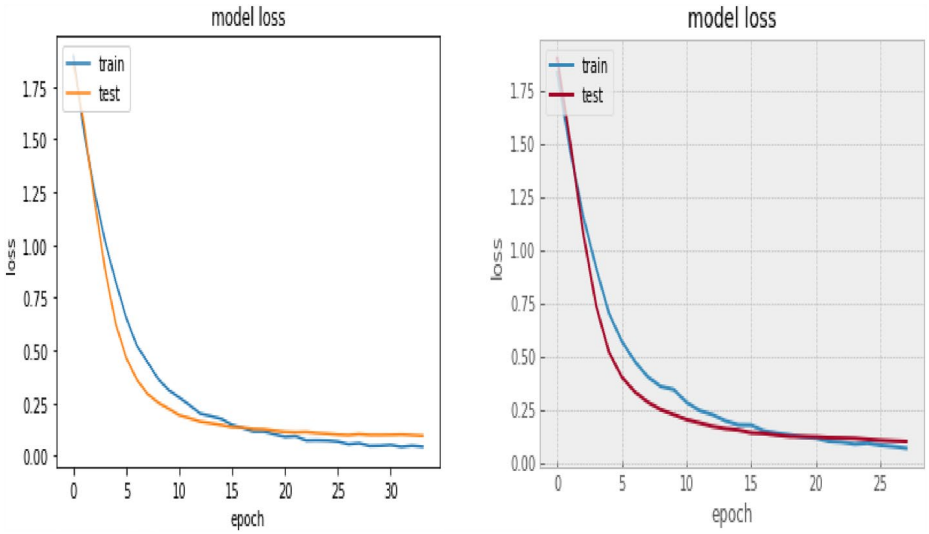


Figure 13. Model loss curve.

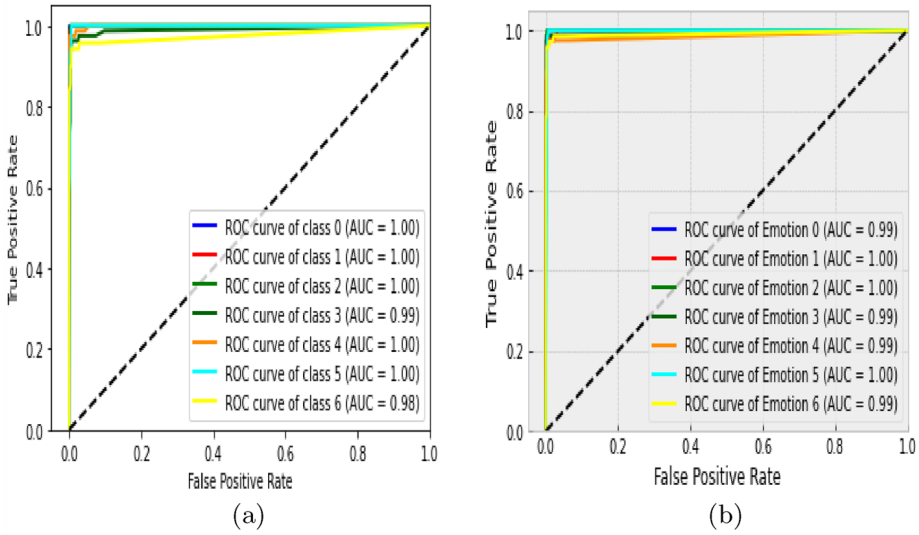


Figure 14. ROC curve.

Emotion	Class	AUC score (%)
Angry	0	100
Disgust	1	100
Fear	2	100
Happy	3	99
Neutral	4	100
Sad	5	100
Surprise	6	98

Table 1. Return of Characteristics Description.

Publication	Model	Dataset	Accuracy (%)
2017[47]	MFCC+SVM	TESS	96.00
2018[48]	DNN+GRU	TESS	95.82
2019[49]	MFCC+CNN	TESS	81.00
2021[50]	IMF+SVM+KNN	TESS	93.30
2022[51]	DNN+NCA+MLP	TESS	96.10
Proposed	Attention-based DCNN+RNCA+RF	TESS	97.8

Table 2. Comparison of our proposed with other methods.

Emotion	Without attention(%)	Without RNCA(%)	Attention + RNCA(%)
Angry	98	100	96.00
Sad	91	91	95.82
Surprise	86	89	81.00
Happy	98	100	93.30
Fear	96	100	96.10
Neutral	95	98	96.10
Disgust	94	100	96.10
Average	94.00	96.85	97.71

Table 3. Emotional level comparison of the significance of attention and feature selection.

TESS dataset. Seven classes of emotions comprised of anger, sad, happy, fear, neutral, disgust and surprise that reflect human major emotions were accurately classified. Besides, by contrasting our proposed model in this study with other methods that have recently been put forward, obviously, our model outperforms many of them in speech emotion classification tasks.

Moreover, the computational cost peculiar to most deep learning tasks is prevented in this study, simply because our based model for the attention network requires no training and the total number of trainable parameters has been reduced to the barest minimum (101,480) out of the total parameters of 14,017,704. The number of floating-point operations per seconds(FLOPs), and the model's (size of 98MB) memory requirement have been reduced to minimize complexity because the top layer of the VGGNet has been frozen. The average time taken for each emotional utterance to be classified by the proposed model is 0.12. However, though, the result obtained from this study has undoubtedly provided some insight for researchers on the application of attention mechanism with feature selection for SEC tasks, we recommend that future work can be carried out using a sequential network, more pre-trained based network, low-level features and introduction of other speech emotion dataset.

Data availability

Benchmarked publicly available dataset, Toronto English Speech Set (TESS) is used.

Received: 13 January 2023; Accepted: 16 July 2023
Published online: 25 July 2023

References

1. Costantini, G., Parada-Cabaleiro, E., Casali, D. & Cesarini, V. The emotion probe: On the universality of cross-linguistic and cross-gender speech emotion recognition via machine learning. *Sensors*https://doi.org/10.3390/s22072461 *(2022).*
2. Chimthankar, P. P. Speech Emotion Recognition using Deep Learning. http://norma.ncirl.ie/5142/1/priyankaprashantchimthankar.pdf (2021)
3. Saad, H. F.and Mahmud, Shaheen, M., Hasan, M., Farastu, P. & Kabir, M. Is speech emotion recognition language-independent? Analysis of english and bangla languages using language-independent vocal features. arXiv:2111.10776 (2021)
4. Burghardt, G. M. A place for emotions in behavior systems research. *Behavioural Process.*https://doi.org/10.1016/j.beproc.2019.06.004 *(2019).*
5. Mustaqeem, & Kwon, S. The emotion probe: On the universality of cross-linguistic and cross-gender speech emotion recognition via machine learning. *Appl. Soft Comput.*https://doi.org/10.1016/j.asoc.2021.107101 *(2021).*
6. Ba'abbad, I., Althubiti, T., Alharbi, A., Alfarsi, K. & Rasheed, S. A short review of classification algorithms accuracy for data prediction in data mining applications. *J. Data Anal. Inform. Process.* **09**, 162–174. https://doi.org/10.4236/jdaip.2021.93011 (2021).
7. Choudhary, G. R., Meena, G. & Mohbey, K. Speech emotion based sentiment recognition using deep neural networks. *J. Phys. Conf. Ser.* **2236**(1), 012003. https://doi.org/10.1088/1742-6596/2236/1/012003 (2022).
8. Wani, T., Gunawan, T., Qadri, S., Kartiwi, M. & Ambikairajah, E. A comprehensive review of speech emotion recognition systems. *IEEE Access* **9**, 47795–47814. https://doi.org/10.1109/ACCESS.2021.3068045 (2021).
9. Cowen, A. & Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Nat. Acad. Sci. U.S.A.* **114**(38), 7900–7909. https://doi.org/10.1073/pnas.1702247114 (2017).
10. Oaten, M., Stevenson, R. J. & Case, T. I. Disgust as a disease-avoidance mechanism. *Psychol. Bull.* **135**(2), 303–321. https://doi.org/10.1037/a0014823 (2009).
11. Anvarjon, T., Mustaqeem, & Kwon, S. Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. *Sensors (Switzerland)* **20**(18), 1–16. https://doi.org/10.3390/s20185212 (2020).
12. Kwon, S. A CNN-assisted enhanced audio signal processing. *Sensors*https://doi.org/10.3390/s20185212 *(2020).*
13. Staudemeyer, R. & Morris, E. Understanding LSTM–a tutorial into Long Short-Term Memory Recurrent Neural Networks. arXiv:1909.09586 (2019)
14. Atila, O. & Şengür, A. Attention guided 3d CNN-LSTM model for accurate speech based emotion recognition. *Appl. Acoustics*https://doi.org/10.1016/j.apacoust.2021.108260 *(2021).*
15. Farooq, M., Hussain, F., Baloch, N., Raja, F. & Zikria, Y. Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. *Sensors (Switzerland)* **20**(21), 1–18. https://doi.org/10.3390/s20185212 (2020).
16. Hajarolasvadi, N. & Demirel, H. 3d CNN-based speech emotion recognition using k-means clustering and spectrograms. *Entropy*https://doi.org/10.3390/e21050479 *(2019).*
17. Zhu, L., Chen, L., Zhao, D., Zhou, J. & Zhang, W. Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN. *Sensors (Switzerland)*https://doi.org/10.3390/s17071694 *(2017).*
18. Wang, Z. & Tashev, I. Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. *IEEE Int. Conf. Acoustics Speech Signal Process.* **17**(7), 5150–5154. https://doi.org/10.1109/ICASSP.2017.7953138 (2017).
19. Pascanu, R., Gulcehre, C., Cho, K. & Bengio, Y. How to construct deep recurrent neural networks. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, p. 1–13 (2014)
20. Kerkeni, L. *et al.* Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. *Speech Commun.* **114**, 22–35. https://doi.org/10.1016/j.specom.2019.09.002 (2019).
21. Lieskovská, E., Jakubec, M., Jarina, R. & Chmulík, M. A review on speech emotion recognition using deep learning and attention mechanism. *In Electronics (Switzerland)*https://doi.org/10.3390/electronics10101163 *(2021).*
22. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 (1997).
23. Zhao, Z. *et al.* Self-attention transfer networks for speech emotion recognition. *Virtual Real. Intell. Hardw.* **3**(1), 43–54. https://doi.org/10.1016/j.vrih.2020.12.002 (2021).
24. Puri, T., Soni, M., Dhiman, G., Khalaf, O. & Khan, I. Detection of emotion of speech for Ravdess audio using hybrid convolution neural network. *Hindawi J. Healthc. Eng.*https://doi.org/10.1155/2022/8472947 *(2022).*
25. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15 (2015)
26. Qamhan, M., Meftah, A., Selouani, S., Alotaibi, Y., Zakariah, M. & Seddiq, Y. Speech emotion recognition using convolutional recurrent neural networks with attention model. Canadian Conference on Electrical and Computer Engineering 2020-Augus(Cii), 341–350 (2020). https://doi.org/10.1109/CCECE47787.2020.9255752
27. Chen, M., He, X., Yang, J. & Zhang, H. 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Process. Lett.* **25**(10), 1440–1444. https://doi.org/10.1109/CCECE47787.2020.9255752 (2018).
28. Zhao, Z., Zheng, Y., Zhang, Z., Wang, H., Zhao, Y., Li, C.: Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNS and FCNS for speech emotion recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018-Septe(September), 272–276 (2018). https://doi.org/10.21437/Interspeech.2018-1477
29. Du, Q., Gu, L., Zhang, W. & Huang, S. Poster abstract: Attention-based LSTM-CNNS for time-series classification. In *SenSys 2018 - Proceedings of the 16th Conference on Embedded Networked Sensor Systems*, 410–411 (2018). https://doi.org/10.1145/3274783.3275208
30. Seo, M. & Kim, M. Fusing visual attention CNN and bag of visual words for cross-corpus speech emotion recognition. *Sensors* **20**, 5559. https://doi.org/10.3390/s20195559 (2018).
31. Zhang, J., Xing, L., Tan, Z., Wang, H. & Wang, K. Multi-head attention fusion networks for multi-modal speech emotion recognition. *Comput. Ind. Eng.* **168**, 108078. https://doi.org/10.1016/j.cie.2022.108078 (2022).
32. Zhang, H. *et al.* Pre-trained deep convolution neural network model with attention for speech emotion recognition. *Front. Physiol.*https://doi.org/10.3389/fphys.2021.6432028 *(2021).*
33. Chen, S. *et al.* The impact of attention mechanisms on speech emotion recognition. *Sensors*https://doi.org/10.3390/s21227530 *(2021).*
34. Zhao, Z. *et al.* Self-attention transfer networks for speech emotion recognition. *Virtual Real. Intell. Hardw.*https://doi.org/10.1016/j.vrih.2020.12.002 *(2021).*
35. Zhou, S. & Beigi, H. A transfer learning method for speech emotion recognition from automatic speech recognition. arXiv:2008.02863 (2021)
36. Singh, Y. & Goel, S. A systematic literature review of speech emotion recognition approaches. *Neurocomput. Elsevier*https://doi.org/10.1016/j.neucom.2022.04.028 *(2022).*
37. Atsavasirilert, K., Theeramunkong, T., Usanavasin, S., Rugchatjaroen, A., Boonkla, S., Karnjana, J., Keerativittayanun, S. & Okumura, M. A light-weight deep convolutional neural network for speech emotion recognition using mel-spectrograms. In *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing* (ISAI-NLP) (2019)

117

38. Zhou, Q. *et al.* Cough recognition based on MEL-spectrogram and convolutional neural network. *Front. Robot. AI*https://doi.org/10.3389/frobt.2021.580080 *(2021).*
39. Chen, Q. & Huang, G. A novel dual attention-based BLSTM with hybrid features in speech emotion recognition. *Eng. Appl. Artif. Intell.*https://doi.org/10.1016/j.engappai.2021.104277 *(2021).*
40. Bilal, M. Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition. arXiv:1907.06083v2 (2019)
41. Tursunov, A., Mustaqeem, Choeh, J. Y. & Kwon, S. Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. *Sensors*https://doi.org/10.3390/s21175892 *(2021).*
42. Ho, N., Yang, H., Kim, S. & Lee, G. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. IEEE Access **2020**(8) (2020)
43. Zhou, A., Luktarhan, N. & Ai, Z. Research on webshell detection method based on regularized neighborhood component analysis (RNCA). *Symmetry*https://doi.org/10.3390/sym13071202 *(2021).*
44. Malan, N. & Sharma, S. Feature selection using regularized neighbourhood component analysis to enhance the classification performance of motor imagery signals. *Comput. Biol. Med.*https://doi.org/10.1016/j.compbiomed.2019.02.009 *(2019).*
45. Duville, M., Alonso-Valerdi, L. & Ibarra-Zarate, D. Mexican emotional speech database based on semantic, frequency, familiarity, concreteness, and cultural shaping of affective prosody. *Data*https://doi.org/10.3390/data6120130 *(2021).*
46. Dupuis, K. & Kathleen Pichora-Fuller, M. Recognition of emotional speech for younger and older talkers: Behavioural findings from the Toronto emotional speech set. *Can. Acoust.*https://doi.org/10.3389/fphys.2021.6432028 *(2011).*
47. Verma, D. M. Age driven automatic speech emotion recognition system. *IEEE Int. Conf. Comput. Commun. Autom*https://doi.org/10.1109/CCAA.2016.7813862 *(2017).*
48. Praseetha, V. & Vadivel, S. Deep learning models for speech emotion recognition. *J. Comput. Sci.*https://doi.org/10.3844/jcssp.2018.1577.1587 *(2018).*
49. Gao, Y. Speech-Based Emotion Recognition. https://libraetd.lib.virginia.edu/downloads/2f75r8498?filename=1_Gao_Ye_2019_MS.pdf (2019)
50. Krishnan, P., Joseph Raj, A. & Rajangam, V. Emotion classification from speech signal based on empirical mode decomposition and non-linear features. *Complex Intell. Syst.*https://doi.org/10.1007/s40747-021-00295-z *(2021).*
51. Akinpelu, S. & Viriri, S. Robust feature selection-based speech emotion classification using deep transfer learning. *Appl. Sci.* **12**, 8265. https://doi.org/10.3390/app12168265 (2022).

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to S.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 4

Lightweight Deep Learning Framework for Speech Emotion Recognition

4.1 Brief Review

This chapter presents the paper entitled Lightweight Deep Learning Framework for Speech Emotion Recognition. This chapter extends research on chapter three by presenting a deep learning model that overcomes the limitation of the deep learning technique's over-reliance on the tunning of millions of parameters before achieving a significant accuracy of emotion classification. The model utilizes three convolutional layers of a Visual Geometric network with the best-performing classifier. It eliminates redundant layers that do not contribute to the extraction of emotional features but rather increase the number of parameters. A novel method that performs the classification of emotional utterance from speech signals across three datasets and can be implemented on low-memory devices is devised in this paper.

Paper status: Published in Published in IEEE Access.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 June 2023, accepted 10 July 2023, date of publication 20 July 2023, date of current version 28 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3297269

RESEARCH ARTICLE

Lightweight Deep Learning Framework for Speech Emotion Recognition

SAMSON AKINPELU, SERESTINA VIRIRI, (Senior Member, IEEE), AND ADEKANMI ADEGUN
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban 4041, South Africa
Corresponding author: Serestina Viriri (viriris@ukzn.ac.za)

ABSTRACT Speech Emotion Recognition (SER) system, which analyzes human utterances to determine a speaker's emotion, has a growing impact on how people and machines interact. Recent growth in human-computer interaction and computational intelligence has drawn the attention of many researchers in Artificial Intelligence (AI) to deep learning because of its wider applicability to several fields, including computer vision, natural language processing, and affective computing, among others. Deep learning models do not need any form of manually created features because they can automatically extract the prospective features from the input data. Deep learning models, however, call for a lot of resources, high processing power, and hyper-parameter tuning, making them unsuitable for lightweight devices. In this study, we focused on developing an efficient lightweight model for speech emotion recognition with optimized parameters without compromising performance. Our proposed model integrates Random Forest and Multi-layer Perceptron(MLP) classifiers into the VGGNet framework for efficient speech emotion recognition. The proposed model was evaluated against other deep learning based methods (InceptionV3, ResNet, MobileNetV2, DenseNet) and it yielded low computational complexity with optimum performance. The experiment was carried out on three datasets of TESS, EMODB, and RAVDESS, and Mel Frequency Cepstral Coefficient(MFCC) features were extracted with 6-8 variants of emotions namely, Sad, Angry, Happy, Surprise, Neutral, Disgust, Fear, and Calm. Our model demonstrated high performance of 100%, 96%, and 86.25% accuracy on TESS, EMODB, and RAVDESS datasets respectively. This revealed that the proposed lightweight model achieved higher accuracy of recognition compared to the recent state-of-the-art model found in the literature.

INDEX TERMS Deep learning, convolutional neural network, speech emotion, lightweight, human–computer interaction.

I. INTRODUCTION

The social structure, the demand for talent, and human-machine interaction have all altered because of the rapid growth of deep learning from artificial intelligence (AI), data science, and IoT technology [1], [2], [3]. Speech is an exceptionally straightforward and seamless mode of human interaction that has been proven to effectively and swiftly communicate information in this era. People now dedicate a lot of time to learning how to speak to a variety of smart devices and interact with them through speech signals. Currently, a wide range of voice assistants, including Amazon Alexa, Microsoft Cortana, and Samsung Bixby recognize human-generated information through interactive real-time intelligent conversation and realize automatic operation in accordance with the speech content [4].

The computer has given birth to many innovations through which traditional learning and communication have given way to animation, visual art, artificial reality, and other forms of computer-aided means of expression. As a result of the advancement in technology, the learning environment has evolved, and researchers' interest in the intelligent learning environment built around AI has increased significantly. Teachers conduct their educational activities online over the Internet in an adaptive learning environment, and students can readily learn new information through the network. However,

The associate editor coordinating the review of this manuscript and approving it for publication was Giacomo Fiumara.

all these teacher-learner communications are not without expression of one emotion or the other. Studies from the psychological domain have revealed that different emotions that emerge throughout the learning process can impact the learning outcome. Research has also demonstrated that while negative emotions [5], [6] like disgust, fear, and sadness can impede cognition, good emotions like happiness and joy that are produced throughout the learning process are beneficial to increasing learning interest [7], [8], [9].

Humans communicate naturally through speech. Communication through speech utterance has become a major necessity in human life by which messages and ideas are conveyed [10], [11]. Therefore, interacting with the computer rather than typing on a keyboard has become more pleasant for people. Speech Emotion Recognition (SER) refers to the process by which a computer is made to recognize the inherent emotion present in speech signals. If the SER system is successful, the computer will be able to engage with people on an entirely new level. For instance, in an e-learning system, the computer can identify the speaker's emotions and provide helpful answers by suggesting simpler learning steps than the ones already in place [12]. Additional uses of SER include health care systems, aiding fighter pilots in combat, mental disorder treatment, and caring for the elderly in society.

Deep learning differs from the conventional machine learning approach that is handcrafted in nature (Fig. 1). The machine learning approach usually consists of a separate segment for its feature extraction before introducing a machine learning algorithm [13] while deep learning does not require handcrafted feature extraction approach because the algorithm such as Convolutional Neural Network (CNN) searches what specific features is best for classifying images [14]. In other words, the machine learning approach tends to break the problem down into constituents' parts at first and later aggregate the result at the output stage whereas the deep learning model proffers a solution to a problem in an end-to-end manner.

Recent years have seen a rise in the use of deep learning models for the extraction of speech signals' emotions [15], [16]. For feature extraction, convolutional neural networks (CNNs) have been very effective [17], [15]. Researchers have also thought about combining CNNs with other machine learning techniques, such as support vector machines, to increase the effectiveness of emotion identification models [18], [19]. The development of compact models for usage on mobile and embedded systems has also received attention.

In general, deep learning generally requires a very huge volume of training data to avoid overfitting though not very much feasible in most SER systems (limited dataset) [20]. Researchers have attempted to perform a deeper convolution layer to increase accuracy and improve performance, however, they have discovered that resources for computation, like memory storage, have proven to be a significant roadblock. Computational complexity [21] has been the major drawback in this case because of several convolutional layers. Therefore, a lightweight model which is the focus of this study becomes paramount in increasing the efficiency and performance of the SER model on low low-memory devices, without comprising accuracy.

The following are the paper's main contributions: Speech emotion recognition is made possible by: (i). A lightweight model using deep learning techniques for feature extraction and the best-performing classifier for classification (ii) Reduction in the number of computing resources needed for deep learning model feature extraction through architectural depth optimization. (iii) Evaluation of our proposed lightweight model against existing deep learning models on Central Processing Unit(CPU) and low-memory devices, which performs significantly better in terms of accuracy of recognition and execution time than a complete deeply learn architecture.

II. RELATED WORKS

The creation of models for recognizing emotions from speech signals has garnered more attention in recent years. Both conventional techniques like hidden Makov Model(HMM) and Support Vector Machine (SVM) and more contemporary ones like CNNs, Recurrent Neural Networks (RNNs), and Long Short-Term Memory(LSTMs) have been investigated. In comprehending conversations more effectively, researchers have also considered leveraging transfer learning, domain adaptation, and Natural Language Processing (NLP) approaches, as well as utilizing the time-frequency information contained in voice signals for effective recognition of emotion.

Scholars have also introduced several traditional methods, such as HMMs and Random Forest, to extract acoustic information from speech signals [22]. Researchers have looked into recurrent neural networks (RNNs) and long short-term memory (LSTM) networks as viable methods for classifying emotions from voice inputs [23]. Also, some studies have examined how various forms of data augmentation [24] affect emotion classification models.

In addition to the approaches discussed above, researchers have also used transfer learning and domain adaptation strategies to improve the performance of emotion recognition models [25]. Other research has looked at exploiting the time-frequency information of speech signals [26], [27]. Furthermore, there have been studies exploring the use of biometric features of speech signals for emotion recognition [28], [29].

Several studies investigated the use of generative models for voice signal emotion identification [30], [31], [32], [33]. The use of NLP methods to extract features from conversations has also received attention [34]. Researchers have also looked into harnessing processes for self-attention to better recognize salient aspects in human speech [35] for accurate recognition of emotion. Moreover, experiments investigating the integration of deep learning and unsupervised learning techniques to enhance emotion identification models have been conducted [36].

The achievement of the Deep Neural Network(DNN) in SER task cannot be overemphasized, though with a few peculiar limitations. Unlike any other image recognition task, speech signal differs in terms of environment, style, language, and content of the speaker. Also, DNN is prone to learning a high-level feature from common Low-Level Descriptors (LLD), a technique that is less sufficient in extracting all emotional features from speech signals. This is what paved the way for the use of the Mel Frequency Cepstral Coefficient (MFCC) by researchers in representing speech signals [37]. Two axes (vertical and horizontal) exist in any MFCC representation. While the horizontal axis carries information that is time-domain specific, the vertical axis usually carries information that has to do with the frequency of the signal. Thus, positioning MFCC has a unique representation of speech signals that possess essential speech emotional features. Therefore, CNN has achieved improved performance in the SER domain because it extracts emotional features from MFCC in an automatic manner.

In Muyawei et al. [38], a distributed CNN and bidirectional recurrent neural network were utilized in obtaining emotional features from human raw speech. They adopted the attention mechanism technique of focusing on the most useful section of emotion and achieved a weighted accuracy of 64.08% on the IEMOCAP dataset. The author in [39] proposed a combination of attention-based RNN and convolutional neural networks for SER. Their method achieved a significant performance on IEMOCAP and FAU datasets [40] as indicated in the result obtained. A parallelized CRNN (Convolutional Recurrent Neural Network) is proposed by Jiang et al., [41] to acquire more salient emotional features from human speech. In their methodology, LSTM was utilized to learn frame-level features and CNN was used to learn other features from the log Mel-spectrogram, thereafter, all the features were fused. Finally, a softmax classifier was adopted to classify emotion on four public speech emotion datasets. Their experimental result showed superior performance in previous work. Prau et al. [42] proposed presented a neural network classifier and RNN for speech emotion recognition. The author applied a denoising technique in removing noise from the speech dataset, through the median filter. However, RNN is prone to dependency problems, even though it has excellent performance on time series data. Also, no record of testing their model on any of the publicly available datasets.

Chimthankar [43] proposed an innovative deep learning technique that combines CNN and LSTM on MFCC features extracted from four popular speech datasets (TESS, RAVDESS, SAVEE, and CREMA-D). Their model achieved 67.58% validation accuracy and 71.28% testing accuracy on a German based (audio samples) separate dataset which was not introduced to the model during training. However, an improvement in the performance of the recorded experimental result can still be achieved and computational complexity peculiar to the CNN model. Atila et al. [44] proposed a 3D CNN-LSTM with an attention mechanism model for SER. Four features including MFCC, fractal dimension, etc. were used in their study. The method showed an improved result, but the number of parameters generated tends to increase complexity which may not be suitable for low-memory devices.

Aggarwal et al. [45] applied principal component analysis (PCA)-DNN and pre-trained VGG-16 model as a two-way feature extraction approach for speech emotion recognition. Their extensive experimental result over two datasets with in-depth analysis yielded an improved accuracy on the RAVDESS dataset alone. The author showed that their model achieved better performance compared to the one-way feature extraction approach with DNN. However, better performance over only one dataset poses a limitation on the generalizability of this work. A gender-dependent CNN model for SER was proposed in [46]. The author captured two different emotional intensities (normal and strong) on six emotions (Sad, Fear, angry, happy, disgust, and calm) from the RAVDESS speech dataset, with an improved result. MFCC features and its variant was used in their study with an in-depth performance comparison showing a relative improvement over the baseline system that utilized emotional features such as Chromogram, mel-spectrogram, MFCC, and Spectral contrast.

Building a lightweight model with limited data has been a challenging task in SER. The author in [47] proposed a lightweight architecture for speech emotion recognition using convolution that is separable, inverted residuals and attention mechanism. They achieved 71.72% and 90.1% on two well-established datasets (IEMOCAP and EMODB) respectively. Atsavasirilert et al., [48] also presented a lightweight DCNN model for SER using the EMOD dataset only and they achieved the highest accuracy of 87.16% accuracy. Speech emotion recognition based on lightweight CNN was proposed in Anvarjon et al. [49]. Deep features from the spectrogram were extracted from IEMOCAP and EMODB datasets, and they achieved 77.01% and 92.02% accuracy respectively on both datasets. Inspired by the need to build a more lightweight deep learning architecture for SER, coupled with the fact that much research has a lesser focus on SER for low-memory devices, this paper proposed an optimized and state-of-the-art lightweight methodology for speech emotion recognition with low computational complexity and higher recognition accuracy.

III. PROPOSED METHODOLOGY

The proposed lightweight architecture is designed to carry out speech emotion recognition tasks directly from speech signals irrespective of environmental background or language. One major pre-processing required is to reshape the MFCC speech feature image to 224 × 224 as input to our model for standardizing all the speech datasets used. The remaining part of this section gives detail of how feature extraction is performed using VGGNet and classifiers used in classifying emotion as shown in Figure 1. The input to our model is at first subject to a 2D convolutional layer with 2 strides for

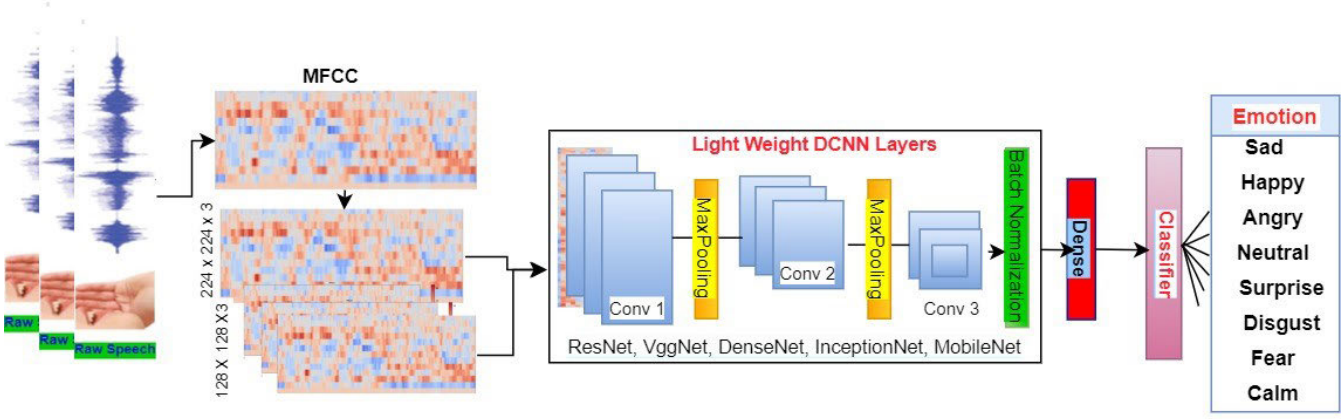


FIGURE 1. Lightweight SER architectural framework.

the extraction of distinct feature representations. A higher-level feature produced through the discriminative features is passed as input to the classifier for eventual speech emotion recognition. The result from other experimental studies is showcased as well to explain the rationale behind the choice of VGGNet and random forest classifier.

A. SPEECH FEATURE EXTRACTION

In this study, the emotional feature utilized is MFCC. Much research in SER has adopted MFCC as an efficient feature when it comes to emotion recognition. It can adequately describe the human vocal tract in speech utterance especially when speech sounds exhibit different kinds of emotion [50]. The mechanism behind the MFCC [51] extraction from the raw speech is to pass the audio spectrum into a Meyer filter bank to filter the audio spectrum in the frequency domain by the characteristics of human perception of sound frequency. Equation 1 illustrates the actual relationship between the filter's center frequency $Mel(f)$ and frequency f.

$$Mel(f) = 2595 \log\left(1 + \frac{f}{700}\right) \tag{1}$$

where f is the frequency in Hz

The MFCC approach is based on the idea of Mel frequency, which is one of the most popular and useful ways to characterize parameters and correct for convolutional channel distortion. The Fast Fourier Transform (FFT), Meyer-filter-bank, Normalization, Framing, Windowing, and Discrete Cosine Transform (DCT) are all parts of the MFCC extraction procedure as shown in Figure 2. For the extraction of the MFCC feature in this study, we first scanned through all the audio (WAV) files of our dataset, RAVDESS, TESS, and EMODB, and all the voice paths were saved. This was necessary to label each file using the path truncation approach, read the audio files to acquire information about them, and thereafter extract the MFCC from the audio files. We extracted 13 MFCC features as a standard practice for SER task and the dataset used. At the pre-emphasis stage, we employed a first-order high-passed filter of 100KHz to emphasize components of higher frequency. Each pre-processed audio signal is divided into equal lengths of frames. We utilize a hamming window of 25ms length, 16kHz sampling frequency in carrying out the filtering process of the short fast Fourier transform (STFT). The windowing is performed using equation 2 to prevent spectral leakage. The final MFCC coefficients were obtained by applying the DCT to the log filter-bank energies. The higher-order coefficients were left out and we maintained the first 13 coefficients as our features. The choice of MFCC features in this study is due to the fact that they are highly compact in terms of speech signal representation (emotion-rich feature), robust to noise variability, and computationally efficient which are beneficial to the lightweight deep learning model as compared to spectrogram.

$$w(n) = \begin{cases} 0.5 - 0.5\cos\left[2\pi n/(n-1)\right] & 0 \le n = N-1 \\ 0 & other \end{cases} \tag{2}$$

where f is the frequency in Hz

Speech signal has a peculiarity of continuous form in the time domain, and to better capture this continuous nature, a first-order and second-order difference method on the MFCC is utilized as shown in equation 3:

$$dt = \frac{\sum_{n=1}^{N} n(c_{t+n} - C_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{3}$$

where c_t represents the speech signal data point.

B. CONVOLUTIONAL NEURAL NETWORK LAYERS

The deep learning model with CNN provides advantages in terms of feature extraction. With the help of its distinctive convolutional kernel, it is possible to extract both local and global emotional information efficiently. To increase recognition accuracy, the pooling operation can simultaneously adjust to varying speech speeds and shifts in speech positions, to handle temporal variations in speech duration and align the extracted features effectively. CNN, on the other hand, has its foundation in sharing of weight and local

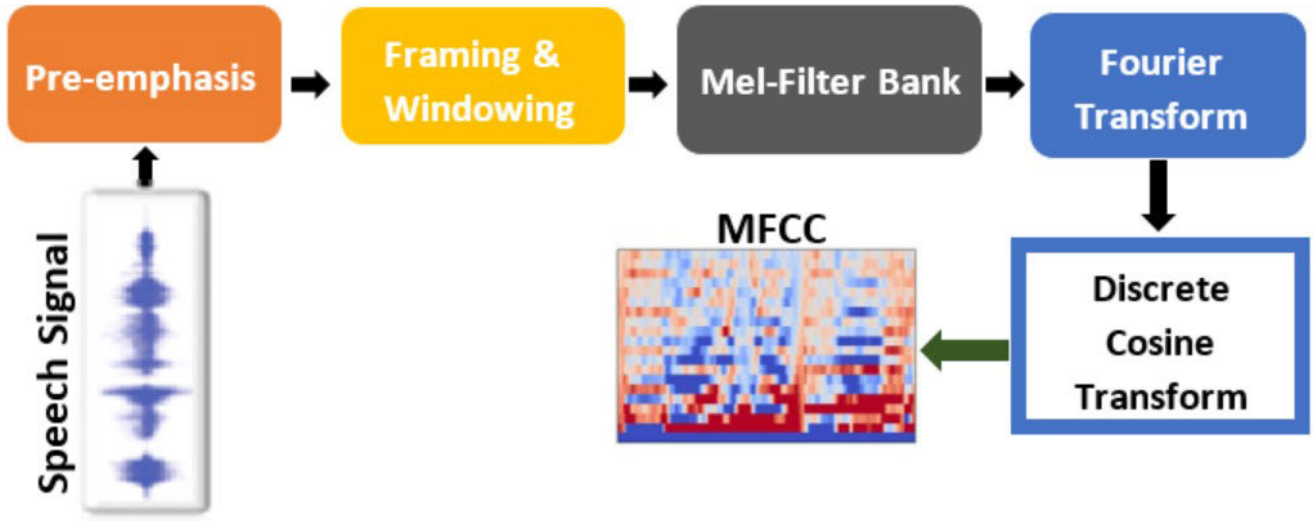


FIGURE 2. MFCC feature extraction from raw speech.

receptive fields [52]. The CNN model with fewer parameters, per the rules, requires comparatively less training data. Undoubtedly, a deeper convolutional network may extract high-dimensional aspects of emotion more effectively, but this typically comes at a hefty computational cost and is unsuitable for lightweight devices.

While some pre-trained deep learning models (VggNet, ResNet, MobileNet, etc.) and their variants enhance the flow of gradients and propagation of features through the sum of an identity function, a recent study [53] has revealed that a vast number of convolutional layers contribute very little to the outcomes. There are tons of trainable parameters generated by this. Contrarily, DenseNet places a strong emphasis on feature reuse through dense connectivity to address the issue and boost parameter effectiveness. Meanwhile, extensive connectivity enhances the flow of gradients and propagation of features even more.

Our proposed lightweight model is built on the VGGNet architecture for feature extraction, as shown in Figure 1. Figure 3 illustrates a comparison between the original VGGNet and our lightweight model. Three convolution blocks, one layer of batch normalization, one dense layer, and an input layer make up the proposed model. While the third convolution blocks include three convolution layers and one dropout layer, the first two convolution blocks only have two convolution layers and a max pooling layer. After passing through each convolution block, the input MFCC image's channel number steadily rises while its width and height are halved. The channel initially goes from 3 to 64, then to 128, 256, and finally to 512 channels, in that order. In generating the output image of $C_w/2$ and $C_h/2$, the max pooling operation divides the input channel's width, C_w, and height, C_h, into nearby pixels of 2×2 size. The fourth and fifth convolution block from the pre-trained VGGNet model has been eliminated from the proposed model to reduce the memory overhead. The top layer consisting of two fully connected and flatten layers were frozen from the conventional VGGNet. We adopted the weight from the original pre-trained VGGNet model on ImageNet which has been rained over four million images to prevent our model from training from scratch. The proposed model is designed to ensure adequate learning capacity with no overfitting [29], [54]

After the fourth convolution block, a layer of batch normalization is introduced as a way of normalizing the output and preventing model overfitting. Any layer of the neural network can undergo batch normalization, with the main goal being to achieve stable activation values that will lessen the inner covariate shift and prevent the over-fitting issue. The normalization of each dimension in the m-dimensional input, $x = (x^{(1)} \ldots x^{(d)})$, is computed in equation 4-6.

$$\hat{x}^{(m)} = \frac{x^{(m)} - E\left[x^{(m)}\right]}{\sqrt{Var\left[x^{(m)}\right]}} \tag{4}$$

where $x^{(m)}$ denotes each activation and

$$E\left[x^{(m)}\right] = \frac{1}{n}\sum_{j=1}^{n} x_j^m \tag{5}$$

denotes the mean while

$$Var\left[x^{(m)}\right] = \frac{1}{n}\sum_{j=1}^{n}\left(E\left[x^{(m)}\right]\right)^2 + \epsilon \tag{6}$$

denotes the variance and ϵ is the numerical constant added for stabilizing the output.

Proposed			Original VGGNet		
Layer(type)	**Output Shape**	**Parameters**	**Layer(type)**	**Output Shape**	**Parameters**
input_1(Input Layer)	[(None,224,224,3)]	0	input_1(Input Layer)	[(None,224,224,3)]	0
block1_conv1(Conv2D	[(None,224,224,64)]	1792	block1_conv1(Conv2D)	[(None,224,224,64)]	1792
block1_conv2(Conv2D)	[(None,224,224,64)]	36928	block1_conv2(Conv2D)	[(None,224,224,64)]	36928
block1_pool(MaxPooling2D)	[(None,112,112,64)]	0	block1_pool(MaxPooling2D)	[(None,112,112,64)]	0
block2_conv1(Conv2D)	[(None,112,112,128)]	73856	block2_conv1(Conv2D)	[(None,112,112,128)]	73856
block2_conv2(Conv2D)	[(None,112,112,128)]	147584	block2_conv2(Conv2D)	[(None,112,112,128)]	147584
block2_pool(MaxPooling2D)	[(None,56,56,128)]	0	block2_pool(MaxPooling2D)	[(None,56,56,128)]	0
block3_conv1(Conv2D)	[(None,56,56,256)]	295168	block3_conv1(Conv2D)	[(None,56,56,256)]	295168
block3_conv2(Conv2D)	[(None,56,56,256)]	590080	block3_conv2(Conv2D)	[(None,56,56,256)]	590080
block3_conv3(Conv2D)	[(None,56,56,256)]	590080	block3_conv3(Conv2D)	[(None,56,56,256)]	590080
Dropout_6(Dropout)	[(None,56,56,256)]	0	block3_conv3(Conv2D)	[(None,56,56,256)]	0
batch_normalization_3	[(None,56,56,256)]	1024	block4_conv4(Conv2D	[(None,28,28,512)]	1180160
dense_1(Dense)	[(None,7,7,512)]	1799	block4_conv4(Conv2D)	[(None,28,28,512)]	2359808
			block4_conv4(Conv2D)	[(None,28,28,512)]	2359808
			block4_pool(MaxPooling2D)	[(None,14,14,512)]	0
			block5_conv1(Conv2D	[(None,14,14,512)]	2359808
			block5_conv1(Conv2D)	[(None,14,14,512)]	2359808
			block5_conv1(Conv2D)	[(None,14,14,512)]	2359808
			block4_pool(MaxPooling2D)	[(None,7,7,512)]	0

FIGURE 3. The proposed model and original VGGNet architecture.

At the final layer, we added a dense layer that aggregates the input from the preceding layer and output feature vector according to the number of emotions (sad, angry, happy, disgust, calm, etc.) to be recognized. Output from this layer is fed into the classifier for the final recognition of emotion.

C. EMOTION RECOGNITION CLASSIFIERS

In recognition of emotion from this study, two classifiers were employed in this study which is Random Forest and Multilayer perceptron. The output from the dense layer after several convolutions is passed to the classifier for accurate classification of emotion into various categories. The description of the classifier is as follows:

1) RANDOM FOREST(RF) CLASSIFIER

Random forest is an ensemble learning technique for classification that is based on the class that most decision trees have chosen as their target as shown in Figure 4. In this situation, the features produced by convolution layers can be handled by downscaling variables since random forest excels at managing enormous input variables. Any size can be used because there is no need for a cross-validation set to guarantee an impartial estimate.

2) MULTI-LAYER PERCEPTRON (MLP) CLASSIFIER

In this study, a multi-layer perceptron (MLP) classifier is used for recognizing emotion as it receives input from the last layer of our proposed model. This feedforward artificial neural network model functions as a feedforward network-based classifier, mapping a set of appropriate outputs from a set of input datasets. Each layer that makes up an MLP is completely coupled to the layer below it. The nodes of the layers represent neurons with nonlinear activation functions,

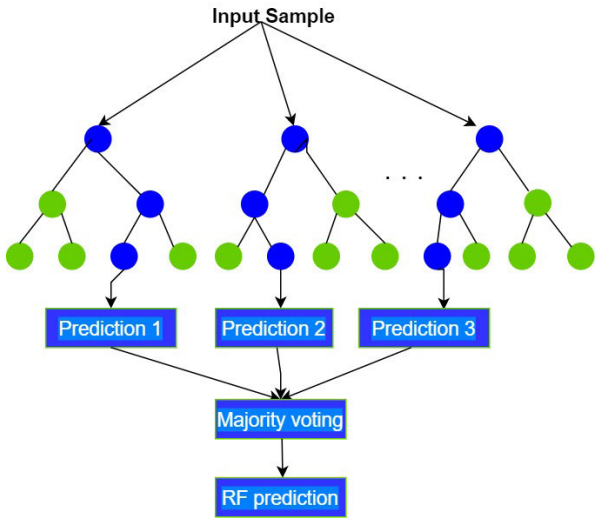


FIGURE 4. Random forest classifier structural view.

except the input layer nodes. We employed three hidden layers in our MLP classifier, with the first layer, second and third layers containing 128, 64 and 32 hidden units respectively.

IV. EXPERIMENTS AND RESULTS

We present significant details about the experimental setting, our experiments, and the analyses' findings in this section of the paper.

A. TESS DATASET

Many SER tasks have employed the Toronto English Speech Set [55], or TESS for short, one of the largest publicly accessible datasets. TESS speech samples were captured in 2010 at Northwestern University's Auditory Laboratory.

TABLE 1. Description of TESS speech dataset.

Emotion	Audio files Used	Percentage Ratio (%)
Angry	400	14.28
Sad	400	14.28
Disgust	400	14.28
Fear	400	14.28
Happy	400	14.28
Neutral	400	14.28
Surprise	400	14.28

TABLE 2. Description of RAVDESS speech dataset.

Emotion	Audio files Used	Percentage Ratio (%)
Angry	192	13.33
Sad	96	6.66
Fear	192	13.33
Boredom	192	13.33
Disgust	192	13.33
Happy	192	13.33
Neutral	192	13.33
Calm	192	13.33

Two actresses were instructed to speak a few of the 200 words during the spontaneous occurrence, and their voices were captured, creating a comprehensive collection of 2800 speech utterances. There were seven various emotions observed in the scenario, including joyful, furious, fear, disgust, pleasant, surprise, sad, and neutral. The description of TESS is illustrated in Table 1.

B. RAVDESS DATASET

A new English-language scripted emotional corpus, known as RAVDESS or Ryerson's audio-visual dataset of emotional song and speech [56], was completed in 2018. It is the most widely used dataset for identifying emotionally charged songs and speech. The recommended corpus, which consists of eight distinct emotions, was recorded by 24 professional individuals-12 men and 12 women-speaking scripts with altered emotions. The RAVDESS speech corpus is now mostly utilized for comparative analysis, which demonstrates the model's generalization as a result of the frequent use of emotions. It comprises 1440 sounds in all, collected at a sample rate of 48000 *Hz*. Table 2 provides a full description of the categories, audio utterances, and participation rate in percentage.

C. EMODB DATASET

This data set, also known as the Berlin emotion dataset [60] or the EMO-DB, is one of the most widely used. There are 535 speech utterances with seven different emotions in this well-known and popular speech emotion dataset. Ten expert people read prescript sentences and record different emotions for the suggested dataset, five males and five females. In the EMO-DB corpus, time is recorded with a sampling rate of 16*kHz* and an average of 2 to 3 seconds. A large number of emotion recognition techniques are based on the EMO-DB corpus, which is widely utilized in the SER domain.

TABLE 3. Description of EMODB speech emotion dataset.

Emotion	Audio files Used	Percentage Ratio (%)
Angry	127	23.73
Sad	62	11.58
Disgust	46	8.59
Neutral	79	14.76
Happy	71	13.27
Fear	69	12.89
Boredom	81	15.14

TABLE 4. Hyperparameter setting.

Hyperparameter	Value
Input size	224x224 x3
Optimizer	Adam
Learning rate	5e-5
Loss function	Sparse categorical crossentropy
Classifier random state	50
Number of estimator	100

An overview of selected emotions, total utterances, and participation ratio is shown in Table 3.

1) EXPERIMENTAL SETUP

The experiment for this study out was carried on 8GB RAM, 64bit OS, Intel core i7 device with Python 3.9 programming software. Table 4 shows the summary of the hyperparameter setting for the experiment. In selecting the best lightweight model architecture, we experimented with several deep learning lightweight architectures. Among them are DenseNet, VGG16, InceptionNetV3, ResNet, and MobileNet. Out of these, the VGGNet outperformed the rest in terms of accuracy and speed of emotion recognition. The feature map produced by the proposed lightweight model (Figure 5) for speech emotion recognition, it indicates whether the model is creating discriminative and meaningful representations. We can qualitatively evaluate if the model is capturing relevant speech features that are instructive for emotion recognition by looking at the feature maps. We gain insights into the learned representations and see how the model alters the input features to capture pertinent patterns and features for the speech emotion recognition task by visualizing the feature maps at various layers. Table 5 shows the performance of our model in comparison with other architecture tested, where the EMODB dataset (represented as DT1), RAVDESS dataset (represented as DT2) and TESS dataset (represented as DT3) respectively with highest accuracy recorded on DT3 when estimators value is 100 and random state set to 50. The dataset is split into an 80:20 ratio for the training and testing set. Both the test and training sets of data had their pixel values normalized to range from 0 to 1. Besides, the model size is minimal with optimum accuracy compared to the existing ones.

2) EXPERIMENTAL RESULTS

The performance comparison of our proposed model with other studies is highlighted in Table 7 following the speech

FIGURE 5. Feature map from the first convolutional layer.

TABLE 5. Lightweight model selection.

Model	Size in MB (Lightweight)	No. of Parameters	Dataset	Accuracy(%)	Average Accuracy
DenseNet	7.04	7,043,654	DT1 DT2 DT3	86.13 79.24 91.60	85.65
VGGNet	**7.64**	**7,640,903**	**DT1** DT2 DT3	**96.03** 86.25 100.00	**94.09**
InceptionNet	9.64	9,604,544	DT1 DT2 DT3	62.30 75.60 88.20	75.36
MobileNet	4.80	4,806,855	DT1 DT2 DT3	48.50 81.32 31.96	53.92
ResNet	9.12	9,116,032	DT1 DT2 DT3	87.12 83.90 98.00	89.67

TABLE 6. Low-memory device configuration.

Device	Specification
Hardware	Raspberry Pi 4 Model B
Processor	Quad-core Cortex-A72 (ARMv8) 64-bit SoC @ 1.5GHz
RAM	2GB LPDDR4
Operating System	Raspbian OS
Software Frameworks	TensorFlow Lite, Python 3.7
Compilation	Model optimized using TensorFlow Lite for ARM architecture

TABLE 7. Performance comparison.

Reference	Year	Dataset	Accuracy Reported (%)
[57]	2018	**EMODB**	82.82
[41]	2019		84.49
[58]	2019		88.99
[59]	2020		85.57
[60]	2020		90.01
[49]	2020		92.02
[61]	2021		93.00
[62]	2022		90.01
Proposed	**2023**		**96.03**
		RAVDESS	
[63]	2019		75.79
[64]	2019		67.14
[59]	2020		77.01
[65]	2022		81.82
Proposed	**2023**		**86.25**
		TESS	
[66]	2017		96.00
[67]	2018		89.96
[68]	2019		89.96
[69]	2021		93.30
[70]	2022		96.10
Proposed	**2023**		**100.00**

emotion database used. Our model outperforms other proposed methods in all three datasets from an accuracy and computational complexity point of view. The highest accuracy recorded by other researchers on EMODB was 93.00%, 81.82% on RAVDESS, and 96.10% on TESS, but our model supersedes them all. Crucial to the recognition of speech emotion in real time-application is time, and our proposed model has an average clock of 0.07 seconds (CPU time) in the classification of a single emotional utterance because the number of parameters has been reduced drastically. Besides, our model outperforms what was obtained in [57] when comparing the model size(323.46 to 7.94) and accuracy(82.82% to 96.03%) of emotion recognition on the EMODB dataset. To evaluate the performance and efficiency of our lightweight SER model on low-resource devices, we conducted an experiment using the hardware and configuration as shown in Table 6. We obtain an average processing time of 0.15 seconds per utterance and 0.02 seconds standard deviation. The processing time measurement includes all the steps from pre-processing to emotion recognition. Specifically, it incorporates feature extraction, model loading, and forward pass through the model.

3) DISCUSSION

The significance of feature extraction in the speech emotion recognition model is very important. However, some features

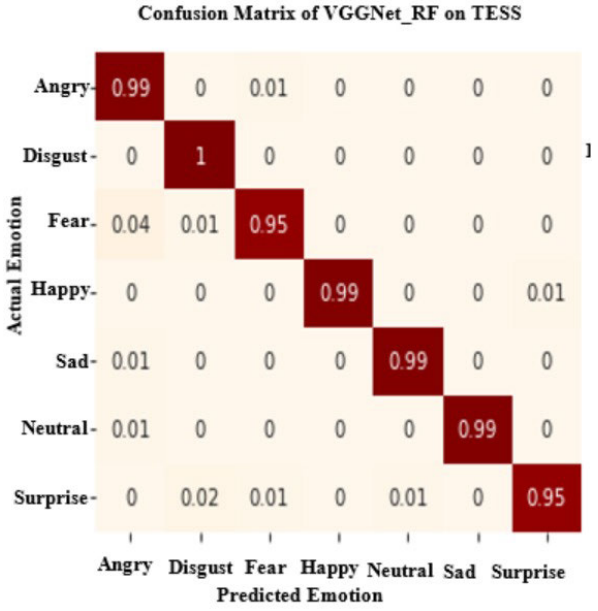


FIGURE 6. Confusion matrix of the proposed model on TESS datasets with an average recognition accuracy of 98%.

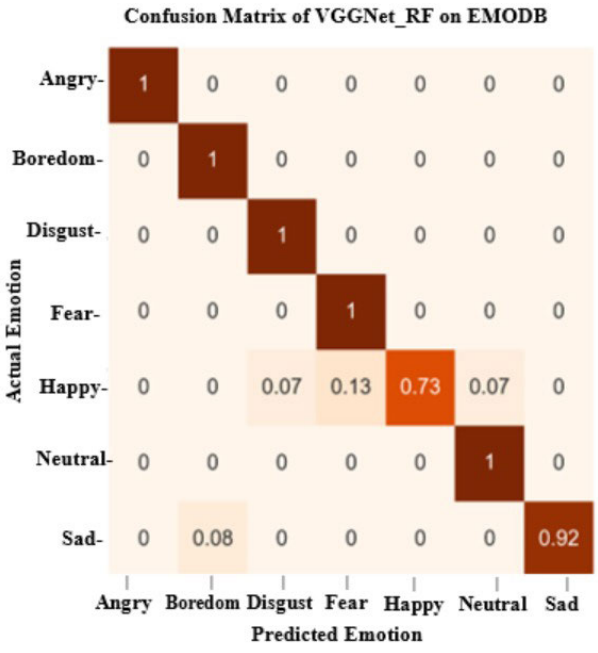


FIGURE 7. Confusion matrix of the proposed model on EMODB datasets with an average recognition accuracy of 95.5%.

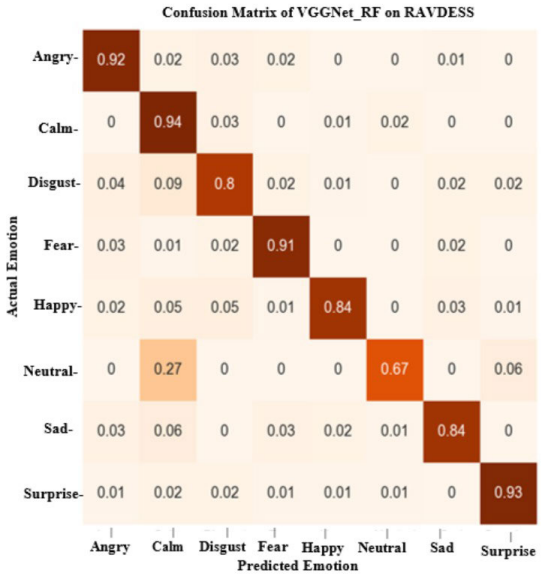


FIGURE 8. Confusion matrix of the proposed model on RAVDESS datasets with an average recognition accuracy of 85.6%.

can be difficult to extract owing to environmental factors and the expected number of features to be extracted. From the input MFCC image, the features extracted by each layer of the convolutions differ. The deeper the convolution, the higher the number of features that will be extracted. Many a time these features range from generic ones to more specific ones (high-level features). However, computational complexity in terms of resources, space, and training time has been a major drawback. Our lightweight VGGNet utilized the transfer learning techniques, where weights from the original pre-trained VGGNet model that has been trained on millions of datasets was transferred. Two major criteria were used in optimizing our proposed model, which are training-time and accuracy. Our extensive experiment indicated that at the third convolutional block with additional drop out, batch, and dense layer, an optimal accuracy was recorded. Hence, the remaining convolutional layer and block from the main VGGNet have been expunged.

The confusion matrix for this study is shown in Figure 6-9. With our Lightweight model, we achieved the highest accuracy of recognition (100%) on TESS with MLP classifier with 7 emotions. With the confusion matrix, there is a deep insight that showcases the misperception between the predicted emotional classes and the actual emotional classes, along with other emotions at corresponding rows. The two axes that existed in the confusion matrix represent expected predictions (x-axis) and actual predictions (y-axis).

Our proposed SER lightweight model achieved overall recognition accuracy of 100%, 96.03%, and 86.25% for TESS, EMODB, and RAVDESS speech datasets respectively. To further the investigation of the model performance, Figure 10 illustrate the emotional-level prediction for each of the dataset. Angry, boredom, and fear emotional class shows the highest recognition rate of 100% on the EMODB dataset with both classifiers, while only Angry indicates the highest recognition accuracy with both classifiers for the TESS dataset. For RAVDESS, angry and calm emotions show the highest accuracy of 94% from both classifiers. The lowest accuracy recorded was on Neutral emotion with a random forest classifier on the RAVDESS dataset. The experimental results obtained have in no doubt, shown the efficiency and

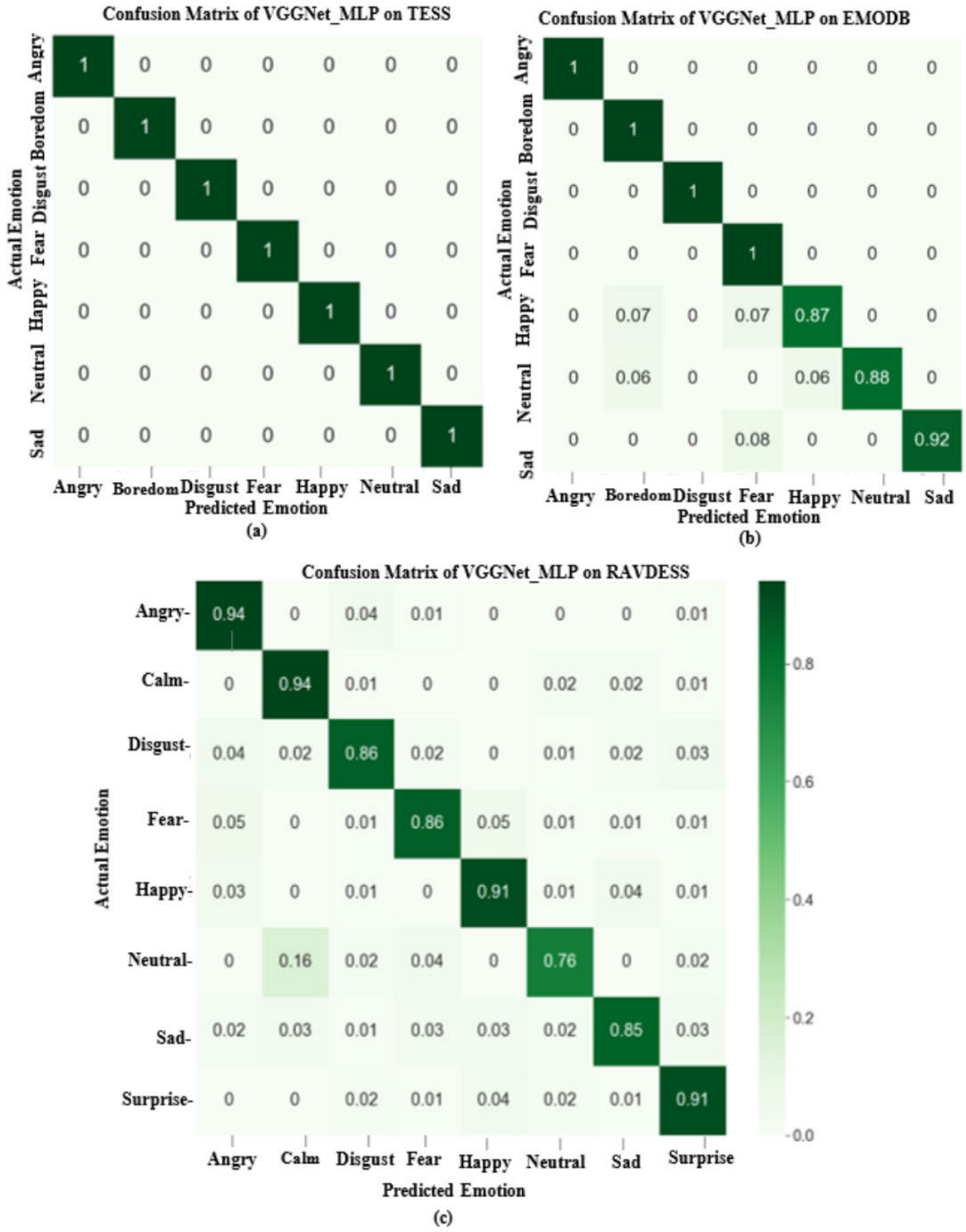


FIGURE 9. Confusion matrix of the proposed model on three datasets with MLP Classifier with 100%, 90.83% and 87.87% average recognition accuracy respectively.

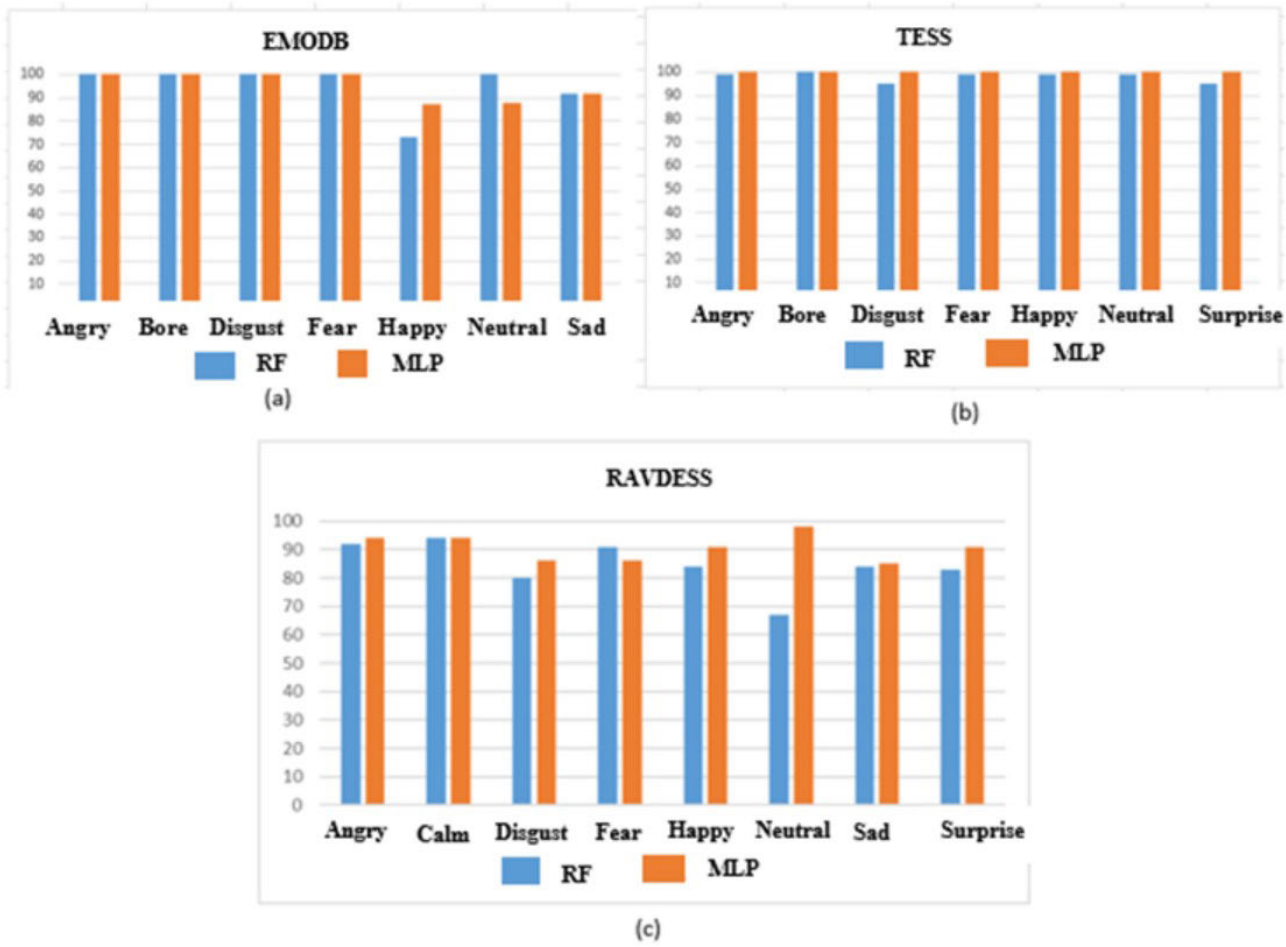


FIGURE 10. Emotion class prediction comparison on two classifiers with three datasets.

robustness of our proposed lightweight model for SER, with an improved performance over state-of-the-art techniques.

V. CONCLUSION

In this work, we have proposed a lightweight speech emotion recognition model that uses an end-to-end approach for feature extraction and classification through the best-performing classifier. Our VGGNet SER model has been optimized such that model size, computational complexity, and recognition time has been simplified, while the accuracy of recognition has been improved. The robustness and efficiency of our model on three popular datasets, TESS, EMODB, and RAVDESS achieved recognition accuracy of 100%, 96.03%, and 86.02% respectively. Our extensive experiments with the proposed model have proven that it is suitable for real-life application because it requires lesser time for recognition of emotion and has the capability for generalization. The proposed lightweight model has improved on emotion recognition systems especially for low-memory devices because of its moderate size (7.94Mb). However, in the future, we intend to investigate the possibility of integrating our model with other deep learning architecture (self-attention and transformer), and audio pre-trained models to extract more salient features from speech to improve emotion recognition. Besides, we will also take a closer at experimenting larger size emotional speech database to establish the performance of the model in SER-related tasks. The result of our experiment on a low-memory device highlights the practicality and efficiency of our proposed approach for real-time applications on such devices, however, we have a limitation on the number of these devices available at our disposal as at the time of this study, but we intend to explore more low-memory devices in our future research. The link to our implementation code to foster collaboration can be found here: https://github.com/samsoftcom1/Speech-Emotion2023.git and https://www.kaggle.com/code/samsonadebisi/speech-emotion-recognition-using-attention-network/.

REFERENCES

[1] T. S. Ustun, S. M. S. Hussain, L. Yavuz, and A. Onen, "Artificial intelligence based intrusion detection system for IEC 61850 sampled values under symmetric and asymmetric faults," *IEEE Access*, vol. 9, pp. 56486–56495, 2021.

[2] F.-K. Wang, T. Mamo, and X.-B. Cheng, "Bi-directional long short-term memory recurrent neural network with attention for stack voltage degradation from proton exchange membrane fuel cells," *J. Power Sources*, vol. 461, Jun. 2020, Art. no. 228170, doi: 10.1016/j.jpowsour.2020.228170.

[3] Y. Liu and G. Fu, "Emotion recognition by deeply learned multi-channel textual and EEG features," *Future Gener. Comput. Syst.*, vol. 119, pp. 1–6, Jan. 2021, doi: 10.1016/j.future.2021.01.010.

[4] C. Yu, M. Kang, Y. Chen, J. Wu, and X. Zhao, "Acoustic modeling based on deep learning for low-resource speech recognition: An overview," *IEEE Access*, vol. 8, pp. 163829–163843, 2020, doi: 10.1109/ACCESS.2020.3020421.

[5] S. Mekruksavanich, A. Jitpattanakul, and N. Hnoohom, "Negative emotion recognition using deep learning for Thai language," in *Proc. Joint Int. Conf. Digit. Arts, Media Technol. ECTI Northern Sect. Conf. Electr., Electron., Comput. Telecommun. Eng.*, Mar. 2020, pp. 71–74, doi: 10.1109/ECTIDAMTNCON48261.2020.9090768.

[6] X. Lu, "Deep learning based emotion recognition and visualization of figural representation," *Frontiers Psychol.*, vol. 12, p. 818833, 2022, doi: 10.3389/fpsyg.2021.818833.

[7] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[8] S. Lugovic, I. Dunder, and M. Horvat, "Techniques and applications of emotion recognition in speech," in *Proc. 39th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2016, pp. 1278–1283.

[9] T. L. Nwe, S. W. Foo, and L. C. D. Silva, "Speech emotion recognition using hidden Markov model," *Speech Commun.*, vol. 10, no. 4, pp. 603–623, 2003;.

[10] L. Trinh Van, T. D. T. Le, T. Le Xuan, and E. Castelli, "Emotional speech recognition using deep neural networks," *Sensors*, vol. 22, no. 4, p. 1414, Feb. 2022, doi: 10.3390/s22041414.

[11] E. L. R. Ewe, C. P. Lee, L. C. Kwek, and K. M. Lim, "Hand gesture recognition via lightweight VGG16 and ensemble classifier," *Appl. Sci.*, vol. 12, no. 15, p. 7643, Jul. 2022, doi: 10.3390/app12157643.

[12] N. Kim and S. Kim, "A study on user experience of online education programs with elementary schools and art museums in non-face-to-face era," *J. Digit. Converg.*, vol. 19, no. 8, pp. 311–317, 2021.

[13] K. Feng and T. Chaspari, "A Siamese neural network with modified distance loss for transfer learning in speech emotion recognition," 2020, *arXiv:2006.03001*.

[14] O. Fagbuagun, O. Folorunsho, L. Adewole, and H. Akin-Olayemi, "Breast cancer diagnosis in women using neural networks and deep learning," *J. ICT Resour. Appl.*, vol. 16, no. 2, pp. 152–166, 2022, doi: 10.5614/itbj.ict.res.appl.2022.16.2.4.

[15] J. Amherst and J. Jhun, "Speech emotion recognition using biometric features," *IEEE Access*, vol. 8, pp. 12452–12463, 2020.

[16] W. Zheng, Z. Wenming, and Z. Yuan, "Multi-scale discrepancy adversarial network for cross-corpus speech emotion recognition," *Virtual Reality Intell. Hardw.*, vol. 3, no. 1, pp. 65–75, 2021, doi: 10.1016/j.vrih.2020.11.006.

[17] Y. Gou, S. Wang, and C. Feng, "Speech emotion recognition based on parallel convolutional neural networks," in *Proc. 2017 Int. Joint Conf. Neural Netw.*, 2017, pp. 3789–3794.

[18] Z. Wang, H. Seng-Beng, and E. Cambria, "A review of emotion sensing: Categorization models and algorithms," *Multimedia Tools Appl.*, vol. 79, pp. 35553–35582, Jan. 2020, doi: 10.1016/j.vrih.2020.11.006.

[19] A. M. Badshah, "Deep features-based speech emotion recognition for smart affective services," *Multimedia Tools Appl.*, vol. 78, no. 5, pp. 5571–5589, 2019.

[20] S. A. Ajagbe, K. A. Amuda, M. A. Oladipupo, O. F. Afe, and K. I. Okesola, "Multi-classification of Alzheimer disease on magnetic resonance images (MRI) using deep convolutional neural network (DCNN) approaches," *Int. J. Adv. Comput. Res.*, vol. 11, no. 53, pp. 51–60, Mar. 2021, doi: 10.19101/IJACR.2021.1152001.

[21] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: A review," *Complex Intell. Syst.*, vol. 8, no. 3, pp. 2663–2693, Jun. 2022, doi: 10.1007/s40747-021-00637-x.

[22] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, pp. 103–126, Jul. 2020.

[23] Y. Yu and Y.-J. Kim, "Attention-LSTM-attention model for speech emotion recognition and analysis of IEMOCAP database," *Electronics*, vol. 9, p. 713, 2020, doi: 10.3390/electronics9050713.

[24] B. T. Atmaja and A. Sasou, "Effects of data augmentations on speech emotion recognition," *Sensors*, vol. 22, no. 16, p. 5941, Aug. 2022, doi: 10.3390/s22165941.

[25] Q. Lin, H. Feng, and H. Yin, "Emotion recognition from speech using convolutional neural network and transfer learning," *IEEE Access*, vol. 7, pp. 94059–94068, 2019.

[26] X. Wu, W.-L. Zheng, Z. Li, and B.-L. Lu, "Investigating EEG-based functional connectivity patterns for multimodal emotion recognition," *J. Neural Eng.*, vol. 19, no. 1, Feb. 2022, Art. no. 016012, doi: 10.1088/1741-2552/ac49a7.

[27] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[28] C. A. Kumar and K. A. Sheela, "Emotion recognition from speech biometric system using machine learning algorithms," in *Advances in Communications, Signal Processing, and VLSI*, T. Laxminidhi, J. Singhai, S. R. Patri, and V. V. Mani, Eds., vol. 722. Singapore: Springer, 2021, doi: 10.1007/978-981-33-4058-9_6.

[29] S. Kwon, "Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network," *Int. J. Intell. Syst.*, vol. 36, no. 9, pp. 5116–5135, Sep. 2021, doi: 10.1002/int.22505.

[30] S. Sahu, R. Gupta, and C. Espy-Wilson, "Modeling feature representations for affective speech using generative adversarial networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 1098–1110, Apr. 2022, doi: 10.1109/TAFFC.2020.2998118.

[31] A. Shahid, S. Latif, and J. Qadir, "Generative emotional AI for speech emotion recognition: The case for synthetic emotional speech augmentation," 2023, *arXiv:2301.03751*.

[32] B. Pan and W. Zheng, "Emotion recognition based on EEG using generative adversarial nets and convolutional neural network," *Comput. Math. Methods Med.*, vol. 2011, Oct. 2021, Art. no. 2520394, doi: 10.1155/2021/2520394.

[33] J. L. Bautista, Y. K. Lee, and H. S. Shin, "Speech emotion recognition based on parallel CNN-attention networks with multi-fold data augmentation," *Electronics*, vol. 11, no. 23, p. 3935, Nov. 2022, doi: 10.3390/electronics11233935.

[34] Y. Gao, D. Zhang, and H. Li, "Emotion recognition from conversation using natural language processing," *Brain Informat.*, vol. 8, no. 1, p. 5162, 2021.

[35] F. Makhmudov, A. Kutlimuratov, F. Akhmedov, M. S. Abdallah, and Y.-I. Cho, "Modeling speech emotion recognition via attention-oriented parallel CNN encoders," *Electronics*, vol. 11, no. 23, p. 4047, Dec. 2022, doi: 10.3390/electronics11234047.

[36] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, Feb. 2021, doi: 10.3390/s21041249.

[37] A. Winursito, "Improvement of MFCC feature extraction accuracy using PCA," in *Proc. Indonesian Speech Recognit. Int. Conf. Inf. Commun. Technol.*, 2018, pp. 379–383.

[38] L. Muyawei, G. Hernndez, C. Antonio, A. Carlos, W. Xuetian, and G. Hongmin, "Speech emotion recognition using convolutional-recurrent neural networks with attention model," in *Proc. 2nd Int. Conf. Comput. Eng., Inf. Sci. Internet Technol.*, 2017, pp. 341–350, doi: 10.12783/dtcse/cii2017/17273.

[39] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97515–97525, 2019, doi: 10.1109/ACCESS.2019.2928625.

[40] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008, doi: 10.1007/s10579-008-9076-6.

[41] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," *IEEE Access*, vol. 7, pp. 90368–90377, 2019, doi: 10.1109/ACCESS.2019.2927384.

[42] M. Prau, A. Tiwari, R. K. Singh, and R. Yadav, "Speech emotion recognition and features extraction based on NN classifier," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 5, pp. 2852–2857, 2020.

[43] P. P. Chimthankar. (2021). *Speech Emotion Recognition using Deep Learning*. School of Computing National College of Ireland. [Online]. Available: http://norma.ncirl.ie/5142/1/priyankaprashantchimthankar.pdf
[44] O. Atila and A. Şengür, "Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition," *Appl. Acoust.*, vol. 182, Nov. 2021, Art. no. 108260, doi: 10.1016/j.apacoust.2021.108260.
[45] A. Aggarwal, A. Srivastava, A. Agarwal, N. Chahal, D. Singh, A. A. Alnuaim, A. Alhadlaq, and H. Lee, "Two-way feature extraction for speech emotion recognition using deep learning," *Sensors*, vol. 2022, pp. 1–11, Jan. 2022, doi: 10.3390/s22062378.
[46] V. Singh and S. Prasad, "Speech emotion recognition system using gender dependent convolution neural network," *Proc. Comput. Sci.*, vol. 218, pp. 2533–2540, Jan. 2023, doi: 10.1016/j.procs.2023.01.227.
[47] Y. Zhong, Y. Hu, H. Huang, and W. Silamu, "A lightweight model based on separable convolution for speech emotion recognition key laboratory of multilingual information technology," in *Proc. Interspeech*, 2020, pp. 3331–3335.
[48] K. Atsavasirilert, T. Theeramunkong, S. Usanavasin, A. Rugchatjaroen, S. Boonkla, J. Karnjana, S. Keerativittayanun, and M. Okumura, "A lightweight deep convolutional neural network for speech emotion recognition using Mel-spectrograms," in *Proc. 14th Int. Joint Symp. Artif. Intell. Natural Language Process. (iSAI-NLP)*, Oct. 2019, pp. 1–4.
[49] T. Anvarjon and S. Kwon, "Deep-Net: A lightweight CNN-based speech emotion recognition system using deep frequency features," *Sensors*, vol. 20, no. 18, p. 5212, Sep. 2020, doi: 10.3390/s20185212.
[50] S. Chen, M. Zhang, X. Yang, Z. Zhao, T. Zou, and X. Sun, "The impact of attention mechanisms on speech emotion recognition," *Sensors*, vol. 21, no. 22, p. 7530, Nov. 2021, doi: 10.3390/s21227530.
[51] M. M. Lynn, C. Su, and K. K. Maw, "Efficient feature extraction for emotion recognition system," in *Proc. 4th Int. Conf. Converg. Technol. (I2CT)*, Oct. 2018, pp. 1–6, doi: 10.1109/I2CT42659.2018.9058313.
[52] M. Gokilavani, H. Katakam, S. A. Basheer, and P. Srinivas, "RAVDNESS, CREMA-D, TESS based algorithm for emotion recognition using speech," in *Proc. 4th Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Tirunelveli, India, Jan. 2022, pp. 1625–1631, doi: 10.1109/ICSSIT53264.2022.9716313.
[53] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger, "Deep networks with stochastic depth," 2016, *arXiv:1603.09382*.
[54] F. Zhu-Zhou, R. Gil-Pita, J. García-Gómez, and M. Rosa-Zurera, "Robust multi-scenario speech-based emotion recognition system," *Sensors*, vol. 22, no. 6, p. 2343, Mar. 2022, doi: 10.3390/s22062343.
[55] K. Dupuis and M. K. Pichora-Fuller, "Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set (TESS)," *Can. Acoust. Acoust. Can.*, vol. 39, pp. 182–183, 2011.
[56] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391, doi: 10.1371/journal.pone.0196391.
[57] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.
[58] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D Log-Mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.
[59] M. Sajjad and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.
[60] S. Kwon, "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114177.
[61] S. Kwon, "ATT-Net: Enhanced emotion recognition system using lightweight self-attention module," *Appl. Soft Comput.*, vol. 102, Apr. 2021, Art. no. 107101, doi: 10.1016/j.asoc.2021.107101.
[62] E. Guizzo, T. Weyde, S. Scardapane, and D. Comminiello, "Learning speech emotion representations in the quaternion domain," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1200–1212, 2023.
[63] A. Bhavan, P. Chauhan, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowl.-Based Syst.*, vol. 184, Nov. 2019, Art. no. 104886.
[64] A. A. A. Zamil, S. Hasan, S. M. D. J. Baki, J. M. D. Adam, and I. Zaman, "Emotion detection from speech signals using voting mechanism on classified frames," in *Proc. Int. Conf. Robot., Elect. Signal Process. Techn. (ICREST)*, Dhaka, Bangladesh, Jan. 2019, pp. 281–285, doi: 10.1109/ICREST.2019.8644168.
[65] C. Luna-Jimnez, R. Kleinlein, D. Griol, Z. Callejas, J. M. Montero, and F. Fernndez-Martnez, "A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset," *Appl. Sci.*, vol. 12, p. 327, Jan. 2022.
[66] D. Verma and D. Mukhopadhyay, "Age driven automatic speech emotion recognition system," in *Proc. Int. Conf. Comput., Commun. Autom. (ICCCA)*, Apr. 2016, pp. 1005–1010, doi: 10.1109/CCAA.2016.7813862.
[67] V. Praseetha and S. Vadivel, "Deep learning models for speech emotion recognition," *J. Comput. Sci.*, vol. 14, no. 11, pp. 1577–1587, 2018, doi: 10.3844/jcssp.2018.1577.1587.
[68] Y. Gao. (2019). *Speech-Based Emotion Recognition*. [Online]. Available: https://libraetd.lib.virginia.edu/downloads/2f75r8498?filename=1GaoYe2019MS.pdf
[69] P. Krishnan, A. J. Raj, and V. Rajangam, "Emotion classification from speech signal based on empirical mode decomposition and non-linear features," *Complex Intell. Syst.*, vol. 7, no. 4, pp. 1919–1934, 2021, doi: 10.1007/s40747-021-00295-z.
[70] S. Akinpelu and S. Viriri, "Robust feature selection-based speech emotion classification using deep transfer learning," *Appl. Sci.*, vol. 12, no. 16, p. 8265, Aug. 2022, doi: 10.3390/app12168265.

SAMSON AKINPELU received the B.Sc. and M.Tech. degrees in computer science from the University of Kwazulu-Natal, South Africa, where he is currently pursuing the Ph.D. degree in computer science. He has over three years of teaching experience with Federal University Oye Ekiti, Nigeria. His main research interests include artificial intelligence, computer vision, deep leaning, speech emotion recognition, pattern recognition, and natural language processing.

SERESTINA VIRIRI (Senior Member, IEEE) received the B.Sc. degree in mathematics and computer science and the M.Sc. and Ph.D. degrees in computer science. He is a Full Professor of computer science with the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, South Africa. He has been in academia, since 1998. He has published extensively in several artificial intelligence and computer vision-related accredited journals and international and national conference proceedings. His main research interests include artificial intelligence, computer vision, image processing, machine learning, medical image analysis, pattern recognition, and other image processing related fields, such as biometrics, medical imaging, and nuclear medicine. He serves as a reviewer for several machine learning and computer vision-related journals. He has also served on program committees for numerous international and national conferences. He is a Rated Researcher by the National Research Foundation (NRF) of South Africa.

ADEKANMI ADEGUN received the B.Tech., M.Sc., and Ph.D. degrees in computer science. He has a lecturing experience in universities, for nearly ten years. He has also co-supervised M.Sc. and Ph.D. candidates in machine learning fields. He has published extensively in several artificial intelligence and computer vision-related accredited journals and international and national conference proceedings. His main research interests include artificial intelligence, computer vision, image processing, machine learning, medical image analysis, pattern recognition, and natural language processing. He serves as a reviewer for some machine learning and computer vision-related journals.

Chapter 5

Speech Emotion Classification Enhancement using Deep Learning Transformer Models

5.1 Introduction

The published works in the deep learning transformer model-based speech emotion classification are presented in this chapter. It provides a summary of the speech emotion classification model that was used in this study. Pre-processing techniques for speech signals, transformer architecture types, and the dataset utilized are presented.

5.2 An Enhanced Speech Emotion Recognition using Vision Transformer

5.2.1 Brief Review

The paper entitled An Enhanced Speech Emotion Recognition using Vision Transformer: ViTSER improved on the previous paper in terms of memory efficiency with fewer trainable parameters(4,166,151) while extracting emotional features from the speech signals. Mel-spectrogram representation was extracted from raw speech signal with 128M mel bins to map the frequency onto the Mel scale. Each audio sound was split into frames of 25ms, with a 10ms gap between each frame, to avert feature degradation. An ablation study was also conducted on the functional components of the model architecture to ascertain its robust capability to quickly classify emotion on TESS, EMODB, and hybridized TESS-EMODB datasets. The result obtained in the paper proved superior performance when compared with other architectures.

Paper status: Published in Springer, Scientific Reports

scientific reports

Check for updates

OPEN

An enhanced speech emotion recognition using vision transformer

Samson Akinpelu, Serestina Viriri✉ & Adekanmi Adegun

In human–computer interaction systems, speech emotion recognition (SER) plays a crucial role because it enables computers to understand and react to users' emotions. In the past, SER has significantly emphasised acoustic properties extracted from speech signals. The use of visual signals for enhancing SER performance, however, has been made possible by recent developments in deep learning and computer vision. This work utilizes a lightweight Vision Transformer (ViT) model to propose a novel method for improving speech emotion recognition. We leverage the ViT model's capabilities to capture spatial dependencies and high-level features in images which are adequate indicators of emotional states from mel spectrogram input fed into the model. To determine the efficiency of our proposed approach, we conduct a comprehensive experiment on two benchmark speech emotion datasets, the Toronto English Speech Set (TESS) and the Berlin Emotional Database (EMODB). The results of our extensive experiment demonstrate a considerable improvement in speech emotion recognition accuracy attesting to its generalizability as it achieved 98%, 91%, and 93% (TESS-EMODB) accuracy respectively on the datasets. The outcomes of the comparative experiment show that the non-overlapping patch-based feature extraction method substantially improves the discipline of speech emotion recognition. Our research indicates the potential for integrating vision transformer models into SER systems, opening up fresh opportunities for real-world applications requiring accurate emotion recognition from speech compared with other state-of-the-art techniques.

Keywords Human–computer interaction, Deep learning, Speech emotion recognition, CNN, Vision transformer, Mel spectrogram

Human–computer interactions (HCI) can be improved by paying more attention to emotional cues in human speech[1]. The need for speech recognition and enhancement of emotion recognition in achieving more natural interaction and better immersion is becoming more of a challenge as a result of the growing trend in artificial intelligence (AI)[2,3]. Coincidentally, with the development of deep neural networks, research on Speech Emotion Recognition (SER) systems has grown steadily by turning audio signals into feature maps that vividly describe the vocal traits of speech(auditory) samples.[4].

Speech Emotion Recognition (SER) is a classification problem that seeks to classify audio samples into pre-defined emotions. SER has applications in affective computing, psychological wellness evaluation, and virtual assistants, and has become a crucial field of research in human–computer interaction[5]. Speech signals may be used to reliably detect and comprehend human emotions, which enables machines to react correctly and produce more interesting and tailored interactions[6]. By acquiring acoustic features from speech signals[7], such as pitch, energy, and spectral qualities, and using machine learning algorithms to categorize emotions based on these features, has been the concentration of conventional approaches (Fig. 1) to SER[8]. Although these methods have yielded encouraging results, they frequently fail to pick up on nuances in emotional cues and are subject to noise and unpredictability in voice signals.

Researchers have been able to improve SER by using the spectral features of an audio sample as an image input to the impressive advancements in computer vision. Convolutional neural networks (CNNs), in particular, have shown astounding performance in deep learning[9] models for visual tasks like image processing and object detection. The weights of several convolutional layers have been utilized to create feature representations in this architecture[10,11]. Utilizing mel-spectrograms, this method can be used in SERs to convert audio data into visual audio signals based on its frequency components. Then, these representations that resemble images can be trained using a CNN network. Traditional CNN, however, only accepts a single frame as input and does not

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban 4001, South Africa. ✉email: viriris@ukzn.ac.za

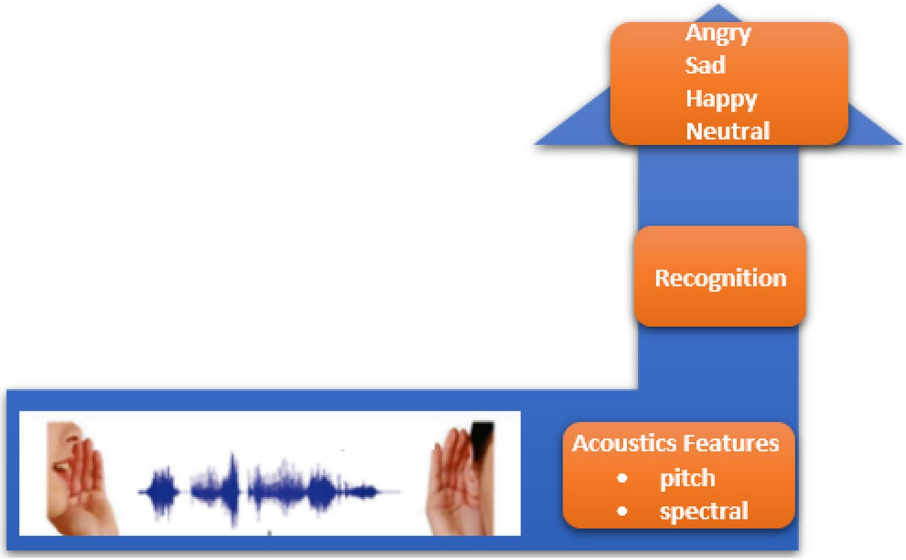


Figure 1. Traditional speech emotion recognition framework.

compute over a timestep sequence, therefore they are unable to remember previous data from the same sample while processing the subsequent timestamp.

Additionally, because of the number of parameters generated by the numerous convolutional layers, they provide large levels of computational complexity[12]. Researchers have been seeking alternative architectures that are more appropriate for handling visual data in the context of SER as a result of this constraint.

The Vision Transformer (ViT) is one such architectural design that has attracted significant interest. The ViT model, which was initially introduced for image classification tasks, completely changed the area of computer vision by exhibiting competitive performance without utilizing conventional CNN building blocks[13]. The ViT model makes use of a self-attention mechanism that enables it to directly learn global features from the input image and capture spatial dependencies. This unique model has demonstrated promising performance in several computer vision applications[14], raising the question of whether it may be leveraged to enhance SER.

In this study, we addressed two core issues. At first, the computational complexity is reduced, we enhanced the accuracy of emotion recognition from speech signals by improving the state-of-the-art performance. We focus mainly on extracting features from the mel-spectrogram[15] and fed it into a novel lightweight ViT model with a self-attention mechanism for accurate speech emotion recognition. The spectrogram image is represented in time and frequency as width and length to enable our proposed model to learn emotionally rich features from speech signals. Computational cost is reduced as a result of fewer over-blotted parameters. The major contributions of this work are highlighted below.

- We proposed a novel lightweight Vision Transformer (ViT) model with self-attention for learning deep features related to emotional cues from the mel-spectrogram to recognize emotion from speech.
- Complexity is reduced for SER through fewer parameters and efficient layers that learn discriminative features from the input image.
- We evaluated the proposed SER model on popular benchmark datasets which include TESS and EMO-DB. The result of our comparative experiments shows an improved performance in SER, confirming its suitability for real-time application.

The remaining part of this paper is split into other sections as follows. Section 2 presents the reviewed literature and related works, Section 3 highlights the proposed methodology and its detailed description. In Section 4, the experimental configuration, result and discussion are presented, while Section 5 illustrates the conclusion and future work to foster research progress in the SER domain.

Review of related works

The study of emotion recognition from speech signals as it plays a crucial role in behavioural patterns and enhances human–computer interaction in the past decade has come a long way. Identification of human emotional conditions from speech samples (natural or synthetic) has formed the basis for the development of Speech Emotion Recognition SER systems. Core among these emotional states are angry, sad, happy, neutral, etc. Researchers began with the conventional approach of recognizing these emotions with the use of orthodox machine learning models which includes Support Vector Machine(SVM)[16–18], Gaussian Mixture Model(GMM)[19],k-nearest Neighbour(KNN)[20] and Hidden Markov Model (HMM)[21] among others. However, these classical machine learning classifiers are bewildered with the problem of high susceptibility to noise and the inability to efficiently handle large audio speech samples.

Therefore, neural network approaches such as Recurrent Neural Networks (RNN)[22] and Long Short Term Memory (LSTM)[23–25] have been proposed by researchers in the SER domain, because of their capability to handle sequence(time series) data and learn temporal information that is critical to emotion recognition using contextual dependencies. The adoption of these two techniques has littered several SER literature, because emotion recognition has been improved upon. However, RNN is prone to gradient descent problems[26]

The common approach to SER in recent came as a result of unimaginable success through deep learning techniques[27,28] and prominent among this approach are Convolutional Neural Networks (CNN)[29], Deep Neural Networks(DNN)[30–32], Deep Belief Networks(DBN)[33] and Deep Boltzman Machine (DBM)[34]. In Zeng et al.[35] spectrogram feature extracted from Rayson Audio-Visual Database of Emotional Speech and Song(RAVDESSS) speech dataset was fed into DNN with gated residual network which yielded 65.97% accuracy of emotion recognition on tested data. In the same vein, a pre-trained VGG-16 convolutional neural network was utilized in Popova et al.[36] and they achieved an accuracy of 71% after extensive experiments. To increase the possibility of improving the recognition rate, the author Issa et al.[37] proposed a novel Deep Convolutional Neural Network (DNN) for SER. Multiple features Similarlyrances were extracted such as Mel Frequency Cepstral Coefficient, spectral contrast, and Mel-Spectrogram, and were fused to serve as their model input. Their method arrived at 71.61% accuracy for recognising eight different emotions from the RAVDESS dataset. Their method was also experimented on EMODB and IEMOCAP datasets for generalizability. However, their CNN model could not efficiently capture the spatial features and sequences peculiar to speech signals. In addressing the foregone, a multimodal approach of deep learning and temporal alignment techniques was proposed by Li et al.[38]. In their method, CNN, LSTM and Attention Mechanism were combined and they achieved the highest accuracy of 70.8% with semantic embeddings.

In recent times as well, the combination of CNN, LSTM or RNN for SER tasks has recorded significant improvement[39]. This approach relies heavily on the extraction of features from raw speech signals with CNN and passing them into the LSTM or RNN for extraction of long-term dependencies features that are peculiar to emotion recognition from auditory utterances[40]. Puri et al.[41] implemented a hybrid approach of utilizing LSTM and DNN on the RAVDESS dataset. They extracted MFCC from raw speech signals and fed it into their model. The ensemble technique of extracting salient features from speech utterances and passing the emotional features into a classifier, irrespective of the language and cultural background of the speakers has also aroused the interest of researchers in the SER field. High-level features from speech signals were extracted using DBN and then later fed into a Support Vector Machine classifier for emotion classification in Schuller et al.[42]. Similarly, Zhu et al.[43] utilized DNN and SVM and experimented with the efficiency of their model on the Chinese Academy of Chinese-based dataset. A separate study by Pawar et al.[44] proposed a deep learning approach for SER. Relevant features were extracted from speech signals using MFCC, as input to train the CNN model. They achieve a significant result of 93.8% accuracy on the EMODB dataset. The author in[45] proposed innovative lightweight multi-acoustic features-based DCNN techniques for speech emotion recognition. In their method, various features such as Zero Crossing Rate(ZCR), wavelet packet transform (WPT), spectral roll-off, linear prediction cepstral coefficients (LPCC), pitch, etc. were extracted and fed into one-dimensional DCNN and they obtained 93.31% on Berlin Database of Emotional Speech(EMODB) and 94.18% on RAVDESS respectively. Badshah et al.[46] presented present a double CNN-based model for SER with spectrogram from an audio signal. They utilized a pooling mechanism and kernel of different sizes with spectrogram input generated using Fast Fourier Transform (FFT). Their approach validates the importance of max-pooling operation in CNN.

The introduction of audio transformer to speech paralinguistics has contributed immensely to emotion recognition from speech signals. It involves analysis and synthesis of speech signals with features that are non-verbal[47]. Chen et al[48] proposed a novel full-stack audio transformer (WavLM) for speech analysis using a speech denoising approach for learning general speech representations from huge unannotated data. The performance of their proposed transformer model, benchmarked on the SUPERB dataset achieved a state-of-the-art result and improved many speech-related tasks such as speech emotion recognition and speaker verification or identification. Xu et al.[49] proposed a novel speech transformer-based that incorporated self-attention and local dense synthesizer attention (LDSA) for extracting both local and global features from speech signals. In a bid to enhance the efficiency of end-to-end speech recognition models while lowering computing complexity, the technique eliminates pairwise interactions and dot products and limits attention scope to a narrow region surrounding the current frame. A novel hybrid-based audio transformer, named Conformer-HuBERT was implemented by Shor et al.[50]. Their mechanism achieves a significant milestone in emotion recognition from speech signals and other paralinguistic tasks by learning from many large-scale unannotated data. Again, Chen et al.[51] proposed a novel SpeechFormer technique that combines the distinctive features of speech signals into transformer models. A hierarchical encoder that uses convolutional and pooling layers to shorten the input sequence is one of the three components of the framework. Another is a local self-attention module that records dependencies inside a predetermined window size, and a global self-attention module that records dependencies across various windows. Paraformer is another novel speech transformer model for non-autoregressive end-to-end speech recognition that employs parallel attention and parallel decoder approaches, introduced by Gao et al.[52]. The framework enables independent prediction of each output token without reliance on prior output tokens, and permits each decoder layer to handle all encoder outputs concurrently without waiting for previous decoder outputs. The study demonstrates that Paraformer achieves faster inference speed and higher accuracy on multiple speech recognition datasets compared to existing non-autoregressive models.

In the immediate past, efforts towards improving the efficiency of deep learning model performance and conquering the challenge of long-range dependencies peculiar to the CNN-base model for SER have been increased. The state-of-the-art transformer model has been introduced into SER[53]. A parallel architecture that utilized the ResNet and Transformer model was proposed in Han et al.[54]. Vijay et al.[55] implemented an audio-video multimodal transformer for emotion recognition. They adopted three self-attention and block embedding

to capture relevant features from spectrogram images. Their model achieved 93.59%, 72.45%, and 99.17% on RAVDESS, CREMA-D and SAVEE datasets respectively, but huge computing resources were required because of the architecture. Not quite long after, Slimi et al.[56] proposed a transformer-based CNN for SER, with hybrid time distribution. They leverage the superior capability of the transformer and achieve a promising result of 82.72% accuracy. However, such a model is prone to high computational complexity due to huge parameters. The ability of CNN-based models to recognize long-range dependencies in speech signals is constrained by the fact that they frequently operate on fixed-size input windows. Speech emotion frequently displays temporal dynamics outside of the speech sequence's local regions. Therefore, we proposed a lightweight Vision Transformer (ViT) model comprised of a self-attention mechanism[57] that enables it to capture global contextual information, making it possible to model long-range dependencies and enhance the representation of emotional speech patterns, hence improving speech emotion recognition.

Additionally, while a couple of research studies have looked at how to include visual cues in speech emotion recognition, they frequently treat visual and auditory modalities independently, resulting in an insufficient fusion of information or features. This study seeks to leverage the synergistic effects of multimodal information, enabling a more thorough comprehension of emotions and enhancing the accuracy of the SER system by using the ViT model[58,59], capable of capturing salient features from the speech signal.

Proposed method

In this section, we delve into the overview of our proposed model (Fig. 2) for SER. We highlighted the overall details from speech collection, pre-processing, feature extraction, and feeding of ViT with feature vectors that eventually lead to emotion recognition.

Speech pre-processing

When background noise cannot be tolerated, pre-processing the speech sound is a crucial step. These systems, such as speech emotion recognition (SER) require effective feature extraction from audio files, where the majority of the spoken component consists of salient characteristics connected to emotions. This study used pre-emphasis and silent removal strategies to reach its goal[60]. Pre-emphasis uses Eq. (1) to increase the high-frequency parts of speech signals. The pre-emphasis technique can improve the signal-to-noise ratio by enhancing high frequencies in speech while leaving low frequencies untouched through the Finite impulse response (FIR) mechanism.

$$H(z) = 1 - \alpha z^{-1}, \alpha = [1, -0.97] \tag{1}$$

where z is the signal and α is the energy level change across the frequency

Contrariwise, Eq. (2) is used in signal normalization to ensure that speech signals are equivalent despite any differences in magnitude.

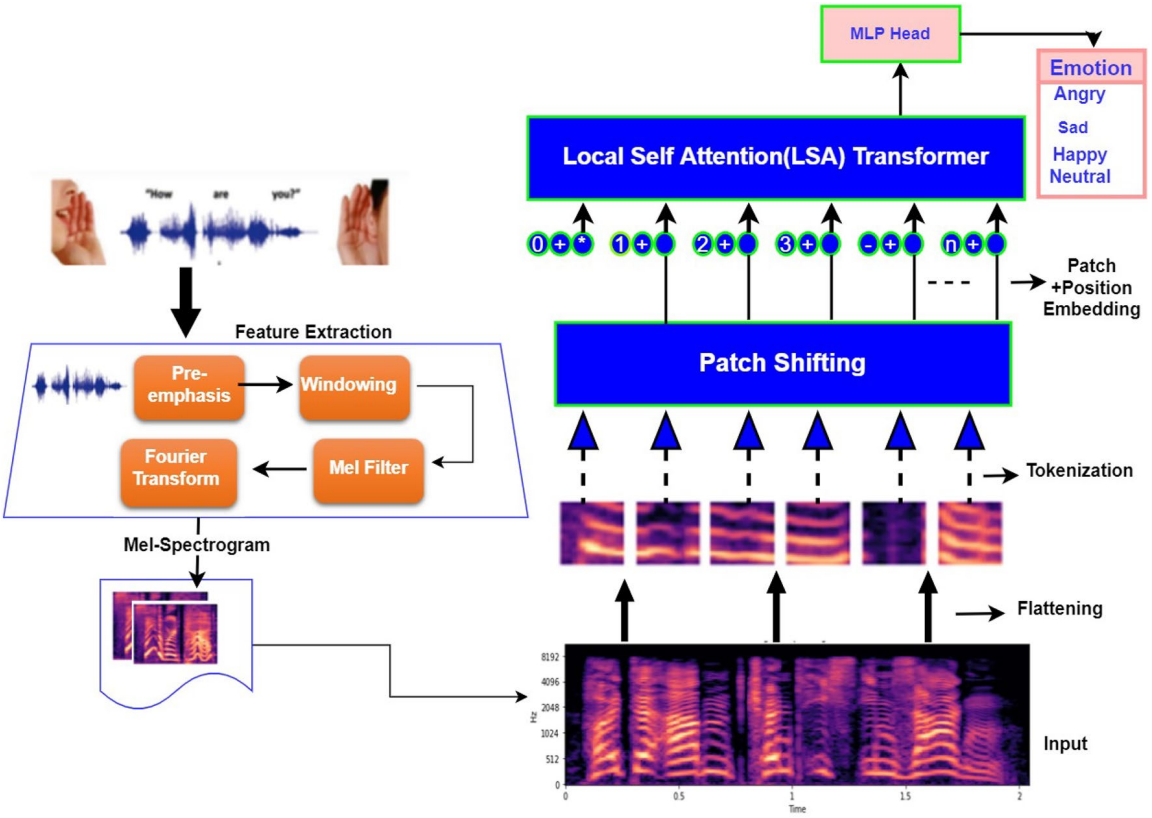


Figure 2. Propose Vision Transformer Architectural Framework.

$$S_{Ni} = \frac{S_i - \mu}{\sigma} \tag{2}$$

where the signal's mean and standard deviation are represented by μ and, σ respectively, while the signal's $i^t h$ portion is denoted by the S_i. The normalized $i^t h$ component of the signal is referred to as SN_i.

Extraction of mel-spectrogram Feature

The quality of the feature set heavily influences the recognition performance of the model. As a result, inappropriate features could produce subpar recognition outcomes. To achieve acceptable recognition performance in the context of Deep Learning (DL), extracting a meaningful feature set is a vital task. According to[61], feature extraction is a crucial step in deep learning since the SER model's success or failure depends heavily on the variability of the features it uses to do the recognition task. If the derived traits have a strong correlation with the emotion class, recognition will be accurate, but if not, it will be challenging and inaccurate. The performance of recognition in SER is strongly influenced by the quality of the feature set.

The process of mel-spectrograms (Fig. 3) feature extraction involves pre-emphasis, framing, windowing and the discrete Short Time Fourier Transform. In our method, we generate a mel-spectrogram image by converting each speech sound sample into a 2D time-frequency matrix. We perform the discrete Short-Time Fourier Transform (STFT) computation for this. We employ an STFT length of 1024, hop size of 128, and 1024 window size (using Hanning as the window function). Additionally, we used 128*Mel* bins to map the frequency onto the Mel scale. Each audio sound was split into frames of 25 ms, with a 10 ms gap between each frame, to avert information degradation. After the framing and windowing, we applied several mel-filter banks and the mel denotes the ears' perceived frequency, which is computed using Eq. 3.

$$Mel(f) = 295 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{3}$$

where f represent the real frequency and $Mel(f)$ represent the corresponding frequency of perception.

Vision transformer

Vision transformers are becoming the standard in the NLP (Natural Language Processing) domain. The attention mechanism is an important element of such a model. It may extract useful features from the input using a typical query, key, and value structure, where the similarity between queries and keys is pulled out by matrix multiplication between queries and keys. In order to effectively extract many scales, multiple resolutions, and high-level spatial features, vision transformers use a multi-head attention mechanism. The global average pooling system is then used to up-sample and concatenate the dense feature maps that have been produced. To be able to successfully learn and extract the intricate features relevant to emotion recognition in mel-spectrogram image, the method makes use of both local and global attention, as well as global average pooling. As illustrated in our proposed architecture, the entire model ranges from flattening to the classification of emotion. The input image is broken up into patches of defined size, fattened and linearly embedded, added to position embedding, and then transferred to the Transformer encoder.

The Vision Transformers have much less image-specific inductive bias than CNNs, hence, we leverage its capability to classify seven human emotions: angry, sad, disgust, fear, happiness, neutral and surprise as shown in our model. Our proposed vision transformer model for SER is not heavy, unlike many baseline models. It comprises 4, 166, 151 total and trainable parameters, with 0 non-trainable parameters, thereby it reduces

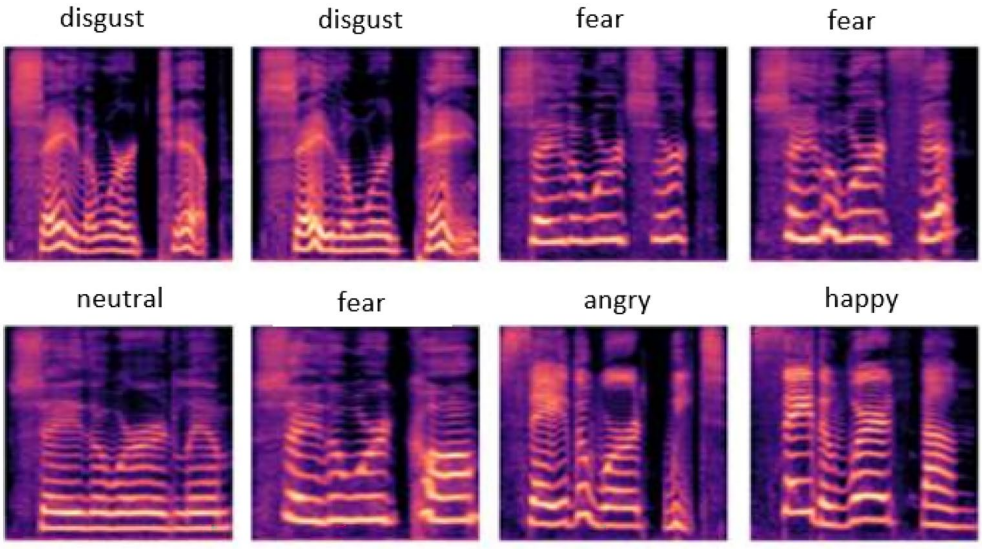


Figure 3. Mel-spectrogram of selected emotion.

computational complexity. In the first stage, a feature vector of shape $(n+1, d)$ is created by embedding an input image(spectrogram) of shape (height, width, and channels) into it[62]. Then, in raster order, the image is splatted into n square patches of shape (t, t, c), where t is a pre-defined value. Patches are then flattened, producing n line vectors with the shape $(1, t^2 * c)$. The flattened patches are multiplied by a trainable embedding tensor of shape $(t^2 * c, d)$ that learns to linearly project each flat patch to dimension d. Our model dimension is 128, with 32 patch sizes.

The ViT model's functional components and corresponding functions in the model architecture are succinctly summarized by the functional components as shown in Table 1. Collectively, they improve the ViT model's ability to identify spatial dependencies and extract relevant representations from speech signals for recognition of speech emotions.

Core module analysis of ViT

The proposed ViTSER model in this study utilizes two core audio transformer modules which are self-attention and multi-head attention. The first mechanism is self-attention, which computes representations for the inputs by relating various positions of input sequences. It employs three specific inputs: values (V), keys (K), and queries (Q). The result of a single query is calculated as the weighted sum of the values, with each weight being determined by a specially constructed query function that uses the associated key. Here, we employ an efficient self-attention method that is based on Dot-product[63]as computed in Eq. 4.

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{QK^T}{\sqrt{d_k}})\mathbf{V} \tag{4}$$

where the softmax function is prevented from entering regions with extremely small gradients by using the scalar $\frac{1}{\sqrt{d_k}}$.

Secondly, another core module of the audio transformer is multi-head attention, which is used to simultaneously exploit several attending representations. The calculation of multi-head attention is h times scaled Dot-Product Attention, where h is the number of heads. Three linear projections are used before each attention for transforming the queries, keys, and values, respectively, into more discriminating representations. Next, as shown in Eq. 5, each Scaled Dot-Product Attention is computed separately and its outputs are concatenated.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)\mathbf{W}^O \tag{5}$$

where $head_i = Attention(\mathbf{QW}_i^Q, \mathbf{KW}_i^K, \mathbf{VW}_i^V)$

We employed an activation function known as Gaussian Error Linear Unit (GELU), a high-performing activation function in many speech-related tasks and NLP[64] as compared to RELu (Reactivation Linear Unit). Rather than gating inputs by their sign as in ReLUs, the GELU non-linearity weights inputs according to their value. The GELU activation function is $x\Phi(x)$, for an input x is defined from Eq. 6.

$$GELU(x) = x\Phi(x) = x.\frac{1}{2}\left[1 + erf(x/\sqrt{2}\right] \tag{6}$$

where $\Phi(x)$ denotes the standard Gaussian cumulative distribution function.

Experimental result

In this section, the full details of how we carried out our extensive experiment and evaluation of our model are highlighted. To demonstrate the significance and robustness of our model for the SER utilizing speech spectrograms, we effectively validate our system in this part using two benchmark TESS and EMODB speech datasets. Using the same phenomena, we evaluated the effectiveness of our SER system and contrasted it with other baseline SER systems. The next sections go into further detail on the datasets that were used, the accuracy matrices, and the results of the study.

SN	Components	Description
1	Patch Embeddings	The initial representation of linearly projected image patches with each patch having a vector representation.
2	Positional Enconding	Provides the input embeddings positional information, which enables the model to comprehend the spatial relationship between several patches.
3	Transformer Encoder	Comprised of feedforward neural network modules and several layers of self-attention that capture high-level features and long-range dependencies from the speech signal.
4	Self-Attention	A method for capturing the dependencies among the various patches in the input sequence which enables the model to focus on relevant information across the whole input sequence.
5	Layer-Normalization	Stabilizes training and enhances generalization by normalizing each layer's activations.
6	Dropout	Regularization method that, during training, randomly sets a portion of the input units to zero thereby, increases the robustness of the model and helps avoid overfitting.

Table 1. Functional Components of the ViTSER.

Datasets

TESS

The Toronto English Speech Set, or TESS for short, one of the largest freely available datasets, has been used in numerous SER projects. The Auditory Laboratory at Northwestern University recorded TESS speech samples in 2010[65]. During the spontaneous event, two actors were given instructions to pronounce a couple of the 200 words. Their voices were recorded, providing a comprehensive collection of 2800 speech utterances. Seven different feelings were seen in the scenario: happy, angry, scared, disgusted, pleasant, surprised, sad, and neutral. Figure 4 provides an illustration of the TESS description based on each emotion's contribution to the whole speech dataset.

EMODB

EMOD is one of the most predominantly utilized datasets, commonly known as the Berlin emotion dataset[66] or the EMO-DB. This well-known and well-liked dataset of speech emotions contains 535 voice utterances expressing seven different emotions. Five men and five women, all experts, read prescriptive words and recorded various emotions for the suggested dataset. Time is captured with a sampling rate of 16 kHz and an average duration of 2 to 3 seconds in the EMO-DB corpus. Every utterance has the same temporal scale, allowing the entire speech to fit within the window size. The EMO-DB corpus, which is widely used in the SER field, forms the foundation for several emotion recognition algorithms. Figure 5 illustrates the summary of the overall utterances, participation rate, and selected emotions.

Model implementation

The primary framework for the model implementation uses PyTorch[67] components. We modified the size of the images during the pre-processing stage to accommodate the dimensions of $224x224$ on three separate channels (corresponding to the RGB channels); we have delved into more depth about speech data preprocessing in the previous section. The experiment was carried out on a computing resource that includes $GPU - 10900K@3.70Ghz$, 64GB RAM, and the Google Colab platform. We utilized the Adam optimizer with sparse categorical cross entropy loss function and $3.63E - 03$ as the learning rate during the training phase. We obtained optimum accuracy at 75 epoch. Finally, using a simple momentum of 0.9, we accelerated training and variable learning by the experiment's chosen optimizer. Two public datasets (TESS and EMODB) are used, with the combination of the two datasets to form the third set of datasets (TESS-EMOD) for assessing the

Figure 4. TESS dataset emotion distribution.

Figure 5. EMODB dataset emotion distribution.

performance and generalizability of our model. The overall description of the hyperparameters utilized in this work is highlighted in Table 2

Evaluation metrics

Standard metrics are typically used to evaluate the effectiveness of deep learning models for emotion identification tasks. Based on several performance criteria, including precision, recall, accuracy, and F1-score as provided in Eqs. (6)–(9), the proposed method's results are contrasted. Precision and recall reflect the qualitative and quantitative performance of the proposed SER system, whilst accuracy represents the percentage of accurate predictions out of the total number of cases analyzed. Recall (sensitivity) measures the proportion of actual positive cases from all actual positive cases, while precision measures the proportion of true positive (TP) cases from all predicted positive cases. The harmonic mean of the precision and recall are provided by the F1-score[68].

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TN}{TN + FN} \tag{8}$$

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{TP + TN}{TP + TN + FP + FN}\right) \tag{9}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

Furthermore, we adopted the confusion matrix metric which gives a more meaningful insight into the outcome of our experiment. It uses variables such as FP (false positive), FN (false negative), TP (true positive), and TN (true negative)[69] in depicting the combinations of true and predicted classes from a given speech dataset.

Results of experiments and discussion

This section describes the result of our extensive experiments carried out to assess the performance of our proposed model for speech emotion recognition tasks. The collection of tests is utilized to assess how well the model recognizes unknown speech utterances. The system generalization error is approximately represented by the model prediction error[70]. The cross-validation estimation approach is used in this study to thoroughly assess each dataset. The database's data is divided into two categories: training data and testing data. There are k fragments to the original data in which the k part of the data is utilized for training, while one portion is used as test data. K-fold cross-validation is a term used to describe the test procedure, which is carried out k times across various portions of all the data[71]. For an in-depth assessment of our technique, we applied a well-known 5-fold cross-validation assessment method. The visual representation of the model loss is shown in Fig. 6. The uniqueness of our proposed model as displayed in the figure, indicates its effectiveness as the loss decreases on both training and testing data. The highest loss value for the three experiments were 0.13, 0.2 and 0.25 on TESS, EMODB and TESS-EMODB respectively.

According to the speech databases used, which include a variety of emotions-seven distinct ones-selected following Ekman's[72] postulation. We investigated the proposed model and presented the emotional level prediction performance in Tables 3, 4 and 5 together with the resulting confusion matrices. Our model's prediction performance displays precision, recall, F1-Score, weighted results, and un-weighted results, which amply demonstrates the model's superiority over state-of-the-art techniques. According to the detailed classification(emotional level prediction) report, it is obvious that the highest recognition was obtained for precision, F1-score and recall on neutral emotion with 100% from the TESS dataset, followed by disgust with 99% from EMODB respectively, and the least recall rate was recorded on boredom with 76%.

Hyperparameter	Value
Number of Epochs	75
Learning rate	3.63E-03
Activation function	GELU
Embedded dropout rate	0.1
Trainable parameters	4,166,151
Patch size	32
MLP dimension	128
Optimizer	Adam
Loss Function	Flattened Loss of Cross Entropy

Table 2. Hyperparameters employed for this study.

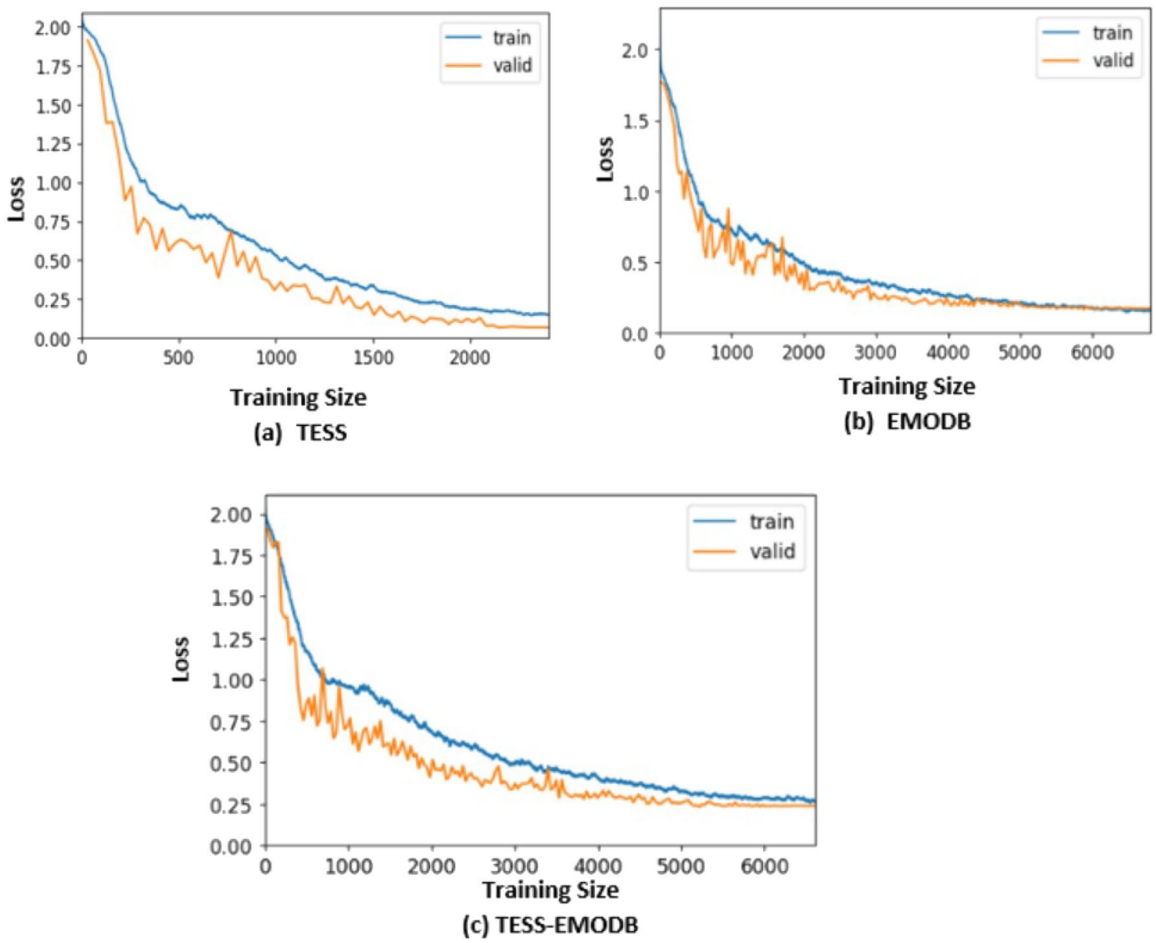


Figure 6. The figure illustrates the proposed model's performance loss curve for the three benchmarked datasets. (**a**) Loss diagram on TESS dataset (**b**) Loss diagram on EMODB dataset and (**c**) Loss diagram on TESS-EMODB dataset.

Emotion	Precision (%)	Recall (%)	F1-score (%)
Angry	100	98	99
Disgust	99	99	99
Fear	99	99	99
Happy	96	96	96
Neutral	100	100	100
Sad	100	98	99
Surprise	92	97	94
Accuracy	–	–	98
Weighted Average	98	98	98

Table 3. Emotional level prediction for TESS dataset.

We summarized the classification report in the above tables for each emotion using 3 metrics on 6 emotions as shown in Fig. 7. Our method demonstrates higher performance than the state-of-the-art approach in terms of the overall recognition of emotions, especially for disgust, neutral, sad and fear respectively. Our model recognizes the emotions from the frequency pixels and salient features to enhance recognition accuracy and mitigate the overall computational cost. Most of the baseline models detected disgust emotions with low accuracy because of their paralinguistic content, however, our model outperformed others with high precision and recall of 99% with only happy emotion demonstrating the least recognition of 82% recall.

In furtherance of our investigation, we obtain a confusion matrix for the three datasets to show a class-wise recognition accuracy as shown in Fig. 8. We achieved the highest recognition accuracy from the confusion matrix on angry, neutral and disgust with 99%, 98% and 95% respectively. Only boredom emotion showed the least

Emotion	Precision (%)	Recall (%)	F1-score (%)
Angry	90	96	92
Boredom	88	76	81
Disgust	99	93	96
Fear	95	85	89
Happy	82	89	85
Neutral	88	96	92
Sad	94	94	94
Accuracy	–	–	91
Weighted Average	92	91	91

Table 4. Emotional level prediction for EMODB dataset.

Emotion	Precision (%)	Recall (%)	F1-score (%)
Angry	87	96	91
Disgust	92	95	93
Fear	97	93	95
Happy	94	87	90
Neutral	97	92	94
Sad	100	99	99
Accuracy	–	–	93
Weighted Average	94	93	93

Table 5. Emotional level prediction for TESS-EMODB dataset.

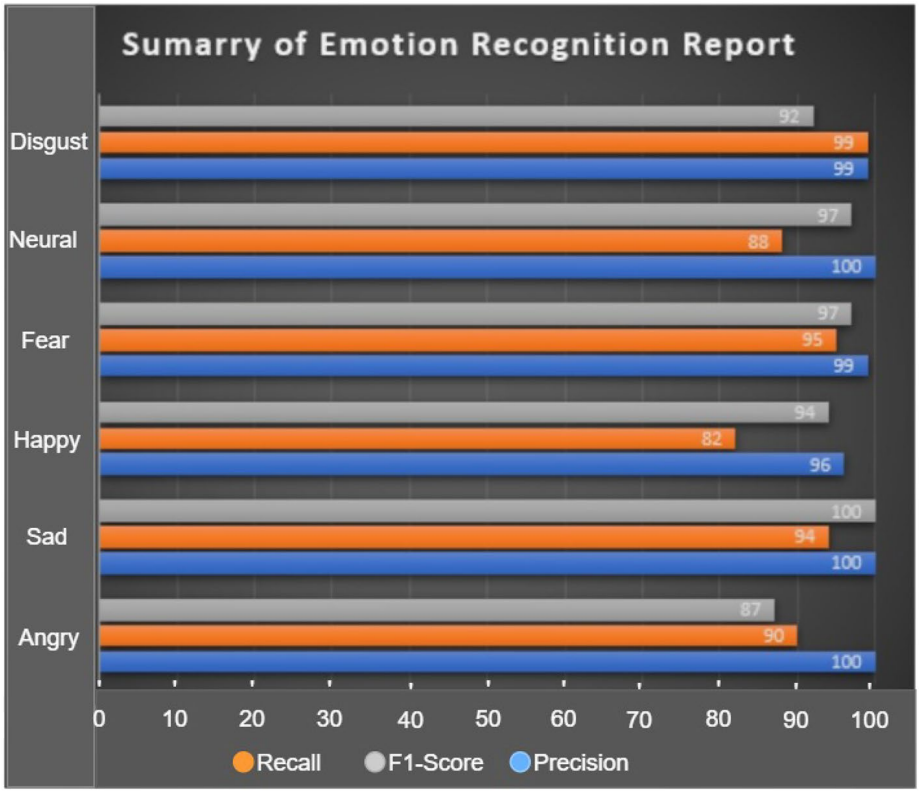


Figure 7. Summary of classification report for F1-Score, Recall and Precision.

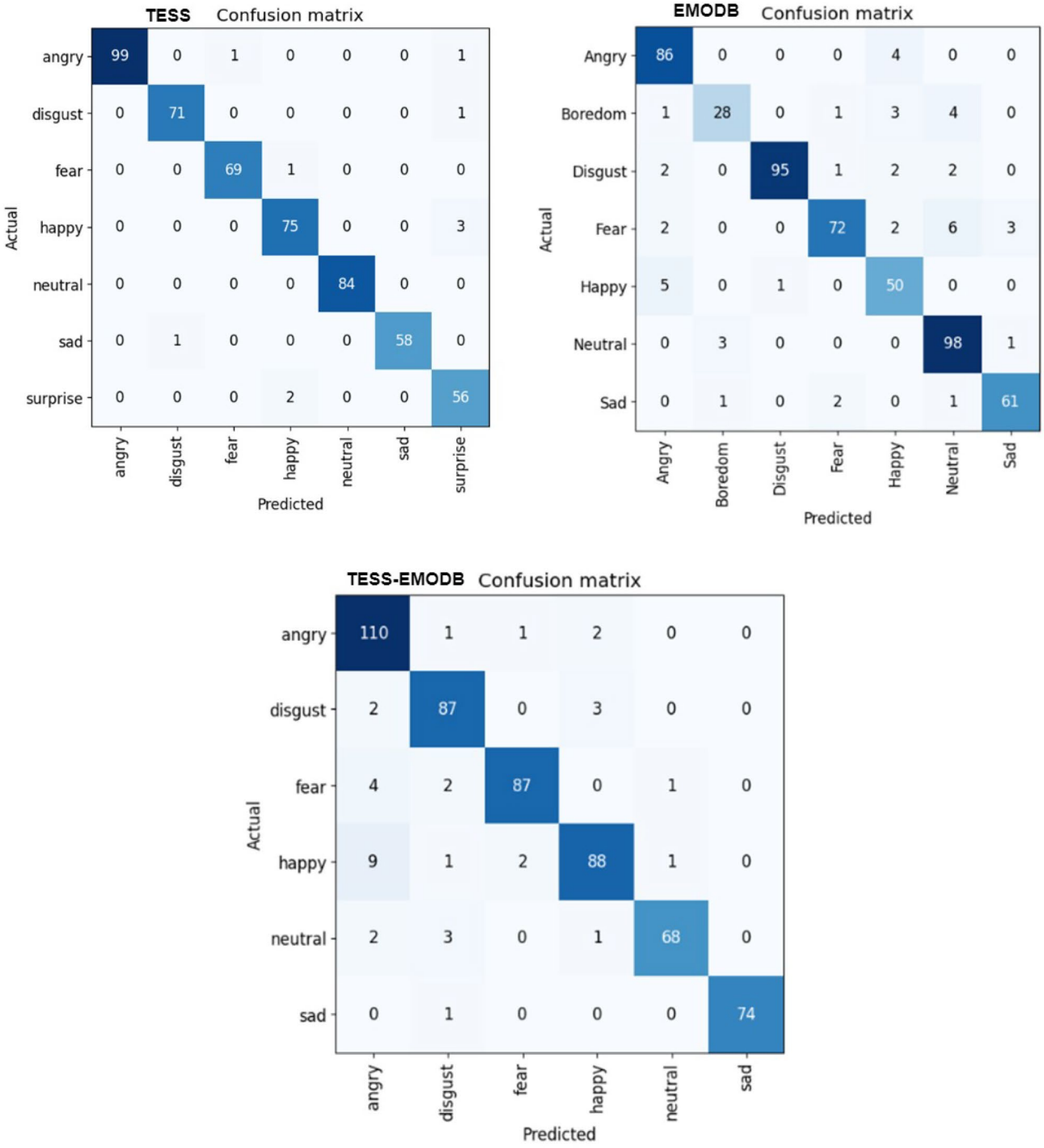


Figure 8. Confusion matrix for TESS, EMODB and TESS-EMODB.

recognition from our confusion matrix, but the classification report recorded a vivid minimum recognition of 76.0% recall and 88.0% precision. The hybrid dataset of TESS-EMODB recorded the lowest accuracy 74% on sad emotion and a 100% overall for angry emotion for six emotions, which further established the robustness of our proposed model for SER.

The simplicity of the model architectural design has in doubt contributed to its performance in enhancing the SER recognition rate, thereby, reducing misclassification of emotion and making it suitable for real-time applications in monitoring human behavioural patterns. The novelty of the model inappropriately recognizing emotion from speech utterances(mel-spectrogram) is also confirmed with selected emotion as shown in Fig. 9. Only three emotions out of about thirty selected for the test had wrong predictions, but twenty-seven of the rest were rightly predicted as the actual emotion. The first label represents the actual emotion, while the second label directly under it is the predicted.

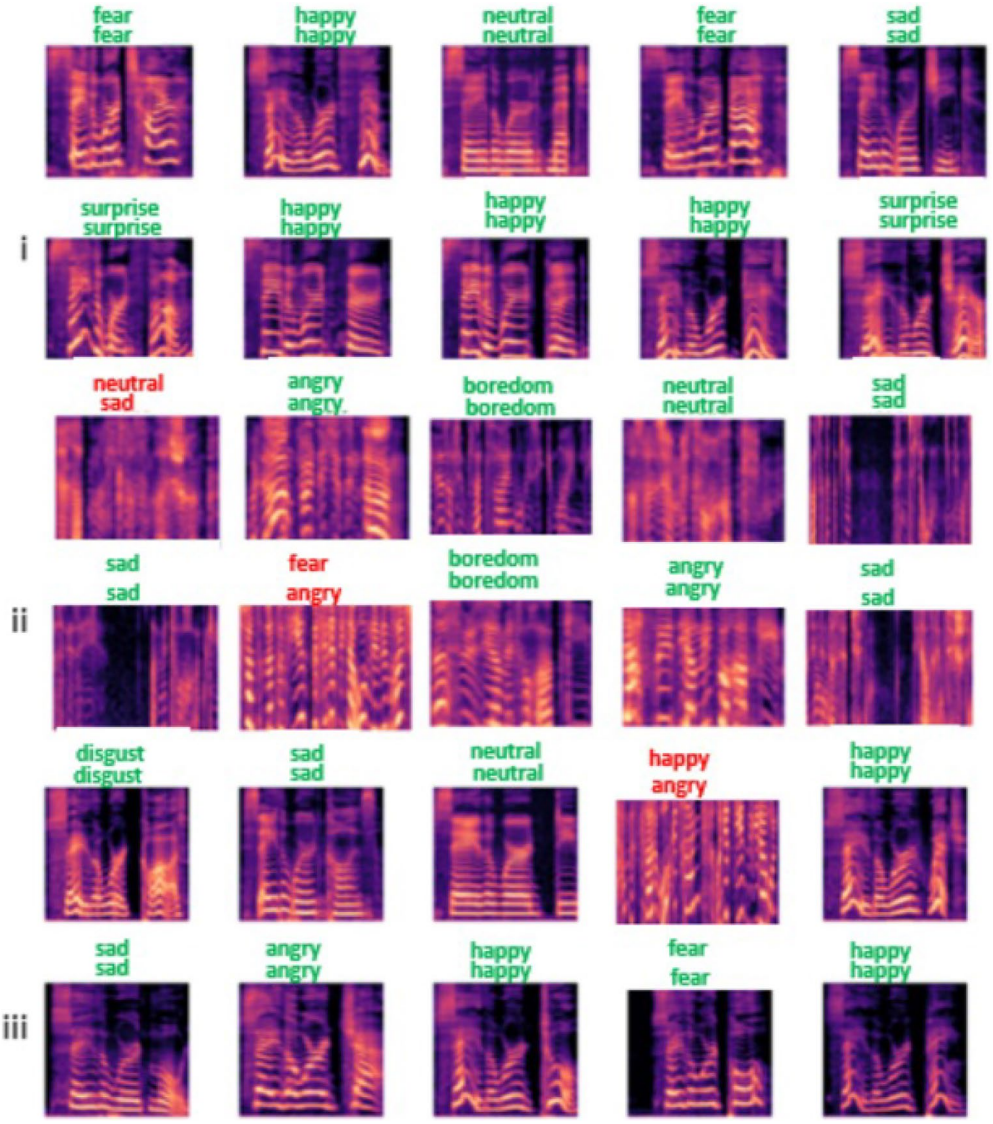


Figure 9. Test sample of emotion recognition output of the proposed model on three datasets: (i) represents recognition output on TESS dataset (ii) represents recognition output on EMODB dataset (iii) represent recognition output on TESS-EMODB dataset.

Performance evaluation

The comparative experiment aimed to evaluate the exact role that the Vision Transformer (ViT) model contributed to enhancing the speech emotion recognition ability that we observed. To carry out this extensive experiment, we substituted other deep learning-based architectures for the ViT model in our proposed framework, as shown in Table 6.

Though, while processing visual data in a similar way to the ViT model, they did not possess the distinctive architectural features of the ViT in capturing long-range dependencies efficiently. Two speech datasets used in this work are represented by the SDT1 and SDT2. The comparative study's results, which showed that the ViT model could enhance speech emotion recognition with fewer parameters while still achieving higher accuracy than other architectures, provided significant fresh insight. The apparent decrease in accuracy when utilizing other architectures highlights the significance of the self-attention mechanism of the ViT model in detecting nuanced spatial relationships that are essential for comprehending emotional nuances in human speech.

The comparative analysis of our proposed model's superior performance with other existing methods was carried out as illustrated in Table 10, using the selected speech emotion database, to demonstrate further our SER method's generalizability and suitability for real-time applications. The proposed method demonstrates the recent success of deep learning transformer in the SER domain, which recognized all the emotions with high accuracy, including even the neutral emotion, using an unambiguous architecture. In the table, we reveal the surpassing results of the proposed system, which are substantially greater than other methods, indicating the efficiency of our method. We carried out ablation experiments as indicated in Tables 7, 8, 9, with a focus on various patch sizes of the spectrogram image and removal of the embedded dropout layer component of the proposed

Architectures	Number of Parameters	Dataset	Accuracy
ResNet	9,116,032	SDT1	87.12
		SDT2	83.90
MobileNet	4,806,855	SDT1	48.50
		SDT2	81.32
InceptionNet	9,604,544	SDT1	62.30
		SDT2	75.60
DenseNet	7,043,654	SDT1	86.13
		SDT2	79.24
ViTSER	**4,166,151**	**SDT1**	**91.03**
		SDT2	**98.00**

Table 6. Deep learning architectures comparative experiments on EMODB and TESS Datasets. Bold highlights our proposed model and its results.

Dataset	A	D	F	H	N	S	Sr	OVA(%)
TESS	0.86	0.92	0.95	0.87	0.94	0.98	0.97	92
	0.90	0.93	0.89	0.86	0.95	0.98	0.95	
	0.88	0.92	0.92	0.86	0.94	0.98	0.96	
EMODB	A	B	D	F	H	N	S	
	0.88	0.86	0.85	0.91	0.92	0.84	0.96	89
	0.92	0.81	0.86	0.94	0.87	0.94	0.84	
	0.90	0.84	0.85	0.93	0.90	0.88	0.89	

Table 7. Ablation Experiment 2 on TESS and EMODB: Removal of dropout layer from the model architecture: *A* Angry, *H* Happy, *S* Sad, *D* Disgust, *N* Neutral, *F* Fear, *B* Boredom, *Sr* Surprise, *B* Boredom, *P* Precision, *R* Recall, *F1* F1-Score.

Size	Metrics	A	D	F	H	N	S	Sr	OVA(%)
14	P	0.83	0.90	0.93	0.90	0.92	0.98	0.97	91
	R	0.90	0.91	0.91	0.80	0.96	0.96	0.95	
	F1	0.86	0.91	0.92	0.85	0.94	0.97	0.96	
16	P	0.84	0.92	0.99	0.90	0.90	0.98	0.97	92
	R	0.94	0.96	0.88	0.82	0.97	0.91	0.98	
	F1	0.89	0.94	0.93	0.86	0.93	0.94	0.98	
28	P	0.86	0.96	0.98	0.90	0.95	0.98	0.98	94
	R	0.97	0.93	0.91	0.87	0.95	0.99	0.95	
	F1	0.91	0.94	0.95	0.88	0.95	0.98	0.97	
32	P	1.00	0.99	0.99	0.96	1.00	1.00	0.92	**98**
	R	0.98	0.99	0.99	0.96	1.00	0.98	0.97	
	F1	0.99	0.99	0.99	0.96	1.00	0.99	0.94	

Table 8. Ablation Experiment 1 on various patch sizes of audio spectrogram representation with TESS dataset: *A* Angry, *H* Happy, *S* Sad, *D* Disgust, *N* Neutral, *F* Fear, *B* Boredom, *Sr* Surprise, *P* Precision, *R* Recall, *F1* F1-Score. Bold highlights our proposed model and its results.

model. The first experiment result obtained from Table 7 shows that the removal of the embedded dropout layer as a functional component of the model significantly reduces the speech emotion recognition accuracy. The accuracy dropped by 6%, and 2.03% on TESS and EMODB datasets respectively. Likewise, the second ablation experiment's results from the two datasets with varying patch sizes indicated that the model declined in overall accuracy(OVA) as the patch size decreased. However, 14 and 32 represent the minimum and maximum patch sizes utilized in the experiments(Tables 8 and 9). It was obvious during the experiment that patch sizes above 32 increase the computational complexity, therefore we stopped at 32 which yielded an optimum accuracy without any need for parameter tuning (Table 10).

Size	Metrics	A	D	F	H	N	S	Sr	OVA(%)
14	P	0.87	0.83	0.80	0.86	0.86	0.86	0.96	86
	R	0.94	0.74	0.86	0.93	0.81	0.87	0.81	
	F1	0.90	0.78	0.82	0.90	0.84	0.87	0.88	
16	P	0.85	0.78	0.90	0.93	0.80	0.74	0.91	92
	R	0.96	0.67	0.90	0.92	0.74	0.90	0.76	
	F1	0.90	0.72	0.90	0.93	0.77	0.81	0.83	
28	P	0.87	0.89	0.81	0.86	0.86	0.86	0.88	86
	R	0.93	0.63	0.88	0.90	0.86	0.88	0.84	
	F1	0.90	0.74	0.84	0.88	0.86	0.87	0.86	
32	P	0.90	0.88	0.99	0.95	0.82	0.88	0.94	**91**
	R	0.96	0.76	0.93	0.85	0.89	0.96	0.94	
	F1	0.92	0.81	0.96	0.99	0.85	0.92	0.94	

Table 9. Patch size ablation experiment on EMODB dataset: B- Boredom. Bold highlights our proposed model and its results.

Year	Author & References	Method	Dataset	Accuracy (%)
2018	Chen et al.[73]	CNN+Attention	EMODB	82.82
2019	Jiang et al.[74]	CRNN	EMODB	84.49
2019	Meng et al.[75]	BiLSTM	EMODB	88.99
2020	Mustaqeem et al.[76]	CNN	EMODB	85.57
2020	Kwon,[77]	CNN	EMODB	90.01
2022	Guizzo et al.[78]	Quantarion CNN	EMODB	88.47
2022	Wen et al.[79]	Transfer Learning	EMODB	84.14
2023	**Proposed**	**ViTSER**	**EMODB**	**91.03**
2017	Verma, et al.[80]	SVM	TESS	96.00
2018	Praseetha et al.[81]	DNN	TESS	89.96.
2019	Gao[82]	CNN	TESS	81.00
2021	Krishnan et al.[83]	Decomposition	TESS	93.30.
2021	Chimthankar[84]	DNN	TESS	96.00.
2022	Akinpelu & Viriri[85]	VGGNet+RF	TESS	96.10.
2022	Guizzo et al.[78]	Quantarion CNN	TESS	97.00
2022	Choudhary et al.[86]	DNN	TESS	87.10
2023	**Proposed**	**ViTSER**	**TESS**	**98.00**

Table 10. Comparison with other baseline studies using TESS and EMODB dataset. Bold highlights our proposed model and its results.

Conclusion

In this research, a novel Vision Transformer model based on the mel-spectrogram and deep features was developed for the problem of speech emotion recognition. To assure accuracy, a simple MLP head attention with 128 dimensions was utilized to extract the deep features. With flattening, tokenizer, 32 patch size, position embedding, self-attention, and MLP head layers for enhancing SER, we developed a vision transformer model. The computational complexity was minimized due to the compactness of our model architecture, which is responsible for reducing an excessive number of parameters. To demonstrate the efficacy along with the significance and generalization of the model, its performance was assessed using two benchmark datasets: TESS and EMO-DB as opposed to[25]. The proposed system outperformed the state-of-the-art in terms of prediction results. Extensive experiments using our model produced astounding recognition accuracy scores of 98% for the TESS dataset, 91% for the EMO-DB, and 93% when the two datasets were combined. In order to recognize all emotions with better accuracy and a smaller model size to produce computationally friendly output, the proposed model improved by 2% and 5% over the state-of-the-art accuracy. The results of the proposed approach demonstrated the capability of Vision Transformer to capture global contextual information, making it possible to model long-range dependencies and enhance the representation of emotional speech patterns, ultimately leading to improved speech emotion recognition. We will concentrate on implementing this kind of system in additional speech recognition-related task systems in the future and go into more detail. Similar to this, we will conduct some tests to evaluate the effectiveness of the proposed method and the obtained results on other datasets, including non-synthetic speech corpora. When combined with other deep learning techniques, the recognition rates are likely to rise. Utilizing

additional speech features such as the Mel-Frequency Cepstral Coefficient (MFCC), Chromagram, and Tonnetz can enhance the investigation as they form part of our future work as well.

Data availability

The two publicly available datasets used or analysed for this study are available at: (i) the Tspace repository (https://tspace.library.utoronto.ca/handle/1807/24487) for the TESS dataset and (ii) Berlin Database of Emotional Speech repository (http://emodb.bilderbar.info/showresults/index.php) for EMODB dataset.

Received: 9 September 2023; Accepted: 2 June 2024
Published online: 07 June 2024

References

1. Alsabhan, W. Human-computer interaction with a real-time speech emotion recognition with ensembling techniques 1d. *Sensors (Switzerland)* **23**(1386), 1–21. https://doi.org/10.3390/s2303138 (2023).
2. Yahia, A. C., Moussaoui, Frahta, N. & Moussaoui, A. Effective speech emotion recognition using deep learning approaches for Algerian Dialect. In *In Proc. Intl. Conf. of Women in Data Science at Taif University, WiDSTaif* 1–6 (2021). https://doi.org/10.1109/WIDSTAIF52235.2021.9430224
3. Blackwell, A. Human Computer Interaction-Lecture Notes Cambridge Computer Science Tripos, Part II. https://www.cl.cam.ac.uk/teaching/1011/HCI/HCI2010.pdf (2010)
4. Muthusamy, K. H., Polat, Yaacob, S. Improved emotion recognition using gaussian mixture model and extreme learning machine in speech and glottal signals. Math. Probl. Eng. (2015). https://doi.org/10.1155/2015/394083
5. Xie, J., Zhu, M. & Hu, K. Fusion-based speech emotion classification using two-stage feature selection. *Speech Commun.* **66**(6), 102955. https://doi.org/10.1016/j.specom.2023.102955 (2023).
6. Vryzas, N., Kotsakis, R., Liatsou, A., Dimoulas, C. & Kalliris, G. Speech emotion recognition for performance interaction. *AES J. Audio Eng. Soc.* **66**(6), 457–467. https://doi.org/10.17743/jaes.2018.0036 (2018).
7. Hemin, I., Chu Kiong, L. & Fady, A. Bidirectional parallel echo state network for speech emotion recognition. *Neural Comput. Appl.* **34**, 17581–17599. https://doi.org/10.1007/s00521-022-07410-2 (2022).
8. Vaaras, E., Ahlqvist-björkroth, S., Drossos, K. & Lehtonen, L. Development of a speech emotion recognizer for large-scale child-centered audio recordings from a hospital environment. *Speech Commun.* **148**(May), 9–22. https://doi.org/10.1016/j.specom.2023.02.001 (2022).
9. Dev Priya, G., Kushagra, M., Ngoc Duy, N., Natesan, S. & Chee Peng, L. Towards an efficient backbone for preserving features in speech emotion recognition: Deep-shallow convolution with recurrent neural network. *Neural Comput. Appl.* **35**, 2457–2469. https://doi.org/10.1007/s00521-022-07723-2 (2023).
10. Haider, F., Pollak, S., Albert, P. & Luz, S. Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods. *Comput. Speech Lang.* **65**, 101119. https://doi.org/10.1016/j.csl.2020.101119 (2021).
11. Oh, S., Lee, J. Y. & Kim, D. K. The design of cnn architectures for optimal six basic emotion classification using multiple physiological signals. *Sensors (Switzerland)* **20**(3), 1–17. https://doi.org/10.3390/s20030866 (2020).
12. Kwon, S. A cnn-assisted enhanced audio signal processing. *Sensors (Switzerland)*https://doi.org/10.3390/s20010183 *(2020).*
13. Dutta, S. & Ganapathy, S. Multimodal transformer with learnable frontend and self attention for emotion recognition. In *In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Singapore, 23-27 May* 6917–6921 (2022). https://doi.org/10.1109/ICEIC57457.2023.10049941
14. Chai, J., Zeng, H., Li, A. & Ngai, E. W. T. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Mach. Learn. Appl.* **6**(August), 100134. https://doi.org/10.1016/j.mlwa.2021.100134 (2021).
15. Atsavasirilert, K., Theeramunkong, T., Usanavasin, S., Rugchatjaroen, A., Boonkla, S., Karnjana, J., Keerativittayanun, S. & Okumura, M. A light-weight deep convolutional neural network for speech emotion recognition using mel-spectrograms. In *In 2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (2019)
16. Jain, M., Narayan, S., Balaji, K. P., Bharath, K., Bhowmick, A., Karthik, R. & Muthu, R. K. Speech emotion recognition using support vector machine. arXiv:2002.07590. (2013)
17. Al Dujaili, M. J., Ebrahimi-Moghadam, A. & Fatlawi, A. Speech emotion recognition based on svm and knn classifications fusion. *Int. J. Electr. Comput. Eng. (IJECE)* **11**, 1259–1264 (2021).
18. Mansour, S., Mahdi, B. & Davood, G. Modular neural-svm scheme for speech emotion recognition using anova feature selection method. *Neural Comput. Appl.* **23**, 215–227 (2013).
19. Cheng, X. & Duan, Q. Speech emotion recognition using Gaussian mixture model. In *In Proceedings of the 2012 International Conference on Computer Application and System Modeling (ICCASM)* 1222–1225 (2012)
20. Lanjewar, R. B., Mathurkar, S. & Patel, N. Implementation and comparison of speech emotion recognition system using gaussian mixture model (gmm) and k- nearest neighbor (k-nn) techniques. *Phys. Rev. E* **49**, 50–57 (2015).
21. Mao, X., Chen, L. & Fu, L. Multi-level speech emotion recognition based on HMM and ANN. In *In Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering* 225–229 (2009)
22. Mirsamadi, S., Barsoum, E. & Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In *In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2227–2231 (2017)
23. Atmaja, B. T. & Akagi, M. Speech emotion recognition based on speech segment using LSTM with attention model. In *In Proceedings of the 2019 IEEE International Conference on Signals and Systems* 40–44 (2019)
24. Xie, Y. *et al.* Speech emotion classification using attention-based lstm. *IEEE/ACM Trans. Audio Speech Lang. Process* **27**, 1675–1685. https://doi.org/10.1109/CCECE47787.2020.9255752 (2019).
25. Ayush Kumar, C., Das Maharana, A., Krishnan, S., Sri, S., Hanuma, S., Jyothish Lal, G. & Ravi, V. Speech emotion recognition using CNN-LSTM and vision transformer. In *In Book Innovations in Bio-Inspired Computing and Applications* (2023)
26. Diao, H., Hao, Y., Xu, S. & Li, G. Implementation of lightweight convolutional neural networks via layer-wise differentiable compression. *Sensors*https://doi.org/10.3390/s21103464 *(2021).*
27. Manohar, K. & Logashanmugam, E. Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm. *Knowl. Based Syst.*https://doi.org/10.1016/j.knosys.2022.108659 *(2022).*
28. Fagbuagun, O., Folorunsho, O. & Adewole, L. Akin-Olayemi: Breast cancer diagnosis in women using neural networks and deep learning. *J. ICT Resour. Appl.* **16**(2), 152–166 (2022).
29. Qayyum, A. B. A., Arefeen, A. & Shahnaz, C. Convolutional neural network (CNN) based speech-emotion recognition. In *In Proceedings of the 2019 IEEE International Conference on Signal Processing, Information, Communication and Systems (SPICSCON)* 122–125 (2019)
30. Harár, P., Burget, R. & Dutta, M. K. Speech emotion recognition with deep learning. In *In Proceedings of the 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)* 137–140 (2017)

148

31. Fahad, S., Deepak, A., Pradhan, G. & Yadav, J. Dnn-hmm-based speaker-adaptive emotion recognition using mfcc and epoch-based features. *Circuits Syst. Signal Process* **40**, 466–489 (2022).
32. Singh, P. & Saha, G. Modulation spectral features for speech emotion recognition using deep neural networks. *Speech Commun.* **146**, 53–69. https://doi.org/10.1016/j.specom.2022.11.005 (2023).
33. G., W., H., L., J., H., D., L. & E., X. Random deep belief networks for recognizing emotions from speech signals. Comput. Intell. Neurosci. 1–9 (2017)
34. Poon-Feng, K., Huang, D. Y., Dong, M. & Li, H. Acoustic emotion recognition based on fusion of multiple feature-dependent deep boltzmann machines. In *In Proceedings of the 9th International Symposium on Chinese Spoken Language Processing* 584–588 (2014)
35. Zeng, Y., Mao, H., Peng, D. & Yi, Z. Spectrogram based multi-task audio classification. *Multimed. Tools Appl.* **78**, 3705–3722 (2017).
36. Popova, A. S., Rassadin, A. G. & Ponomarenko, A. A. Emotion recognition in sound. In *In Proceedings of the International Conference on Neuroinformatics, Moscow, Russia, 2-6 October* 117–124 (Springer, 2017)
37. Issa, D., Fatih Demirci, M. & Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control* **59**, 101894. https://doi.org/10.1016/j.bspc.2020.101894 (2020).
38. Li, H., Ding, W., Wu, Z. & Liu, Z. Learning fine-grained cross-modality excitement for speech emotion recognition. arXiv:2010.12733 (2010)
39. Zhao, J., Mao, X. & Chen, L. Speech emotion recognition using deep 1d and 2d cnn lstm networks. *Biomed. Signal Process. Control* **47**, 312–323. https://doi.org/10.1016/j.bspc.2018.08.035 (2019).
40. Zeng, M. & Xiao, N. Effective combination of densenet and bilstm for keyword spotting. *IEEE Access* **7**, 10767–10775 (2019).
41. Puri, T., Soni, M., Dhiman, G., Khalaf, O. I. & Khan, I. R. Detection of emotion of speech for ravdess audio using hybrid convolution neural network. *Hindawi J. Healthc. Eng. ii*https://doi.org/10.1155/2022/8472947 *(2022).*
42. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F. & Marchi, E. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autismn. In *In Proceedings of the INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France* (2013)
43. Zhu, L., Chen, L., Zhao, D., Zhou, J. & Zhang, W. Emotion recognition from Chinese speech for smart affective services using a combination of svm and dbn. *Sensors* **17**, 1694. https://doi.org/10.3390/s17071694 (2017).
44. Pawar, M. D. & Kokate, R. D. Convolution neural network based automatic speech emotion recognition using mel-frequency cepstrum coefficients. *Multimed. Tools Appl.* **80**, 15563–15587 (2021).
45. Bhangale, K. & Kothandaraman, M. Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. *Electronics (Switzerland)*https://doi.org/10.3390/electronics12040839 *(2023).*
46. Badshah, A. M. *et al.* Deep features-based speech emotion recognition for smart affective services. *Multimed. Tools Appl.* **78**, 5571–5589. https://doi.org/10.1007/s11042-017-5292-7 (2019).
47. Latif, S., Zaidi, A., Cuayahuitl, H., Shamshad, F., Shoukat, M. & Qadir, J. Transformers in speech processing: A survey. http://arxiv.org/abs/2303.11607 16, 1–27 (2023)
48. Chen, S. *et al.* Wavlm: Large-scale self-supervised pre- training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.* **16**, 1505–1518 (2022).
49. Xu, M., Li, S., X., Z.: Transformer-based end-to-end speech recognition with local dense synthesizer attention. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 5899–5903 (IEEE, 2021)
50. Shor, J., Jansen, A., Han, W., Park, D. & Zhang, Y. Universal paralinguistic speech representations using self-supervised conformers. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 3169–3173 (IEEE, 2022)
51. Chen, W., Xing, X., Xu, X., Pang, J. & Du, L. Speechformer: A hierarchical efficient framework incorporating the characteristics of speech. arXiv preprint arXiv:2203.03812 (2022)
52. Gao, Z., Zhang, S., McLoughlin, I. & Yan, Z. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. arXiv preprint arXiv:2206.08317 (2022)
53. Kumawat, P. & Routray, A. Applying TDNN architectures for analyzing duration dependencies on speech emotion recognition. In *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 561–565 (2021). https://doi.org/10.21437/Interspeech.2021-2168
54. Han, S., Leng, F. & Jin, Z. Speech emotion recognition with a ResNet-CNN-transformer parallel neural network. In *In Proceedings of the International Conference on Communications, Information System and Computer Engineering(CISCE)* 803–807 (2021)
55. John, V. & Kawanishi, Y. Audio and video-based emotion recognition using multimodal transformers. In *In Proceedings of International Conference on Pattern Recognition* 2582–2588 (2022)
56. Slimi, A., Nicolas, H. & Zrigui, M. Hybrid time distributed CNN-transformer for speech emotion recognition. In *In Proceedings of the 17th International Conference on Software Technologies ICSOFT* (2022)
57. Chaudhari, A., Bhatt, C., Krishna, A. & Mazzeo, P. L. Vitfer: Facial emotion recognition with vision transformers. *Appl. Syst. Innov.*https://doi.org/10.3390/asi5040080 *(2022).*
58. Arezzo, A. & Berretti, S. SPEAKER VGG CCT: Cross-corpus speech emotion recognition with speaker embedding and vision transformersn. In *In Proceedings of the 4th ACM International Conference on Multimedia in Asia, MMAsia* (2022)
59. Latif, S., Zaidi, A., Cuayahuitl, H., Shamshad, F., Shoukat, M. & Qadir, J. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. arxiv.org/abs/2303.11607 (2023)
60. Alluhaidan, A. S., Saidani, O., Jahangir, R., Nauman, M. A. & Neffati, O. S. Speech emotion recognition through hybrid features and convolutional neural network. *Appl. Sci. (Switzerland)* 13(8) (2023)
61. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* 55 (2012)
62. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *In Proceedings of ICLR 2021 AN* (2021)
63. Dong, L., Xu, S. & Xu, B. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* **2236**(1), 5884–5888. https://doi.org/10.1109/ICASSP.2018.8462506 (2018).
64. Hendrycks, D. & Gimpel, K. Gaussian error linear units (gelus). ArXiv:1606.08415v5 [Cs.LG], 1–10 (2023)
65. Pichora-Fuller, M. K. & Dupuis, K. Toronto emotional speech set (tess). https://doi.org/10.5683/SP2/E8H2MF. (2020)
66. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F. & Weiss, B. A database of german emotional speech (emodb). INTERSPEECH, 1517–1520 (2005)
67. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. Automatic Differentiation in Pytorch. In *In Proceedings of Advances in NIPS* (2017)
68. Xu, Y., Zhang, J. & Miao, D. Three-way confusion matrix for classification. A measure driven view. *Inf. Sci.* **507**, 772–794 (2020).
69. Deng, X., Liu, Q., Deng, Y. & Mahadevan, S. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Inf. Sci.* **340**, 250–261 (2016).
70. Snmez, Y., & Varol, A. In-depth analysis of speech production, auditory system, emotion theories and emotion recognition. In *In Proceedings of the 2020 8th International Symposium on Digital Forensics and Security (ISDFS)* (2020)
71. Shu, L. *et al.* A review of emotion recognition using physiological signals. *Sensors* **18**, 2074. https://doi.org/10.1007/978-3-319-58996-1_13 (2018).

72. Ekman, P. & Davidson, R. J. The Nature of Emotion: Fundamental Questions (Oxford University Press, 1994)
73. Chen, M., He, X., Yang, J., H., Z.: 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Process. Lett.* 25(10), 1440–1444 (2018)
74. Jiang, P., Fu, H., Tao, H., Lei, P. & Zhao, L. Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. *IEEE Access* **7**, 90368–90377. https://doi.org/10.1109/ACCESS.2019.2927384 (2019).
75. Meng, H., Yan, T., Yuan, F. & Wei, H. Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE Access* **7**, 125868–12588 (2019).
76. Mustaqeem, M., Sajjad, M., & K, S. Clustering based speech emotion recognition by incorporating learned features and deep bilstm. *IEEE Access* (2020). https://doi.org/10.1109/ACCESS.2020.2990405
77. Mustaqeem, Kwon, S. Mlt-dnet: Speech emotion recognition using 1d dilated cnn based on multi-learning trick approach. Expert Syst. Appl. 114177 (2021). https://doi.org/10.1016/j.eswa.2020.114177
78. Guizzo, E., Weyde, T., Scardapane, S. & Comminiello, D. Learning speech emotion representations in the quaternion domain. *IEEE/ACM Trans. Audio Speech Lang. Process.* **31**, 1200–1212 (2022).
79. Wen, G. *et al.* Self-labeling with feature transfer for speech emotion recognition. *Knowl. Based Syst.* **254**, 109589 (2022).
80. Verma, D. & Mukhopadhyay, D. Age driven automatic speech emotion recognition system. In *In Proceeding of IEEE International Conference on Computing, Communication and Automation* (2017)
81. Praseetha, V. & Vadivel, S. Deep learning models for speech emotion recognition. *J. Comput. Sci.* 14(11) (2018)
82. Gao, Y. Speech-Based Emotion Recognition. https://libraetd.lib.virginia.edu/downloads/2f75r8498?filename=1GaoYe2019MS.pdf (2019)
83. Krishnan, P. T., Joseph Raj, A. N. & Rajangam, V. Emotion classification from speech signal based on empirical mode decomposition and non-linear features. *Complex Intell. Syst.* **7**(4), 1919–1934. https://doi.org/10.1007/s40747-021-00295-z (2021).
84. Chimthankar, P. P. Speech Emotion Recognition using Deep Learning. http://norma.ncirl.ie/5142/1/priychimtankar.pdf (2021)
85. Akinpelu, S. & Viriri, S. Robust feature selection-based speech emotion classification using deep transfer learning. *Appl. Sci.* **12**, 8265. https://doi.org/10.3390/app12168265 (2022).
86. Choudhary, R. R., Meena, G. & Mohbey, K. K. Speech emotion based sentiment recognition using deep neural networks. *J. Phys. Conf. Ser.* **2236**(1), 012003 (2022).

Author contributions

Conceptualization, S.A. and V.S.; Methodology, A.A. and S.A.; Software, S.A.; Validation, S.V. and A. A; Formal analysis, S.V.; Investigation, S.A.; Resources, S.V.; Data curation, S.A.; Writing original draft preparation, S.A. and A.A; review and editing, S.V.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5.3 An Improved Bilingual Speech Emotion Recognition Using Shift Window Transformer

5.3.1 Brief Review

The Paper entitled SwinTSER: An Improved Bilingual Speech Emotion Recognition Using Shift Window Transformer was proposed to overcome the bi-lingual speech emotional utterance challenge. A deep learning technique that extracts salient emotional features from speech utterances of varied languages was developed in this paper. Speaking utterances are associated with other languages in addition to the cultural background. Therefore, the paper devised deep learning methods that allow models to generalize in a variety of language and cultural contexts, without necessarily concentrating on local features only. The paper utilized a self-attention mechanism that is computed within local windows in a non-overlapping fashion to reduce computational complexity that may result from Bi-Lingual speech utterance for speech emotion classification. An improved bi-lingual speech emotion classification that is based on a shift window transformer network is integrated into deep learning in Chapter 6.

Paper status: Under Review

SwinTSER: An Improved Bilingual Speech Emotion Recognition Using Shift Window Transformer

Samson Akinpelu[1], Serestina Viriri[1*] and Adekanmi Adegun[1]

[1]University of KwaZulu-Natal, School of Mathematics, Statistics and Computer Science, Durban, 4000, South Africa.

*Corresponding author(s). E-mail(s): viriris@ukzn.ac.za;
Contributing authors: 222068579@stu.ukzn.ac.za;
adeguna@ukzn.ac.za;

Abstract

Emotion Recognition from human speech occupies a significant position in Human-Computer Interaction, especially with the recent advancements in Artificial Intelligence and Robotic computing. As the level of interactivity of man-machine increases, intuitive responses that are emotionally based have attracted a lot of research into emotion recognition from speech signals. However, with various machine learning models that littered literature, cross-language efficient speech emotion recognition with extracted features inherent in speech signals with state-of-the-art deep learning techniques is still posing a serious challenge. In this paper, we proposed a deep learning transformer network that is based on a shift window for speech emotion recognition using speech corpus from two different languages. Shift Window Transformer (SWT) is based on hierarchical transformer architecture designed for natural language tasks and has recently become a novel model in computer vision and image processing tasks. The input feature to the model Mel-spectrogram is extracted from two public speech datasets Toronto English Emotion Speech (TEES) and EMOVO. Our proposed transformer model achieved a promising result of 98.3%, 64% and 66% accuracy of recognition on TESS, EMOVO and TESS_EMOVO (hybrid bi-lingual) datasets respectively, after extensive experiment and parameter optimization. Our performance evaluation revealed that the proposed model yielded an improved result in the recognition of six different emotions from human auditory speech compared to others found in the literature.

Keywords: Deep Learning, Transformer, Convolution Neural Network, Emotion, Speech Signals.

1 Introduction

In human conversation, people can make a large range of noises with their mouths that correspond to different linguistic words. While it is evident that humans understand these phrases when they are speaking to one another, speech signal processing is necessary for human-machine interaction. Human auditory voice signals provide useful information about the language used, the main points of the communication, the speaker's articulation, personality trait, a focal point, gender, and other unique features [1].

Expressing feelings and communicating with people can be achieved most effectively through speech. When considering other sources such as text, images, expressions on the face, and EEG (electroencephalogram) signals, Speech is the easiest to obtain emotion recognition [2]. Speech Emotion Recognition (SER) is the process of empowering machines to recognize and categorize speech signals to detect emotional information from speakers' utterances [3]. Understanding and recognizing human emotion can prevent unnecessary aggressions, suicide, depression, and rifts among people. An individual's breathing pattern is influenced by their emotional state, and both their breathing pattern and muscular tension are necessary for that person to produce speech [4]. As a result, the paralinguistic features of speech generated under emotional circumstances differ from those of speech produced under regular circumstances, which has been beckoning for the attention of experts in the SER domain to develop a language-independent and efficient model for recognizing human emotion.

The field of speech emotion recognition has advanced significantly over time, beginning with the advent of digital audio transmission, and continuing with the creation of attention and transformer network techniques that can capture long-term dependencies feature in speech signal[5]. When digital audio transmission first emerged, researchers concentrated on creating mechanisms for adaptable playout latency mechanisms for audio packets sent over the internet. However, with the development of technology came the realization that it was critical to comprehend and identify the emotions expressed through speech, as speech contains important emotional cues that can improve systems for communicating and human-computer interaction [Han, 2022. 83]. As a result, the discipline of speech emotion recognition started to attract significant interest.

Speech emotion recognition (SER) has a wide range of practical uses. It enhances the interaction between people and machines [6]. In addition, it provides online customer assistance to enhance the well-being of those with mental disorders (psycho-medicine). E-learning, telemarketing, call centre customer

service [7][8], criminal investigation, surveillance, avoidable traffic incidents [9], medical applications, etc., are all enhanced by SER. For more than thirty years, SER has piqued the curiosity of several researchers due to its extensive practical applications.

The traditional SER system detects the desired emotion, irrespective of the language, from speech signals through preprocessing of the signals, feature extraction, and classification (using some classifiers) to six fundamental emotional information such as happy, sad, fear, surprise, neutral, and angry [10] as shown in Figure 1. However, as the desire to grow SER domain became more paramount, this conventional approach is bewildered with problems of misclassification of emotion and low accuracy. It was against this backdrop that prompted researchers to explore Machine learning models, Neural network models such as Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Long Short Term Memory (LSTM), etc, to improve speech emotion recognition tasks. Surprisingly, these approaches improved the recognition of emotion from human speech, but not without some complexity issues, which the transformer architecture proffers solutions to [11]. Transformer [12], a scalable natural language processing (NLP) model, has been applied to image classification tasks with outstanding performance. Figure 1 shows a classical structure of emotion recognition. Owing to the exceptional

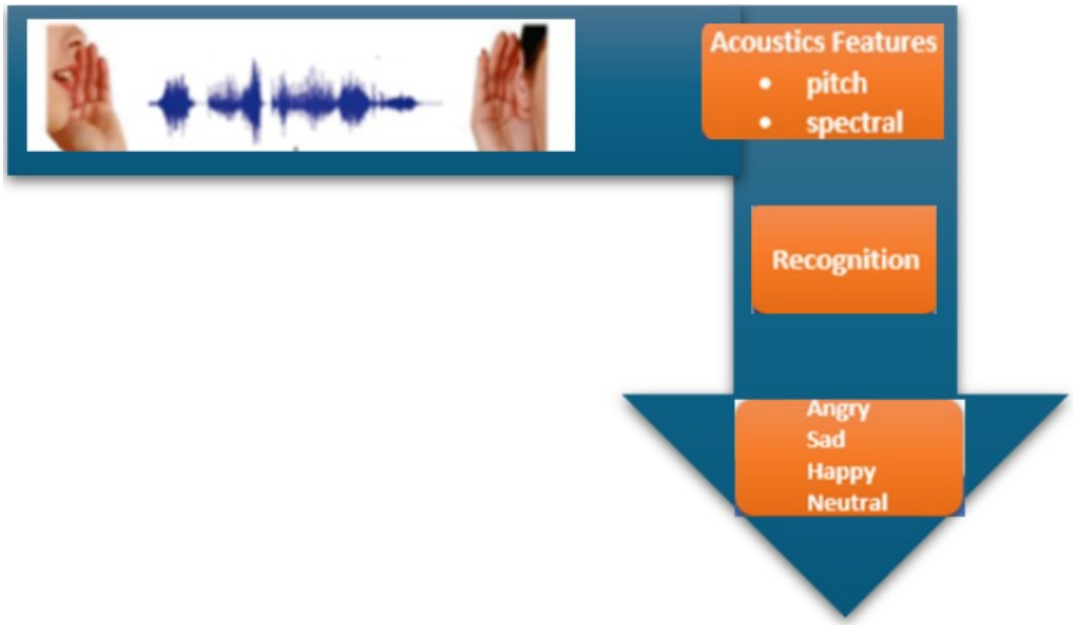


Fig. 1: Conventional Speech Emotion Classification System

performance in a variety of applications, including speech emotion recognition [13], machine translation[14], automatic speech recognition (ASR) [15], speech enhancement, and speech separation [16], to mention a few, transformers have attracted a great deal of research interest in the SER and natural language

processing (NLP) communities [17][18][19][20]. This model has even outperformed conventional recurrent neural networks (RNNs), which have difficulty with extended sequences and the sequence-to-sequence problem caused by the vanishing gradient problem [21].

Transformer models have gained popularity and developed quickly, and this has led to a plethora of research exploring the distinctive features that enable them to perform better than other models, especially the SWT variant of transformers, which stands tall among its peers for its ability to reduce complexity through the computation hierarchical feature maps construct[12]. Swin Transformer creates a hierarchical representation by beginning with small patches and progressively merging nearby patches in deeper Transformer layers. The purpose of the SWT is to improve the robustness and accuracy of emotion recognition across languages by capturing the intricate linguistic nuances and temporal relationships found in bilingual speech. We leverage the potency of this superior transformer framework to improve the recognition of emotion from speech signals using two languages (English and Italian) of speech corpora. The specific contributions of this work are as follows: The main contributions of this study are:

1. The proposed SWT Transformer model extracted Mel spectrograms from raw speech signals. This helped to efficiently extract and represent features, which allowed for capturing salient acoustic features necessary for speech emotion recognition, laying a solid foundation for the modeling process afterwards.
2. Computational complexity is reduced for SER through self-attention that is computed within local windows in a non-overlapping fashion. Fewer parameters, lesser model size and efficient layers that learn discriminative features from the input speech signal features increase the robustness of the model.
3. On bilingual datasets, the proposed model achieved an enhanced recognition accuracy that surpassed current techniques. However, this also brought our attention to the challenges involved with adjusting SER models to a wide range of linguistic and emotional expressions, offering suggestions for further development and underscoring the necessity for models that can accommodate the variations found in real-world applications

The paper is organized as: the review of relevant works on SER methods is presented in Section 2. The methodology used for the tests, feature extraction, transformer model used for the SER, and evaluation metrics are all covered in Section 3. The experimental results and a discussion of the findings that were observed are presented in Section 4. Section 5 provides the conclusion and recommendations for further research.

2 Review of Related Works

The utilization of various conventional machine learning (ML) classifiers trained with speech feature vectors kicked off the journey of the speech emotion recognition (SER) system. These ML classifiers include Support Vector Machine (SVM) [22], Hidden Markov Model (HMM) [23], K-Nearest Neighbours (K-NN), Gaussian Mixture Model (GMM) [24] and Artificial Neural Networks (ANN) [25]. Consequently, various combinations of these primitive classifiers have been presented with the utmost intention of enhancing the classification of emotion in the SER process [26].

Nevertheless, even with the early results reported, standard ML-based approaches have particular constraints when applied to SER [27]. Initially, no specific machine learning technique exists that can effectively extract precise and discriminating features from bilingual speech databases. Secondly, the accuracy of machine learning classifiers that are efficiently modelled suffers when large speech datasets and manually generated features are used for SER Additionally, for varying language content and acoustic conditions, the classification performance substantially declines due to the generic traits of ML-based SER classifiers.

Deep Learning (DL) techniques [28] reduced the manual efforts for the extraction of distinctive features from speech signals, in contrast to the prevalent issues with ML techniques. Artificial neural networks with multiple layers that hierarchically learn features make up deep learning. Deep Neural Networks (DNN) [29], restricted Boltzmann machines (RBM) [30], convolutional neural networks (CNN), long-short-term memory networks (LSTM) [31], recurrent neural networks (RNN) [32] and attention networks [33] are the most widely used speech emotion recognition classifiers based on DL. A spectrogram feature that was extracted from the Rayson Audio-Visual Database of Emotional Speech and Song (RAVDESS) speech dataset was fed into a DNN with a gated residual network in Zeng et al. study [34], yielded a 65.97% accuracy rate for emotion recognition on the tested data. Similarly, Popova et al. [35] used a pre-trained VGG-16 convolutional neural network and, after extensive experiments, they were able to attain an accuracy of 71%.

A novel Deep Convolutional Neural Network (DNN) for SER was proposed by Issa et al. [36] in an attempt to boost the possibilities of enhancing the recognition rate. Several feature vectors, including the Mel-Spectrogram, spectral contrast, and frequency-cepstral coefficient, were obtained from speech utterances and concatenated to provide the model input. Using the RAVDESS dataset, their technique was able to identify eight distinct emotions with an accuracy of 71.61%. For the benefit of generalizability, their approach was validated as well on the EMODB and IEMOCAP datasets. Their CNN model, however, was unable to effectively represent the spatial features and sequences unique to voice signals. Li et al. [37] suggested a multimodal strategy using temporal alignment and deep learning techniques to address the problem. Utilizing CNN, LSTM, and Attention Mechanism, their approach achieved a promising result of 70.8% accuracy with semantic embeddings. For SER

tasks, the use of CNN, LSTM, or RNN in combination has shown notable improvement [38]. This method mainly depends on using CNN to extract features from raw speech signals, which are then passed to an LSTM or RNN to extract long-term dependency features unique to auditory utterance emotion identification[39]. On the RAVDESS dataset, Puri et al. [40] used a hybrid technique that combined LSTM and DNN. They sampled the raw speech signals and extracted the MFCC, then fed it into their model. Researchers in the SER field have also become interested in the ensemble technique, which extracts salient features from speech utterances and feeds the emotional features into a classifier regardless of the speakers' language and cultural background.

Kwon et al. [41], automated extraction of local and global emotional features from acoustic signals was achieved by using an endways multi-learning trick (MLT) based on 1D improved CNN model. Their method utilized a dynamic fusion framework in extracting the discriminative features to enhance the emotion recognition rate. With 73% and 90% accuracy rates, respectively, the proposed model assessed short and long-term relative relationships across two benchmark SER datasets, IEMOCAP and EMO-DB. In comparison, the approach requires more time than other models to train and test the real-time speech signals. An alternative deep learning technique for SER was presented by Pathar et al. in a separate study [42]. The CNN model was trained by extracting relevant features from speech signals with the use of MFCC. Their accuracy of 93.8% on the EMODB dataset is a noteworthy result. A novel lightweight DCNN method for speech emotion identification based on multi-acoustic attributes was developed by the author in [43]. Through their approach, they were able to achieve 93.31% on the Berlin Database of Emotional Speech (EMODB) and 94.18% on RAVDESS, respectively, by extracting and feeding in a variety of features, including Zero Crossing Rate (ZCR), wavelet packet transform (WPT), spectral roll-off, linear prediction cepstral coefficients (LPCC), etc. Using a spectrogram from an audio source, Badshah et al., [44] introduced a parallel CNN-based model for SER. They employed a Fast Fourier Transform (FFT)-generated spectrogram as input for a pooling mechanism with kernels of varying sizes.Their method demonstrates the significance of CNN's max-pooling mechanism. However, their model was evaluated on a monolingual dataset.

In recent times, there has been a surge in efforts to enhance the effectiveness of deep learning models for SER and address the issue of long-range dependencies specific to CNN-based models. In SER, the cutting-edge transformer model has been introduced [45]. Han et al.,'s proposal [46]included a parallel design that made use of ResNet and the Transformer model. An audio-video multimodal transformer was used for emotion recognition by Vijay et al.,[47]. To extract relevant features from spectrogram pictures, they used three self-attention and block embedding techniques. On the RAVDESS, CREMA-D, and SAVEE datasets, their model produced results of 93.59%, 72.45%, and 99.17%, respectively. However, due to the architecture, massive computer resources had to be utilized. The fact that CNN-based models often work with

fixed-size input windows limits their capacity to detect long-range dependencies in speech signals. Temporal dynamics are often seen in speech emotion beyond the immediate areas of the speech sequence. As a result, we proposed a Shift window Transformer (SWT) model that includes both local and self-attention mechanisms [12] that allows it to capture global contextual information. This allows us to model long-range dependencies and improve speech emotion recognition by enhancing the representation of emotional speech patterns, especially on bilingual speech emotional corpora. A Cross-Lingual SER was implemented by the author in Liu et al. [48] [2020] using Transformer-Based Acoustic and Phonetic Embeddings. Even though the study's SER significantly improved, one of its constraints is that it was only able to capture a limited amount of the nuanced temporal variations in emotional speech across languages due to its focus on acoustic and phonetic embeddings rather than temporal dependencies.

Shift Window Transformer was proposed by Huang et al. [49] for Visual Speech Enhancement and emotion recognition. In visual speech challenges, the SWT presents an efficient technique for temporal modeling. Its ability to recognize speech emotions and how well it handles intricate linguistic differences between languages, however, still needs more research. Global self-attention is typically performed by Transformers, computing the relationships between each token and every other token. In relation to the quantity of tokens, the global computation results in quadratic complexity. With SWT, self-attention is computed within local windows in a non-overlapping fashion[12]. While window-based self-attention results in linear complexity and is scalable easily, global self-attention causes quadratic computational complexity in the number of patches. As a result, a novel SWT transformer-based technique was presented in this study to enhance speech emotion recognition and shorten the model training duration while utilizing constrained computational resources. A comparison with baseline deep learning techniques on SER was carried out using a bilingual dataset, and the proposed transformer model showed improved recognition accuracy.

3 Methods and Techniques

This section presents an overview of the model that we propose (Figure 2) for speech utterance-based emotion recognition. To efficiently train the SWT transformer model with feature vectors that ultimately result in speech emotion recognition, we highlighted the key details from the auditory speech database, pre-processing, feature extraction (Mel Spectrogram), and feeding of the model. Ultimately, the proposed transformer model was implemented with the Keras and Tensorflow libraries' aid, making simulations easier to run. Two bilingual datasets were used for evaluation, where accuracy and robustness were the main criteria.

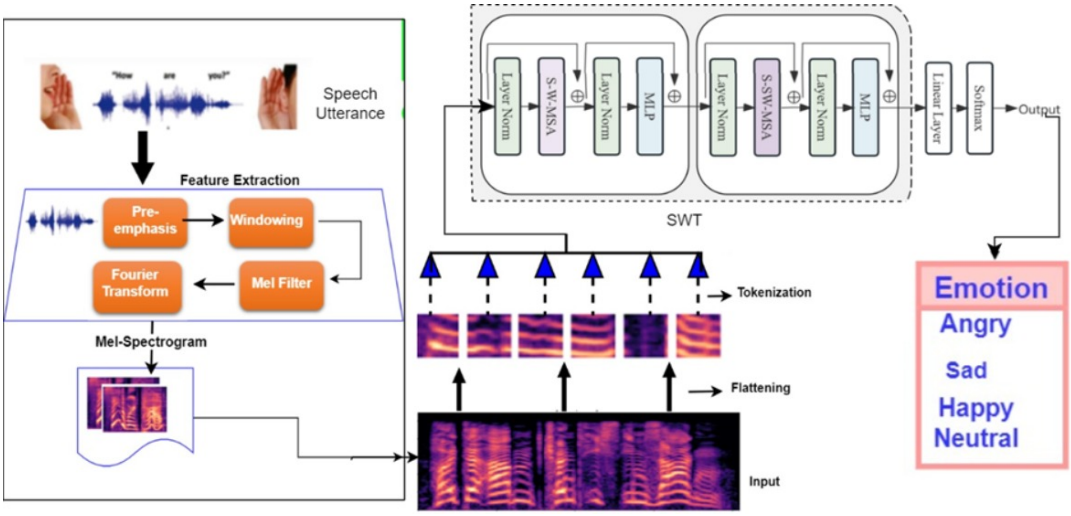


Fig. 2: Architectural view of the proposed SWT model for SER

3.1 Datasets

At the preliminary stage of this study, two public datasets were utilized to verify the performance of our model and establish its generalizability. The two datasets (TESS and EMOVO) are of different languages (Bi-lingual)-English and Italian based. It is common in the literature to see many SER implementations using single language; however, we chose a bi-lingual approach because the language has a role to play in human speech utterance. Moreover, our proposed model is exposed to two different language speech datasets to establish its robustness in capturing emotional cues from speech signals.

3.1.1 TESS Dataset

Many SER initiatives have made use of the Toronto English Speech Set, or TESS for short, one of the biggest publicly available datasets. In 2010, TESS speech samples were recorded by Northwestern University's Audio Laboratory [50]. A few of the 200 words were to be pronounced by the two actors during the unscheduled occurrence. A thorough collection of 2800 verbal utterances was obtained through the recording of their voices. In the scenario, seven distinct emotions were observed: pleased, shocked, terrified, furious, glad, sad, and neutral. The TESS description based on the contribution of each emotion to the entire speech dataset is illustrated in Figure 3.

3.1.2 EMOVO Dataset

There are seven different emotional states in the EMOVO dataset: disgust, fear, angry, joy, surprise, sadness, and neutral. Italian language utterances are included in the first dataset, called EMOVO [51]. Six actors are captured on tape uttering the words in fourteen sentences in this dataset. There are 24 annotators per group for each of the 588 utterances in this dataset, which has

Fig. 3: TESS Emotional speech percentage distribution

been further annotated. All voice recordings were made at the laboratory of the Fondazione Ugo Bordoni.

Fig. 4: EMOVO Emotional Speech percentage distribution

3.2 Speech Feature Extraction

The performance of a model in a recognition task is significantly influenced by the quality of the feature extraction used. Inadequate features may also lead to unsatisfactory recognition results. In the context of Deep Learning (DL), feature extraction is an essential phase in deep learning [52], since the diversity of the features the SER model employs significant impact on the model's success or failure. Recognition will be accurate if there is a substantial correlation between the derived features and the emotional classes; otherwise, it will be complicated and lead to misclassification. The overall quality of the feature set has a big impact on how well the SER model performs. In this study, we utilized a popular feature extraction approach that captures essential features for our transformer model to efficiently recognize emotion from speech signal.

The Mel spectrogram feature, which shows the frequency content of an audio signal as a function of time, was extracted from the speech signal. It is based on the conventional spectrogram, but the frequency scale has been modified to better correspond with human hearing. Auditory spectrograms are often utilized to examine different sounds and identify spoken words phonetically. The time-frequency information of the utterance is provided by it. However, it is more difficult to identify higher frequency differences than lower frequency variations because human perception of frequency does not follow a linear scale. Even if the distance between the two pairs is the same, we can precisely distinguish between 500 and $1000Hz$, but we struggle to distinguish between $10,000$ and $10,500Hz$.

The main processes in the Mel spectrogram feature extraction technique are the following: pre-emphasis, windowing (frame segmentation), Fast Fourier transform (FFT), Mel-filtering, and determining magnitude's log, scale. Equations 1 and 2 are used to obtain the feature extraction process mathematically.

$$w(n) = 0.54 - 0.46\cos\left\{\frac{2}{n-1}\right\} 0 \leq n \leq N-1 \tag{1}$$

$$Y(n) = X(n).W(n) \tag{2}$$

where N = number of samples in each frame Y(n) = Output signal X(n) = Input signal W(n) = Hamming window

Using the hamming window function, the audio signal is first split into short frames to maintain continuity and prevent overlapping. Secondly, each frame is subjected to FFT to convert the signal from the time domain to the frequency domain and reveal the frequency components contained in that frame. Following this, each frame's FFT magnitude spectrum is subjected to a bank of triangular filters that are arranged under the Mel scale to extract energy within distinct Mel-frequency bands. The speech signal's dynamic range is finally compressed by computing and standardizing the log magnitude of the filtered outputs. A 2D representation (fit for model) is created by stacking the resulting Mel spectrogram frames, with the y-axis denoting frequency

bins and the x-axis representing time. The structured, relevant, and informative representation of speech signals provided by Mel spectrograms as features for emotion recognition enables models to efficiently capture and distinguish emotional cues included in the speech utterance.

3.3 Swin Transformer (SWT) Layer

Originally, machine-level translation was intended to be accomplished with the transformer model. Nevertheless, due to its effective performance, it substitutes RNN in NLP and gains prominence [53][54]. To eliminate the recurrent nature of any kind of processing, the transformer model makes use of an alternative internal attention mechanism that is also referred to as the self-attention mechanism. For a given speech utterance, it generates relevant features using a linear transformation. The transformer model consists mainly of two components that are peculiar to its network, which are self-attention (SA) and multi-head self-attention(MSA). The goal of the self-attention (SA) layer is to aggregate global information from the complete input sequence to capture the internal correlation of the input sequence which is a challenging task for traditional recurrent models. The goal of the self-attention (SA) layer is to aggregate global information from the complete input sequence to capture the internal correlation of the sequence which is a challenging task for traditional recurrent models.

The input feature $X(X \in \mathbb{R}^{N \times D})$ comprising of N entities with$D$ dimension each, is usually transformed into a query, key and value (Q, K, V) through a matrix of learnable weight respectively, which can be obtained mathematically from $\mathbb{Q} = \mathbb{X}\mathbb{W}^{Q}, \mathbb{K} = \mathbb{X}\mathbb{W}^{K}, \mathbb{V} = \mathbb{X}\mathbb{W}^{V}$, where W denotes the weights.

Contrarily, Multiple Self-Attention (MSA) blocks comprise Multi-Head Self-Attention (MHSA), as opposed to a single attention computation that exists in single-layer attention network. To make it possible to simulate dependencies between various components in the input sequence, these SA blocks are collectively concatenated channel-wise. For each head in (MHSA), a matrix of learnable weight expressed by $\left\{\mathbb{W}^{Q}m, \mathbb{W}^{K}m, \mathbb{W}^{V}m\right\}$ Where m=0...(n-1) and n represents the total count of MHSA heads, and $\mathbb{W}^{O} \in \mathbb{R}^{m.D_k \times N}$ represents the linear transformation of the head.

In Swin Transformer model, while other layers remain unchanged, a Transformer block's normal multi-head self-attention (MHSA) module is swapped out with one based on shifted windows. A shifted window-based MHSA module and a 2-layer MLP with GELU (Gaussian Error Linear Unit) non-linearity in between makes up a SWT block, as shown in Figure 2. Before every MHSA module, MLP, residual connection and Layer Normalization (LN) layer is applied.

The input Mel spectrogram feature, which was obtained during the architecture's feature extraction phase, is first divided into non-overlapping patches using a patch-splitting module in the SWT model for SER. Every patch is considered a "token," with its feature consisting of a concatenation of the raw

features from the Mel spectrogram for each patch. As depicted in our architecture, the input size is of $H \times W \times C$ dimension, where C represent the channel and $H \times W$ denotes the size of the input. The window size for each patch is set to $C/2$, since a 22 patch size is employed in our experiment, and each patch's feature dimension is expressed as $2 \times 2 \times 3 = 12$. A uniform division of window to standardize the dimension $(C/2, C/2)$ is applied. Thereafter, a linear embedding layer (64 size) is utilized to project the output for appropriate transformation. Patch merging and feature transformation are then utilized to downsize the number of tokens as the network size deepens. At the topmost level of the model, an activation function with one dense layer is used for the final classification of seven emotions. The summary of our proposed model structure with the corresponding parameter size is shown in Table 1.

Table 1: Model Structure Summary

Layer(type)	**Output Shape**	**Param #**
input_1 (InputLayer)	[(None, 32, 32, 3)]	0
random_crop (RandomCrop)	(None, 32, 32, 3)	0
random_flip (RandomFlip)	(None, 32, 32, 3)	0
patch_extract(PatchExtract)	(None, 256, 12)	0
patch_embedding(PatchEmbedding)	(None, 256, 64)	17216
swin_transformer(SwinTransformer)	(None, 256, 64)	50072
swin_transformer_1(SwinTransformer)	(None, 256, 64)	51096
patch_merging(PatchMergin)	(None, 64, 128)	32768
global_average_pooling1d(GlobalAveragePooling1D)	(None, 128)	0
dense_10 (Dense)	(None, 7)	903

3.4 Evaluation Metrics

An evaluation technique that is suitable for a multiclass problem like SER is utilized in this study. This is necessary to ensure that all the seven classes of emotions from three datasets are captured. We adopted standard evaluation metrics of Accuracy, Precision, Recall and F1-Score respectively in our experiment to establish the performance of our proposed model. We also utilize a confusion matrix that indicates the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) of our model prediction over seven emotions from the dataset. The accuracy metric estimates the percentage of emotional utterances that are correctly predicted by the proposed SWTSER model from the whole speech sample. Given a number speech sample denoted by N, accuracy is computed mathematically from equation 3.

$$Accuracy = \frac{1}{N} \sum_{i=1}^{N} \frac{(TP + TN)^i}{(TP + TN + FP + FN)^i} \tag{3}$$

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\frac{(TP)^i}{(TP+FN)^i} \tag{4}$$

$$Presicision = \frac{1}{N}\sum_{i=1}^{N}\frac{(TP)^i}{(TP+FN)^i} \tag{5}$$

While recall estimates the number correct positive prediction made against the wrong prediction by the model, precision estimates the actual positive prediction only as indicated in equation 4 and 5.

3.5 Experiment

The proposed model in this study was implemented on GPU- T4, $10900K$@$3.70Ghz$, 64GB RAM and Google Collab platform. The model implementation parameters were carefully chosen to achieve an enhanced result in SER while minimizing emotion misclassification, as indicated in table 2. The two datasets described in the preceding section were used in carrying out a speaker-independent experiments to evaluate the proposed transformer model. We utilized a split ratio (80 : 20) mechanism for splitting the datasets into training set and test set. This dataset splitting technique is a standardized approach that can help in achieving improved performance in SER models and to prevent overlapping results of emotion classification. Figure 5 - 7 presents the performance of the proposed SWT transformer model applied to the datasets, trained with 128 feature set.

Table 2: Hyperparameters employed for this study

SN	Hyperparameter	Value
1	Window size	7
2	MLP dimension	256
3	Shift size	1
4	Optimizer	Adam
4	Trainable parameters	4,166,151
5	Patch size	2,2
6	Loss function	Categorical Cross Entropy
7	Epoch	150
8	Embedded Dropout	0.02
9	Learning rate	1e-2
10	Number of Heads	8
11	Batch size	128
12	Weight decay	0.0001

3.6 Experimental Result and Discussion

The two major datasets are used in this study with the third one being a hybrid of the two datasets to determine the robustness of the proposed

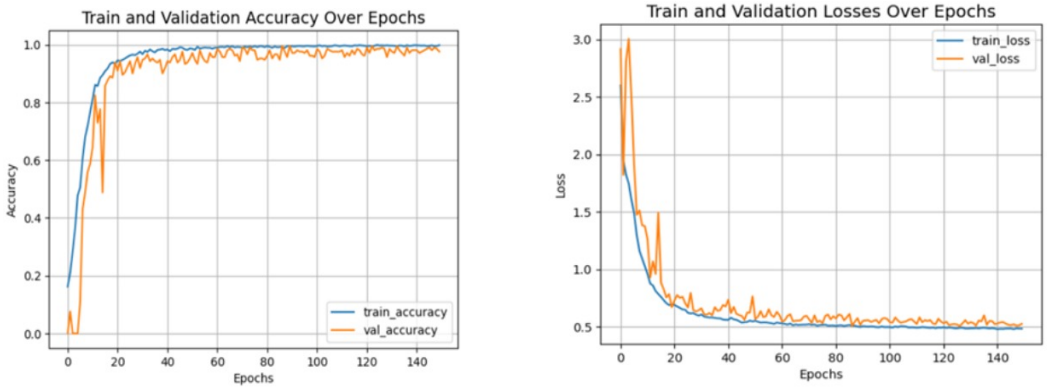


Fig. 5: Validation accuracy and loss plot on TESS

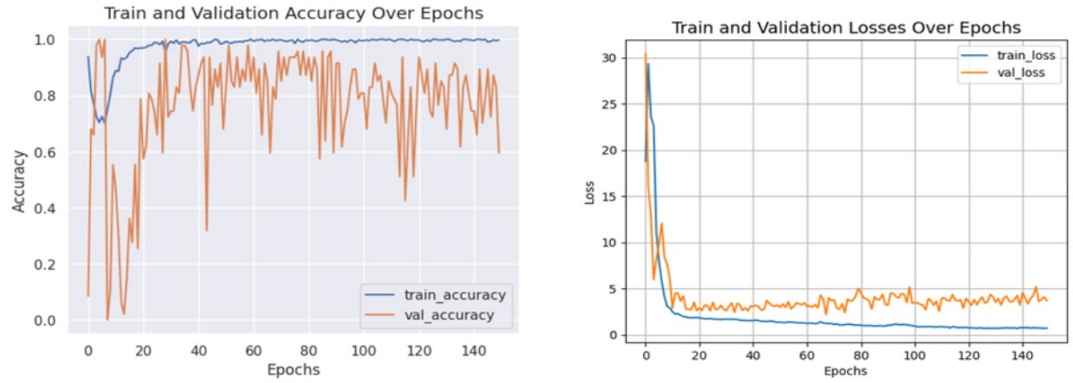


Fig. 6: Validation accuracy and loss plot on EMOVO

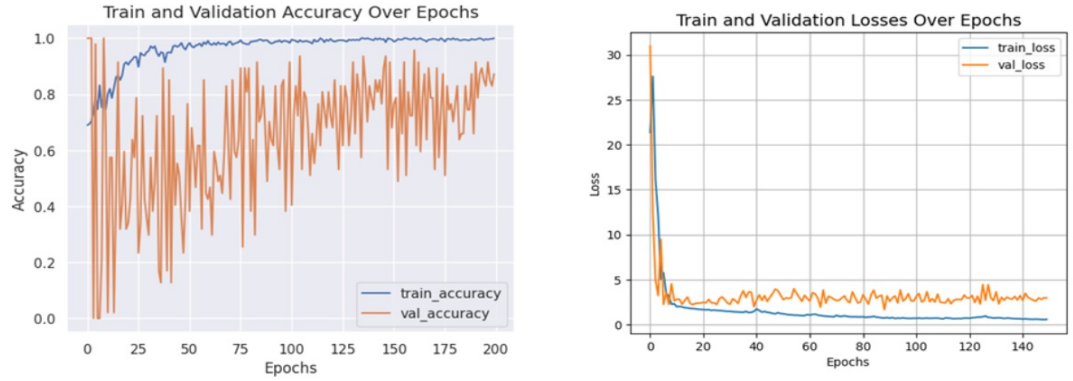


Fig. 7: Validation accuracy and loss plot on bilingual TESS-EMOVO

model to bilingual speech utterances. As previously discussed, the standard valuation metrics were Accuracy, F1-Score, Precision and Recall respectively. After extensive experiment model parameters optimization, the proposed SWT model for speech emotion recognition achieve a state-of-the-earth result on TESS dataset with 98% overall accuracy. We obtained highest precision and recall of 100% (Figure 4a) on four emotions (disgust, fear, happy and surprise)

respectively, while an F1-score of 98% and above was recorded on angry, disgust, fear and happy emotions. This high accuracy indicates that the model improved performance, exceptionally and effectively well, in capturing the variations in emotional expressions present in the TESS dataset, highlighting its robustness in speech emotion recognition.

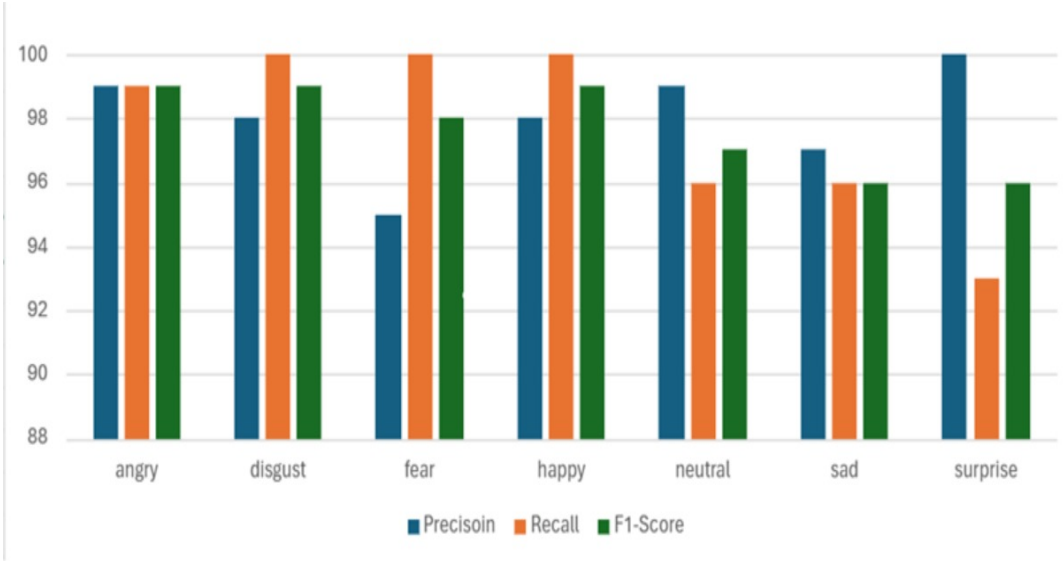


Fig. 8: Emotional level Recognition Report with three metrics (Precision, Recall and F1-Score) on TESS Dataset

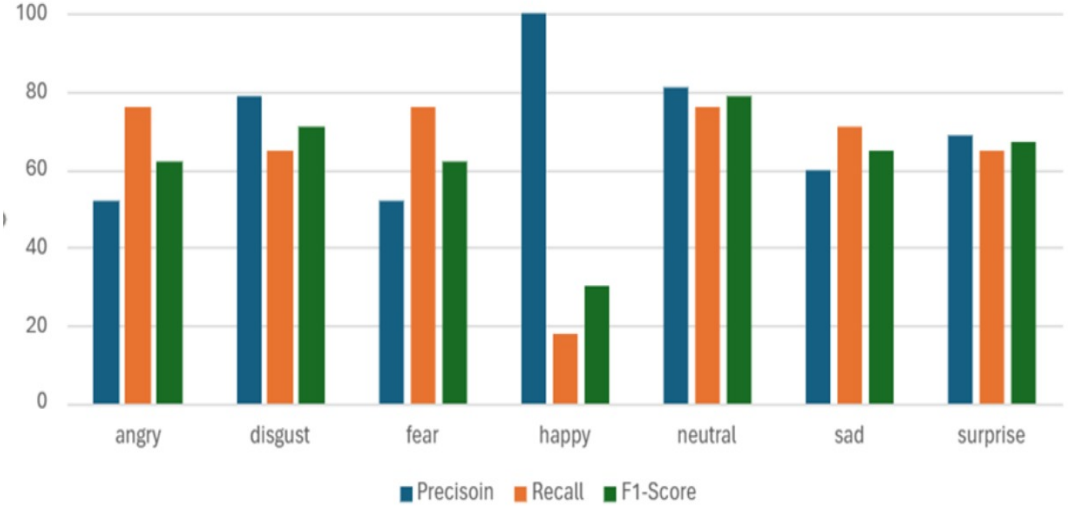


Fig. 9: Emotional level Recognition Report with three metrics (Precision, Recall and F1-Score) on EMOVO Dataset

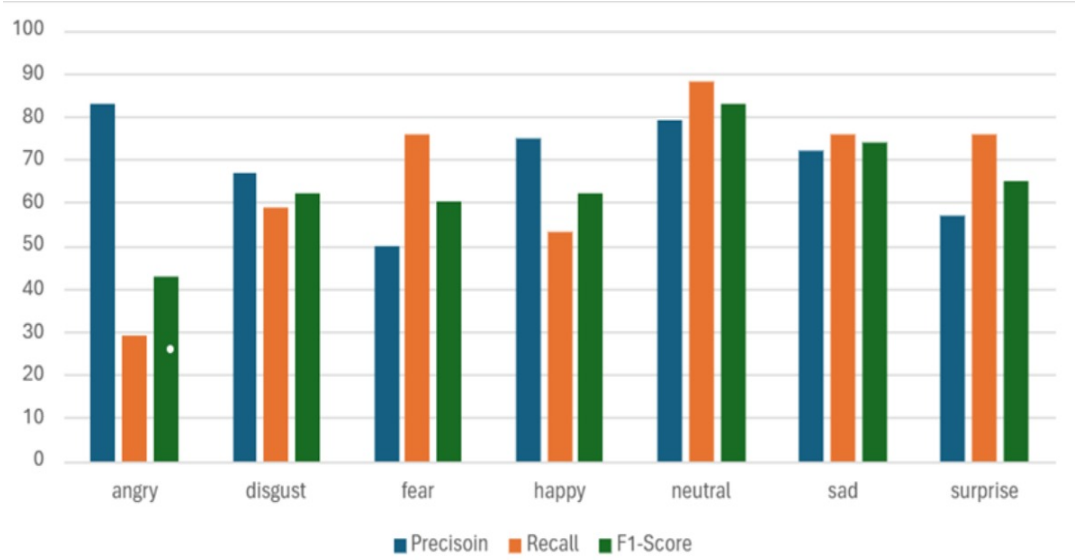


Fig. 10: Emotional level Recognition Report with three metrics (Precision, Recall and F1-Score) on Bilingual TESS-EMOVO Dataset

On EMOVO dataset, a promising result of 64% accuracy was recorded with an F1-score of 60% above on six emotions comprising of angry, disgust, fear, neutral, sad and surprise respectively (Figure 4b). The peculiar reason observed in the drop in accuracy as compared to TESS can be attributed to emotional variability and dataset size. The hybrid bilingual dataset of TESS-EMOVO achieved an accuracy of 66%, with a minimum of 60% and above for F1-Score on six emotions excluding anger as shown in Figure 4c. It shows that the model can handle a variety of linguistic variations and emotional expressions found in a bilingual dataset that has been integrated. The robustness of the transformer model in natural language processing played out in this study as it achieved a promising result on speech utterances of different languages.

Furthermore, we utilized the confusion matrix to evaluate the performance of our proposed model in SER. Figure 6 shows the confusion matrices for the three datasets, respectively. The highest accuracy of 100% and 76% were recorded on fear and disgust emotions with TESS and EMOVO datasets. The model achieved 88% highest accuracy on the TESS-EMOVO bilingual dataset. However, angry and disgust emotions yielded the lowest accuracy, which is a strong indication that the proposed model found it difficult to recognize the two emotions accurately enough as compared to others. In all more than five emotions had the accuracy measure above 70% on the three datasets. We observed that the proposed model performed above average on six, out of seven different emotions on which it was evaluated. This is illustrated in the diagonal values obtained from the confusion matrices.

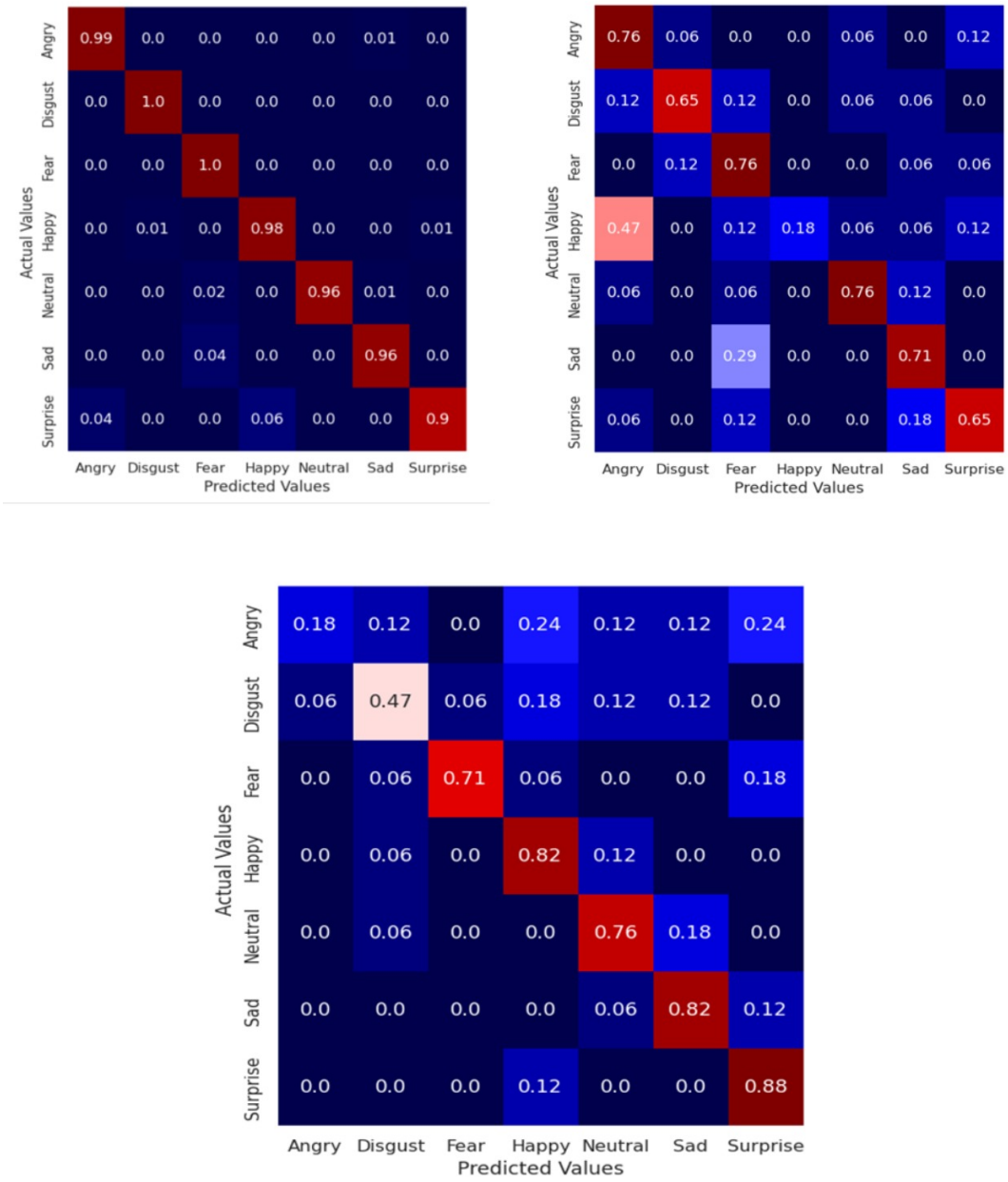


Fig. 11: TESS, EMOVO and TESS-EMOVO Confusion Matrices. (a) TESS; (b) EMOVO; (c) TESS-EMOVO. The predicted emotion is illustrated on x-axis and the actual emotion is illustrated on y-axes.

3.7 Performance Comparison

The strength and uniqueness of the proposed model were assessed by comparing its performance to other methods reported in the literature, both for training and validating the model using the same datasets. A thorough summary of the comparative study of the proposed approach and other studies can be obtained in Table 6. When compared to baseline techniques, the proposed

Table 3: Comparison with other baseline studies using TESS and EMODB dataset

Author & References	Method	Dataset	Accuracy (%)
Fasih & Saturnino[55]	DT, KNN, NB, SVM and RF	EMOVO	40.31
Fasih et al.[56]	ILFS, ReliefF	EMOVO	41.0
Ozseven [57]	KNN, MLP	EMOVO	60.40
Proposed	SWT	EMOVO	62.0
Verma, et al.[58]	SVM	TESS	96.00
Praseetha et al.[59]	DNN	TESS	89.96.
Gao[60]	CNN	TESS	81.00
Krishnan et al.[61]	Empirical Decomposition	TESS	93.30.
Chimthankar [62]	DNN	TESS	96.00.
Akinpelu & Viriri [63]	AttentionRNCA+RF	TESS	96.10.
Guizzo et al. [64]	Quantarion CNN	TESS	97.00
Choudhary et al.[65]	DNN	TESS	87.10
Proposed	**SWT**	**TESS**	**98.00**

SER model's results in the table demonstrate its significance and robustness. In terms of recognition accuracy, the transformer model outperformed the others in terms of total prediction accuracy. Because of its simplified structure, minimal complexity, and capacity to recognize seven emotions in dual language, the proposed model is appropriate for real-world applications. In summary, the proposed SER model improved emotion recognition and demonstrated resilience to linguistic variations.

4 Conclusion

The state-of-the-art novel transformer approach for speech emotion recognition has been evaluated in this work. It employed two distinct datasets for emotion recognition that came from two different languages and a hybridized speech set of dual languages. According to the results, it revealed that small feature sets can yield higher unweighted average recall, especially on bilingual speech utterances. By efficiently capturing temporal dependencies in bilingual speech data, the proposed approach of using a Shift Window Transformer for Bi-Lingual Speech Emotion Recognition seeks to overcome the shortcomings of existing models by improving the state-of-the-art in dual language speech emotion recognition systems by leveraging Transformer architectures. Despite lower performance on Italian-based language, the fusion of both datasets achieved an improved accuracy of 66% emotion classification. The results obtained, highlight how crucial the model robustness is in handling the wide range of emotional expressions and language variations found in everyday scenarios. To improve the model's generalization capabilities across several datasets and emotional settings, additional enhancements of exploring other speech signal

features such as MFCC, Chromogram, Spectral contrast are the major focus in our subsequent work. Also in our future work, we intend to explore the performance of the Shift window-based transformer on multi-lingual datasets and possible dataset augmentation to enhance the accuracy transformer-based model in speech emotion recognition tasks.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

Benchmarked publicly available dataset, Toronto English Speech Set (TESS) and EMOVO are used.

References

[1] de Lope, J., Graña, M.: An ongoing review of speech emotion recognition. Neurocomputing **528** (2023). https://doi.org/10.1016/j.neucom.2023.01.002

[2] Fahad, M., Ranja, A., Yadav, J., Deepak, A.: A survey of speech emotion recognition in natural environment. Digital Signal Process **110**, 102951 (2021)

[3] Tay, Y., Dehghani, M., Bahri, D., Metzler, D.: Efficient transformers: A survey. ACM Computing Surveys **55(6)** (2022)

[4] Uddin, M., Chowdury, M., Khandaker, M., Tamam, N., Sulieman, A.: The efficacy of deep learning-based mixed model for speech emotion recognition. Computers, Materials & Continua **74(1)**, 1709–1722 (2023). https://doi.org/10.32604/cmc.20231177

[5] Dutta, S., Ganapathy, S.: Multi-modal transformer with learnable frontend and self attention for emotion recognition. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Singapore, 23–27 May, 6917–6921 (2022)

[6] Wang, K., Haoran, X., Di, Z., Kee-Lee, C.: Emotion recognition based on convolutional neural networks and heterogeneous bio-signal data sources. Information Fusion **77**, 107–117 (2022)

[7] Lee, C., Mower, E., Busso, C., Narayanan, S.: Emotion recognition using a hierachical binary decision tree approach. Speech Communication **53(9)**, 1162–1171 (2011)

[8] Liang, J., Li, R., Jin, Q.: Semi-supervised multi-modal emotion recognition with cross-modal distribution matching. in Proceedings of the 28th ACm International Conference on Multimedia **42(4)**, 2852–2861 (2020)

[9] Chandaka, S., Chatterjee, A., Munshi, S.: Support vector machine employing cross-correlation for emotional speech recognition. Measurement **42(4)**, 611–8 (2009)

[10] Ekan, P.: Basic emotions. In Book Chapter, Handbook of Cognition and Emotion, San Francisco, USA (199)

[11] Song, Q., Sun, B., Li, S.: Multimodal sparse transformer network for audio-visual speech recognition. IEEE Transactions on Neural Networks and Learning Systems (2022)

[12] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030v2 (2021)

[13] Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Eyben, F., B., S.: Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. Sensors (Switzerland) (2022)

[14] Siddique, L., Aun, Z., Heriberto, C., Fahad, S., Moazzam, S., Junaid, Q.: Transformers in Speech Processing: A Survey. arXiv:2303.11607v1 [cs.CL] (2023)

[15] Dong, L., Xu, S., Xu, B.: Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. in 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), 5884–5888 (2018)

[16] Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., Zhong, J.: Attention is all you need in speech separation. in ICASSP 2021- 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 21–25 (2021)

[17] Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N., Yamamoto, R., X., W.: A comparative study on transformer vs rnn in speech applications. in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). **22(7)** (2019)

[18] Lin, T., Wang, Y., Liu, X., Qiu, X.: A survey of transformers. AI Open (2022)

[19] Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S., Kumar, S.: Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In ICASSP 2020- 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7829–7833 (2020)

[20] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., M., F.: Huggingface's transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6

[21] Dev-Priya, G., Kushagra, M., Ngoc-Duy, N., Natesan, S., Chee, P.: Towards an efficient backbone for preserving features in speech emotion recognition: deep-shallow convolution with recurrent neural network. Neural Computing and Applications **35** (2023). https://doi.org/10.1007/

s00521-022-07723-2

[22] Dujaili, M., Ebrahimi-Moghadam, A., Fatlawi, A.: Speech emotion recognition based on svm and knn classifications fusion. Int. J. Electr. Comput. Eng. (IJECE) **11**, 1259–1264 (2021)

[23] Jain, M., Narayan, S., Balaji, K., Bharath, K., Bhowmick, A., Karthik, R., Muthu: Speech emotion recognition using support vector machine. arXiv:2002.07590 (2013)

[24] Cheng, X., Duan, Q.: Speech emotion recognition using gaussian mixture model. In Proceedings of the 2012 International Conference on Computer Application and System Modeling (ICCASM), Taiyuan, China, 1222–1225 (2012)

[25] Edmondo, T., Marko, G.: A survey of hybrid ann/hmm models for automatic speech recognition. Neurocomputing **37** (2001)

[26] Bharti, D., Kukana, P.: A hybrid machine learning model for emotion recognition from speech signals. In Proceedings of the 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, September, 10–12 (2020)

[27] Dargan, S., Kumar, M., Ayyagari, M., Kumar, G.: A survey of deep learning and its applications: A new paradigm to machine learning. Arch. Comput. Methods Eng. **27**, 1071–1092 (2019)

[28] Manohar, K., Logashanmugam, E.: Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm. Knowledge-Based Systems **246** (2022)

[29] Harar, P., Burget, R., Dutta, M.: Speech emotion recognition with deep learning. In Proceedings of the 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Delhi, India, 2–3 February, 137–140 (2017)

[30] Poon-Feng, K., Huang, D., Dong, M., Li, H.: Acoustic emotion recognition based on fusion of multiple feature-dependent deep boltzmann machines. In Proceedings of the 9th International Symposium on Chinese Spoken Language Processing, Singapore, 584–588 (2014)

[31] Atmaja, B., Akagi, M.: Speech emotion recognition based on speech segment using lstm with attention model. In Proceedings of the 2019 IEEE International Conference on Signals and Systems, Kuala Lumpur, Malaysia,18–19 September, 40–44 (2019)

[32] Mirsamadi, S., Barsoum, E., Zhang, C.: Automatic speech emotion recognition using recurrent neural networks with localattention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March, 2227–2231 (2017)

[33] Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., Schuller, B.: Speech emotion classification using attention-based lstm. IEEE/ACM Trans. Audio Speech Lang. Process **27**, 1675–1685 (2019)

[34] Zeng, Y., Mao, H., Peng, D., Yi, Z.: Spectrogram based multi-task audio classification. Multimed. Tools Appl, 3705–3722 (2017)

[35] Popova, A., Rassadin, A., Ponomarenko, A.: Emotion recognition in sound. In Proceedings of the International Conference on Neuroinformatics, Moscow, Rus-sia, 2–6 October, Springer **25(10)**, 117–124 (2017)

[36] Issa, D., Fatih-Demirci, M., Yazici, A.: Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control **59**, 101894 (2020)

[37] Li, H., Ding, W., Wu, Z., Liu, Z.: Learning fine-grained cross-modality excitement for speech emotion recognition. arXiv:2010.12733 (2010)

[38] Francesco, P., Pasquale, F., Giovanna, M.: Speech emotion recognition with artificial intelligence for contact tracing in the covid-19 pandemic. Cognitive Computation and Systems (2023). https://doi.org/10.1049/ccs2.12076

[39] Zeng, M., Xiao, N.: Effective combination of densenet and bilstm for keyword spotting. IEEE Access **7**, 10767–10775 (2019)

[40] Puri, T., Soni, M., Dhiman, G., Khalaf, O., Khan, I.: Detection of emotion of speech for ravdess audio using hybrid convolution neural network. Hindawi Journal of Healthcare Engineering (2022)

[41] Kwon, S.: Mlt-dnet: Speech emotion recognition using 1d dilated cnn based on multi-learning trick approach. Expert Syst. Appl. **167** (2021)

[42] Pawar, M., Kokate, R.: Convolution neural network based automatic speech emotion recognition using mel-frequency cepstrum coefficients. Multimed. Tools Appl. **80**, 15563–15587 (2021)

[43] Bhangale, K., Kothandaraman, M.: Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. Electronics (Switzerland) **12(4)** (2023)

[44] Badshah, A., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M., Baik, S.: Deep features-based speech emotion recognition for smart affective services. Multimed. Tools Appl **78** (2019)

[45] Andy, T., Shang-Wen, L., L., H.-y.: Tera: Self-supervised learning of transformer encoder representation for speech. IEEE/ACM Tans. Audio, Speech, Language Process **29**, 2351–2366 (2021)

[46] Han, S., Leng, F., Jin, Z.: Speech emotion recognition with a resnet-cnn-transformer parallel neural network. In Proceedings of the International Conference on Communications, Information System and Computer Engineering(CISCE), Beijing, China, 13-16 May (2023)

[47] Vijay, J., Yasutomo, K.: Audioand video-based emotion recognition uisng multimodal transformers. In Proceedings of International Conference on Pattern Recognition, 2582–2588 (2022). https://doi.org/10.1109/ICPR56361.2022.9956730

[48] Syed, A., Siddique, L., Junaid, Q.: Cross-Language Speech Emotion Recognition Using Multimodal Dual Attention Transformers

[49] Zirou, L., Shaoping, S.: Speech emotion recognition based on swin-transformer. Journal of Physics: Conference Series **2508 (2023) 012056** (2023). https://doi.org/10.1088/1742-6596/2508/1/012056

[50] Dupuis, K., Kathleen, M.: Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set. Canadian Acoustics - Acoustique Canadienne **39(3)** (2011). https://doi.org/10.3389/fphys.2021.6432028

[51] Giovanni, C., Iacopo, I., Massimiliano, T.: Emovo corpus: an italian emotional speech database. in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) (2014)

[52] Domingos, P.: A few useful things tok now about machine learning. Commun. ACM (2012)

[53] Liu, M., Ren, S., Ma, S., Jiao, J., Chen, Y., Wang, Z., Song, W.: Gated transformer networks for multivariate time series classification (2021)

[54] Aderberg, M., Simonyan, K., Zisserman, A.: Spatial transformer networks. Adv. Neural Inf. Processing Syst **28** (2015)

[55] Fasih, H., Saturnino, L.: Proceedings of the annual conference of the international speech communication association, interspeech. Computers in Biology and Medicine (2021). https://doi.org/10.21437/Interspeech.2021-1761

[56] Fasih, H., Senja, P., Pierre, A., L., S.: Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods. Computer Speech and Language (2021). https://doi.org/10.1016/j.csl.2020.101119

[57] Ozseven, T.: A novel feature selection method for speech emotion recognition. Applied Acoustics (2019). https://doi.org/10.1016/j.apacoust.2018.11.028

[58] Verma, D., Mukhopadhyay, D.: Age Driven Automatic Speech Emotion Recognition System. In: In Proceeding of IEEE International Conference on Computing, Communication and Automation (2017)

[59] Praseetha, V., Vadivel, S.: Deep learning models for speech emotion recognition. Journal of Computer Science **14(11)** (2018)

[60] Gao, Y.: Speech-Based Emotion Recognition. https://libraetd.lib.virginia.edu/downloads/2f75r8498?filename=1_Gao_Ye_2019_MS.pdf (2019)

[61] Krishnan, P., Joseph Raj, A., Rajangam, V.: Emotion classification from speech signal based on empirical mode decomposition and non-linear features. Complex and Intelligent Systems **7(4)**, 1919–1934 (2021). https://doi.org/10.1007/s40747-021-00295-z

[62] Chimthankar, P.: Speech Emotion Recognition using Deep Learning. http://norma.ncirl.ie/5142/1/priychimtankar.pdf (2021)

[63] Akinpelu, S., Viriri, S.: Speech emotion classification using attention based network and regularized feature selection. Scientific Report **13 11990** (2023). https://doi.org/10.1038/s41598-023-38868-2

[64] Guizzo, E., Weyde, T., Scardapane, S., Comminiello, D.: Learning speech emotion representations in the quaternion domain. IEEE/ACM Transactions on Audio, Speech, and Language Processing **31** (2022). https://doi.org/10.48550/arXiv.2204.02385

[65] Choudhary, R., Meena, G., Mohbey, K.: Speech emotion based sentiment recognition using deep neural networks. Journal of Physics: Conference Series **2236(1)** (2022)

Chapter 6

Results and Discussion

6.1 Overview

This chapter presents a discussion of the overall results of this thesis. This chapter harmonizes the results obtained across all the papers presented in this thesis to show how challenges identified in the emotion classification from speech have been addressed in this thesis. A summary of challenges identified from the introduction and literature review is presented while subsequent sections present a summary of approaches used to resolve the challenges identified.

6.2 Results Summary from the Thesis

The introduction and literature review are contained in chapters 1 and 2 of this thesis. These chapters covered a variety of methods, along with their drawbacks, for classifying emotions from speech signals. The study titled "Deep learning framework for speech emotion classification: A survey of state-of-the-arts" was presented in Chapter 2. This work conducted a comprehensive analysis of the most recent methods for classifying speech emotions and contrasted them with traditional methods. However, some challenges made it difficult for these methods to analyze speech signals and classify emotions. The following are the findings from the literature review and chapter one that highlight the necessity of this study:

1. Deep learning techniques limitations owing to the scarcity of label speech emotion dataset:
   Deep learning techniques are mostly used to hierarchically learn the features corresponding to the emotional cues from speech signals by using huge labeled datasets. In speech emotion classification, where there is a scarcity of labeled datasets, using limited labeled data can lead to overfitting and poor generalization. Generally, they require large training datasets to create effective models. Due to security concerns, human speech datasets are more

difficult to obtain than image datasets. The low sensitivity of the model to emotional information can also result from training deep learning techniques with a smaller amount of data. This issue requires an optimal deep transfer learning approach that can manage the differences in emotional utterance representation and class labels between the source and target data.

2. : Over-reliance of deep learning methods on tuning of millions of parameters:
   Deep learning approaches mainly depend on the appropriate tuning of millions of parameters, which frequently results in overfitting, subpar generalization, and excessive computational resource usage [40]. Additionally, this process raises the memory and processing resource requirements for an effective system. A time-based emotion classification system must build a strong system that can quickly and efficiently handle computer resources, even though the widespread availability of GPUs has been able to counteract this effect. Computational resources should be managed efficiently to promote system portability on low-memory devices for convenient access. For effective performance speech emotion categorization, deep learning architecture still has the opportunity for some advancements and best practices.

3. Quick and accurate automatic recognition of emotion from speech signal:
   The presence of artifacts such as silence fillers and background noise is characteristic of spectrogram image representations. Emotion classification errors may result from these artifacts and noise. Conventional SEC systems tried to address this issue by using pre-processing techniques to remove artifacts before the classification process. However, this possibly increased the computational requirements, processing time, and overall complexity of the system. Additionally, emotions that lead to depression can result in chronic mental illness or premature death. Human well-being will be enhanced by early and accurate emotion classification from speech that is not limited by place, time, or obstruction, unlike facial expression identification. The need to address these issues through an effective framework that captures long-range dependencies for quick classification of emotions from speech signals is pertinent.

4. Misclassification and difficulty in accurate classification of speech utterance with language and cultural variations:
   Despite the recent progress in deep learning for speech emotion Classification, models for cross-cultural adaption still require significant improvement. Speakers' diverse cultural backgrounds contribute to the diversity of speech. Speaking utterances are associated with other languages in addition to the cultural background. This will overcome the shortcomings of existing systems that find it difficult to account for language differences when working with sequence data. Examining and creating deep learning methods that allow models to generalize in a variety of language and cultural contexts, without concentrating on local features alone is required.

5. Over-blotted features as a result of abnormalities in their spectral properties

and irregularities in speech features:
Some speech signals exhibit abnormalities in their spectral properties, leading to over-blotted features. This poses a challenge for deep learning techniques that aim to accurately analyze speech and classify emotions. For instance, because of their striking similarity in energy and pitch characteristics, it may be incorrect to identify the emotions of surprise, happiness, and anger in spoken utterances. When classifying emotion, the systems may become confused by these striking similarities. Furthermore, it is impossible to overemphasize the likelihood of many characteristics arising from overlapping emotions in a multiclass problem such as speech emotion classification. However, using unique feature extraction approaches, the traditional and certain early SEC systems sequence models attempted to address this issue. However, due to the heavy reliance on handcrafted features with low model generalization ability and limited capacity to create high-level representations, their effectiveness has been limited. An effective deep Learning framework that can handle these problems by introducing an extended deep transfer learning with feature selection approaches that minimize misclassification of emotion is required.

6. Selection of high-level emotional relevant features for efficient classification of speech emotion:
Though many novel feature extractions have been developed, however, their performance is limited without a corresponding feature selection approach. Besides, not all speech utterances carry relevant emotional features and extraction of local features by deep network layers alone for accurate emotion classification is not sufficient. Therefore, the incorporation of an advanced feature selection technique with an attention-based deep learning network for the selection of emotionally relevant features, toward efficient classification of emotion from speech signal.

6.2.1 Resolving the challenge of scarce annotated speech emotion dataset

In resolving the challenge associated with a limited annotated dataset, we utilized deep transfer learning (DTL) in sections 3.2 and 3.3. with a notable performance of 96.10% and 94% achieved on two datasets that contain 3,200 spectrogram image representations of speech utterances for the process of classification. The DTL leverages end-to-end DCNN with the exclusion of top layers and combines with various classifiers to build a deep learning model for extracting emotional-related features relevant to emotion classification from auditory speech utterances.

The first paper in section 3.2, titled A Robust Deep Transfer Learning Model for Accurate Speech Emotion Classification, utilized a DTL model that is based on DCNN for the classification of speech emotion. The proposed model comprised speech processing and audio analysis to produce spectrogram image representation

Table 6.1: Performance comparison of two classifiers

Experiment	Specificity	Accuracy	UAR
DCNN-PCA-MLP	**97.4%**	**96.1%**	**98.7%**
DCNN-PCA-RF	96.1%	94.0%	97.4%

of speech signal and VGGNet base architecture for extracting local and high-level features. The architecture is enhanced with a features down-sampling technique of principal component analysis for extracting emotionally related features before it is passed onto the classifier layer for the eventual classification of speech emotion. The outcome shown in Tables 6.1, 6.2, and Figure 6.1 illustrates the achievement of the model. The performance evaluation in comparison with some recent methods used, indicated it outperforms the state-of-the-art approaches.

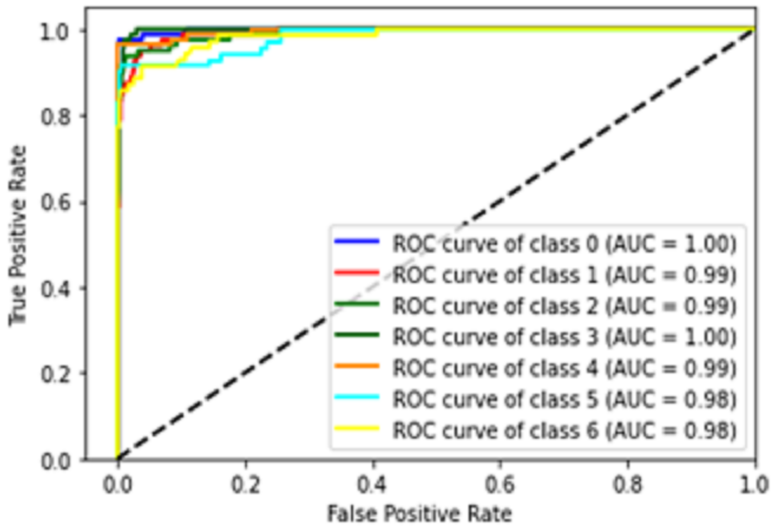


Figure 6.1: Model Accuracy Curve

Table 6.2: Model Performance based on accuracy with some recent state-of-the-art frameworks

Publication	Techniques	Dataset	Reported Accuracy
Praseetha et al.(2018) [40]	DNN-GRU	TESS	89.96%
Venkataramanan et al.(2019) [26]	DNN-LSTM	TESS	70.00%
Krishnan et al.(2021) [48]	IMF-SVM, KNN	TESS	93.30%
Blumentals et al.(2022) [9]	LSTM-FCNN	TESS	86.02%
Proposed Model	**DCNN-PCA-MLP**	**TESS**	**96.10%**

The model has an accuracy of 96.10% on seven different emotions, with a specificity of 97.4% on the TESS dataset.

In the second paper (section 3.3), titled Robust Feature Selection-Based Speech Emotion Classification Using Deep Transfer Learning, the performance of the first

paper was improved through the introduction of feature selection techniques that learn more relevant emotional cues from speech signals. The paper experimented on three datasets, namely TESS and EMODB to establish its generalizability and application in real-life scenarios. An accuracy of 96.1% and 94.90% with a specificity ratio of 98.9% and 100% on both datasets respectively. This can be confirmed by the confusion matrix in Figure 6.2

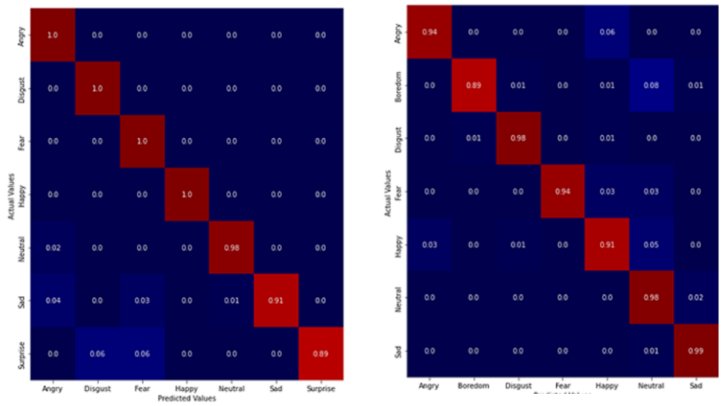


Figure 6.2: Classification of speech emotion on TESS and EMODB datasets: The diagonal value represents the normalized confusion matrix of predicted emotion against the actual emotion

6.2.2 Resolving the challenge of Over-reliance on deep learning methods on the tuning of millions of parameters:

An attempt to resolve the challenge of heavy dependence on turning of millions of parameters by deep learning techniques was proposed in chapter 4 with a paper titled, Lightweight Deep Learning Framework for Speech Emotion Recognition. The research created a unique lightweight DCCN-based VGGNets framework to develop an effective classification system that forgoes learning redundant feature maps to increase classification accuracy. Since it doesn't need redundant convolutional layers, the model has fewer parameters than a comparable conventional CNN. Table 6.3, Fig. 6.3, and 6.4 attest to the remarkable accuracy and reliability of this framework in the classification of speech emotion. Not only that, but it also has the capability of running on low-memory devices without degeneration in efficiency, because of a lesser number of parameters it involved. The model in this paper was benchmarked on three datasets TESS, EMODB, and RAVDESS respectively. It achieved an accuracy of 100%, 96.03%, and 86.25% on the three datasets.

Figure 6.3: Showcasing how the model learned features from MFCC image representation of speech signal.

Table 6.3: Performance comparison of the model in terms of lesser parameter with other CNN model.

Model	**Size(MB)**	**Parameters**	**Dataset**	**Accuracy(%)**	**Average**
DenseNet	7.04	7,043,654	DT1	86.13	85.65
			DT2	79.24	
			DT3	91.60	
VGGNet	**7.64**	**7,640,903**	**DT1**	**96.03**	**94.09**
			DT2	86.25	
			DT3	100.00	
InceptionNet	9.64	9,604,544	DT1	62.30	75.36
			DT2	75.60	
			DT3	88.20	
MobileNet	4.80	4,806,855	DT1	48.50	53.92
			DT2	81.32	
			DT3	31.96	
ResNet	9.12	9,116,032	DT1	87.12	89.67
			DT2	83.90	
			DT3	98.00	

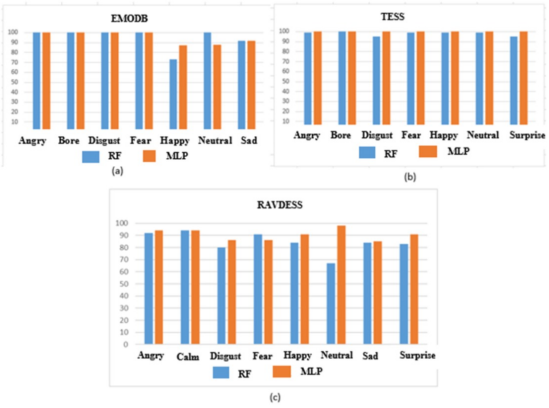


Figure 6.4: Accuracy of emotion classification chart with Neutral emotion showing the highest recognition from the three datasets followed by Surprise, Fear, Disgust, and Calm.

6.2.3 Resolving the challenge associated with quick and accurate automatic classification of emotion from speech signal:

The paper that addressed this challenge is in section 5.2, entitled An Enhanced Speech Emotion Recognition using Vision Transformer. The deep learning model proposed improved on the previous paper in terms of memory efficiency with fewer trainable parameters(4,166,151) while extracting emotional features from speech signals. Mel-spectrogram representation was extracted from raw speech signal with 128M mel bins to map the frequency onto the Mel scale. Each audio sound was split into frames of 25ms, with a 10ms gap between each frame, to avert feature degradation. The model achieved an accuracy of 98%, 94%, and 91% on three datasets, utilized in the study. An extensive experiment for the classification of seven classes of speech emotion was carried out on TESS, EMODB, and a hybridized TESS-EMODB with six emotion classifications. An ablation study was also conducted on the functional components of the model architecture to ascertain its robust capability to quickly classify emotion irrespective of background noise. The result obtained was compared with other architectures and it proved superior performance as shown in Table 6.4 and Figure 6.5 respectively.

Table 6.4: Comparative experiments on two datasets with other deep learning architecture

Architectures	Number of Parameters	Dataset	Accuracy
ResNet	9,116,032	SDT1	87.12
		SDT2	83.90
MobileNet	4,806,855	SDT1	48.50
		SDT2	81.32
InceptionNet	9,604,544	SDT1	62.30
		SDT2	75.60
DenseNet	7,043,654	SDT1	86.13
		SDT2	79.24
ViTSER	**4,166,151**	**SDT1**	**91.03**
		SDT2	**98.00**

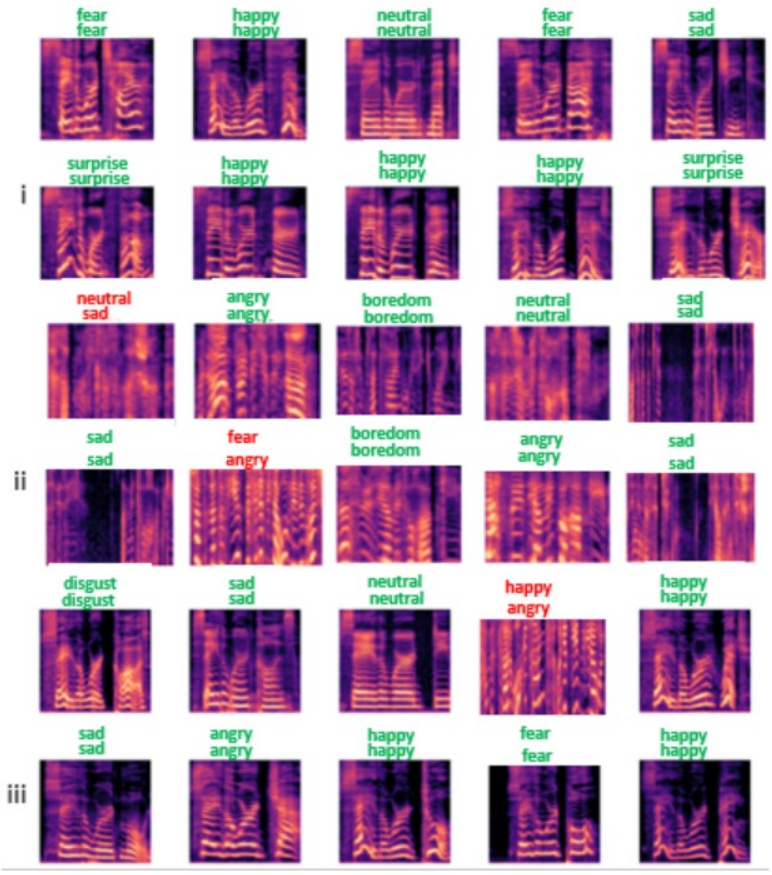


Figure 6.5: Emotion recognition output of the proposed model on three datasets: (i) represents recognition output on the TESS dataset (ii) represents recognition output on the EMODB dataset (iii) represents recognition output on the TESS-EMODB dataset. The first row of the label is the grand truth, while the second row represents the classified emotion.

6.2.4 Resolving the challenge of Misclassification of emotion from speech utterance with language and cultural variations:

The paper entitled SwinTSER: An Improved Bilingual Speech Emotion Recognition Using Shift Window Transformer in section 5.3 attempted to address the challenge here by proposing a deep learning technique that extracts salient emotional features from speech utterances of varied language. Speaking utterances are associated with other languages in addition to the cultural background. Therefore, creating deep learning methods that allow models to generalize in a variety of language and cultural contexts, without necessarily concentrating on local features only formed the focus of the paper. The paper utilized a self-attention mechanism that is computed within local windows in a non-overlapping fashion to reduce computational complexity that may result from Bi-Lingual speech utterance for SEC. A shift-window

transformer that consists of a 2-layer MLP with GELU non-linearity in between and a shifted window-based MHSA module is employed. Layer Normalization (LN), MLP, and residual connection are used before each MHSA module. Fewer parameters with efficient layers that learn discriminative features from the input speech signal features increase the robustness of the model.

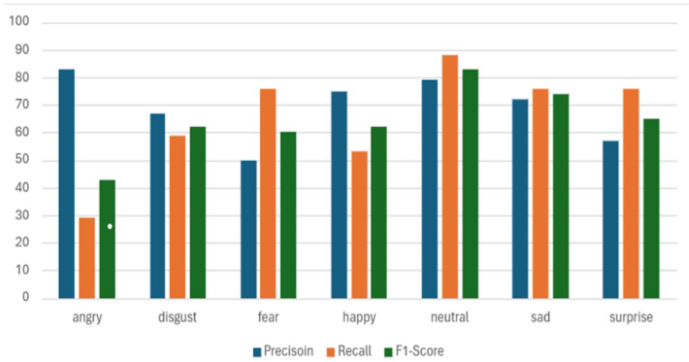


Figure 6.6: Emotional Level Classification Report with three metrics (Precision, Recall and F1-Score) on Bilingual TESS-EMOVO Dataset

Figure 6.6 demonstrates the performance of the model on a combined dataset while Table 6.5 reveals the performance of the transformer model on various patch sizes. This experiment indicated that the lower the patch size of the speech mel-spectrogram image, the lower the accuracy of emotion classification. The highest accuracy obtained was at 32 patch size.

Table 6.5: Ablation study with varying patch sizes of the mel-spectrogram representation with TESS dataset: A – Angry, H - Happy, S - Sad, D - Disgust, N - Neutral, F - Fear, B – Boredom, Sr - Surprise, P - Precision, R- Recall, F1- F1-Score

Size	Metrics	A	D	F	H	N	S	Sr	OVA(%)
14	P	0.83	0.90	0.93	0.90	0.92	0.98	0.97	91
	R	0.90	0.91	0.91	0.80	0.96	0.96	0.95	
	F1	0.86	0.91	0.92	0.85	0.94	0.97	0.96	
16	P	0.84	0.92	0.99	0.90	0.90	0.98	0.97	92
	R	0.94	0.96	0.88	0.82	0.97	0.91	0.98	
	F1	0.89	0.94	0.93	0.86	0.93	0.94	0.98	
28	P	0.86	0.96	0.98	0.90	0.95	0.98	0.98	94
	R	0.97	0.93	0.91	0.87	0.95	0.99	0.95	
	F1	0.91	0.94	0.95	0.88	0.95	0.98	0.97	
32	P	1.00	0.99	0.99	0.96	1.00	1.00	0.92	**98**
	R	0.98	0.99	0.99	0.96	1.00	0.98	0.97	
	F1	0.99	0.99	0.99	0.96	1.00	0.99	0.94	

6.2.5 Resolving the challenge of over-blotted features as a result of abnormalities in spectral properties and irregularities of speech features:

We resolve the issue associated with speech utterance with over-blotted features and irregularities in spectral characteristics in section 3.3 with a paper entitled Robust Feature Selection-Based Speech Emotion Classification Using Deep Transfer Learning. In the paper, a methodology that is based on an efficient emotional feature selection technique is utilized in combination with a deep learning model. The study makes use of a convolutional neural network (CNN) that has already been trained to efficiently extract features from mel-spectrograms that are taken from speech signals. Effective feature extraction is achieved while minimizing computational costs by freezing a large portion of the CNN layers during training. The paper utilizes the Neighborhood Component Analysis (NCA) feature selection algorithm in order to minimize feature dimensionality and prevent misclassification. The problem of over-blotted features brought on by anomalies in spectral properties and irregularities in speech features is addressed by this method, which aids in identifying the most pertinent features. At the top layer of the model, the paper uses Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) classifiers. To accurately classify emotions, these classifiers are applied after feature selection. The proposed method achieved 94.3% on EMO-DB and 97.2% on TESS datasets accuracy respectively after extensive experiments as indicated in Table 6.6 and Figure 6.7.

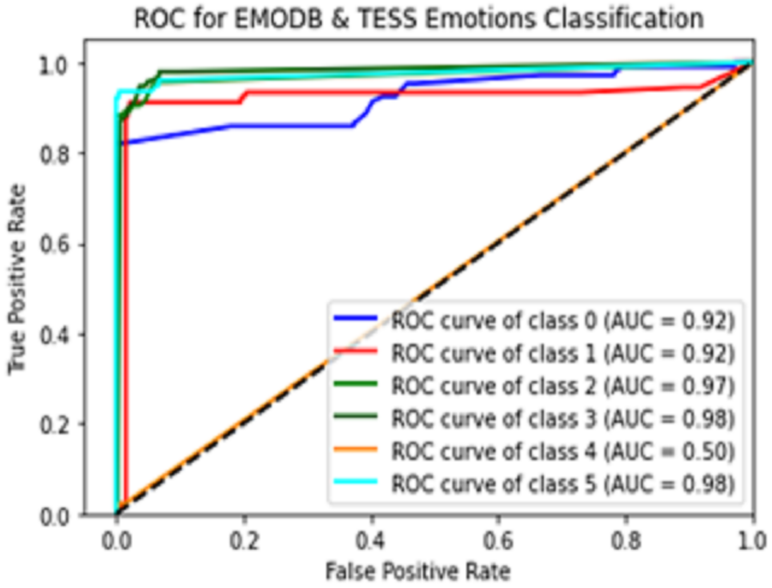


Figure 6.7: ROC curve of classification.

Table 6.6: Performance evaluation with existing deep learning techniques

Dataset	Reference	Methods	Reported Accuracy
TESS	(2018)[40]	DNN-GRU	89.96%
"	(2019)[48]	DNN-LSTM	70.00%
"	(2021)[26]	IMF-SVM, KNN	93.30%
"	(2022)[9]	LSTM-FCNN	86.02%
"	**Proposed**	**DCNN-NCA-MLP**	**96.10**%
EMO-DB	(2019)[33]	DCNN	80.79%
"	(2020)[36]	CNN-BiLSTM	85.50%
"	(2020)[5]	CNN	89.02%
"	(2021)[52]	DCNN-LSTM-Attention	87.86%
"	(2021)[35]	CNN-Attention	93.00%
"	(2021)[51]	LSTM-CNN	93.34%
"	**Proposed**	**DCNN-NCA-MLP**	**94.90**%

6.2.6 Resolving the challenge of Selecting high-level emotion-relevant features for efficient classification of speech emotion:

The paper entitled "Speech Emotion Classification Using Attention Based Network and Regularized Feature Selection" resolves the challenge of selecting high-level emotionally relevant features for efficient speech-emotion classification. The paper proposes utilizing a pre-trained convolutional neural network (CNN) in conjunction with a robust attention-based network to extract global features from speech signals that are emotionally rich as well as local features. By concentrating on prominent features, the attention mechanism improves the model's capacity to identify pertinent emotional cues. In order to enhance feature selection and boost classification accuracy, regularized neighbourhood component analysis (RNCA) is introduced in incorporated. Prioritizing emotionally relevant features, this strategy helps select the most discriminative features while minimizing misclassification. Three key classifiers (SVM, MLP, and Random Forest) are used to comprehensively assess the proposed model on the publicly TESS dataset for speech emotion classification. A substantial improvement is evident in the outcomes, as the model attained a classification accuracy of 97.8%.

Additionally, by freezing the top layer of the VGGNet and lowering the overall number of trainable parameters, the paper focuses on efficiency. This makes the model more useful for real-world applications by lowering its memory requirements and computing costs. The research demonstrates the higher performance of the proposed model in speech emotion classification by comparing it with recent techniques. Furthermore, Figure 6.8 illustrates the confusion matrices of the model's efficacy and efficiency in correctly classifying emotions into seven key categories.

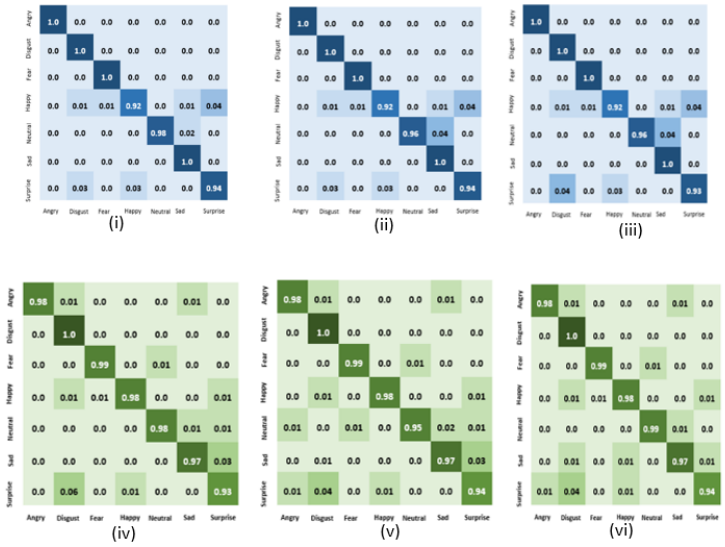


Figure 6.8: : Confusion matrix of speech emotion classification with our model on three classifiers: (i) VGG16+RNCA+RF, (ii) VGG16+RNCA+MLP, (iii) VGG16+RNCA+SVM, (iv) VGG19+RNCA+RF, (v) VGG19+RNCA+MLP, (vi) VGG19+RNCA+SVM

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this study, a deep learning framework for enhancing speech emotion classification has been presented. The deep learning techniques improved the accuracy of recognition and classification of speech emotion from speech signals through efficient feature extraction and selection mechanisms. A comprehensive survey on the current approach for carrying out this challenging task was carried out at first, which also includes the limitations of existing methods. This has led to the development of a new framework for effective and efficient classification of speech emotion from human speech utterances or speech signals.

A deep learning that is based on a deep convolutional Encoder-Decoder network was modelled for the efficient classification of speech emotion. The deep learning model in this study was designed to accurately analyze the complexity of speech features such as noise, language, cultural background, and irregularities in spectral characteristics. The multi-stage approach of the model structure paved the way for it to learn and extract features from the speech signal. The transfer learning resolves the challenge associated with a limited annotated speech emotion dataset. The introduction of efficient feature selection techniques at the middle level of the architecture prevented redundant features and the elimination of irrelevant features that could hamper the accurate classification of emotion. The performance evaluation of the system outperforms the existing method in the classification of speech emotion from speech signals and renders it to be suitable for a typical decision support system for detecting and monitoring mental health conditions like depression and anxiety. By understanding a patient's emotional state, AI-powered systems can provide tailored support and interventions.

This study also proposed a lightweight model for quick classification of speech emotion from human auditory speech, that can be deployed on low-memory devices, for healthcare management and IoT devices. The huge parameters required by most deep learning techniques which usually result in high computing resources

are overcome in this study. The tuning of millions of parameters peculiar to most deep learning architecture is prevented through the implementation of low trainable parameters model-VGGNet, without compromising the efficiency of emotion classification accuracy. The redundant layers of the model design are eliminated, while classifiers are introduced. In this research, the performance evaluation was carried out on the same computing hardware condition as well as the state-of-the-art datasets and the result of extensive comparison shows that the proposed system outperforms existing techniques in terms of computational resources, accuracy, efficiency and generalization.

7.1.1 Contribution to Knowledge

A novel deep learning framework that can efficiently manage complexity in performing the extraction and selection of emotional cues from speech signals, for the accurate classification of seven emotions has been designed and modeled.

7.1.2 Future work

.

This research work has immensely contributed to the growing field of affective computing for the classification of speech emotion, however, the following are the possible areas that can be further explored in this research:

1. Portability: the proposed system can be incorporated or embedded into emotion-aware smart devices, for quick recognition of emotion which will inform real-time decisions.

2. Generalization: Though, four different datasets-TESS, EMODB, RAVDESS, and EMOVO were used in this majorly synthetic research, with mel-spectrogram and MFCC forming the major speech signal representation for emotion classification. Further study can be carried out using more speech datasets of varying languages and cultural backgrounds, especially the non-synthetic dataset, to further improve the model's generalizability.

3. Furthermore, the combination of the speech emotion dataset and other multimodal human expressions of emotion such as facial expression(Cohn Kanade CK and Cohn Kanade Extension CK+) Database) can be implemented on the model to establish its robustness and need for possible fine-tuning.

4. Finally, the deep learning model developed in this research can further be applied to speech recognition and other affective computing-related fields to establish their generalizability as well.

List of References

[1] A. Aggarwal, A. Srivastava, A. Agarwal, N. Chahal, D. Singh, A. Alnuaim, A. Alhadlaq, and H. Lee. Two-way feature extraction for speech emotion recognition using deep learning. *Sensors*, 22(2378), 2022. 6, 8

[2] W. Alsabhan. Human–computer interaction with a real-time speech emotion recognition with ensembling techniques 1d convolution neural network and attention. *Sensors*, 23(1386), 2023. 2

[3] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: a meta- analysis. *Psychol Bull*, 111(2):256–274, 1992. 4

[4] J. Ancilin and A. Milton. Improved speech emotion recognition with mel frequency magnitude coefcient. *Applied Acoustics*, 179, 2021. 5

[5] H. Aouani and Y. B. Ayed. Speech emotion recognition with deep learning. *Int. J. Intell. Syst.*, 176:251–260, 2021. 187

[6] A. Arias, C. Williams, R. Raghvani, M. Aghajani, S. Baez, C. Belzung, L. Booij, Geraldo Busatto, Julian Chiarella, HY. Fu, Agustin Ibanez, J. Liddell, L. Lowe, P. Penninx, B.and Rosa, and A. Kemp. The emotion probe: On the universality of cross-linguistic and cross-gender speech emotion recognition via machine learning. *Neuroscience Biobehavioral Reviews*, 111:199–228, 2020. 2

[7] A. Bashit and D. Valles. A mel-filterbank and mfcc-based neural network approach to train the houston toad call detection system design. *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 22(7):438–443, 2018.

[8] M. Bhatti, Y. Wang, and L. Guan. A neural network approach for human emotion recognition in speech. In *Proceedings of the 2004 IEEE International Symposium on Circuits and Systems (ISCAS)*, Vancouver, BC, Canada, 2004. 5

[9] E. Blumentals and A. Salimbajevs. Emotion recognition in real-world support call center data for latvian language. *Jt. Proc. ACM IUI Work. Helsinki, (Finland)*, 2022. 180, 187

[10] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A database of german emotional speech (emodb). *INTERSPEECH*, pages 1517–1520, 2005. 11

[11] G. Costantini, E. Parada-Cabaleiro, D. Casali, and V. Cesarini. The emotion probe: On the universality of cross-linguistic and cross-gender speech emotion recognition via machine learning. *Sensors*, 22(7), 2022. 8

[12] A. Davletcharova, S. Sugathan, B. Abraham, and A James. Detection and analysis of emotion from speech signals. *Procedia Comput. Sci.*, 58:91–96, 2015. 8

[13] F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *In Proceedings of the Fourth International Conference on Spoken Language Processing. ICSLP, IEEE*, page 1970–1973, Philadelphia, PA, USA, 1996. 4

[14] M. Deriche and A. Absa. A two-stage hierarchical bilingual emotion recognition system using a hidden markov model and neural networks. *Arab. J. Sci. Eng.*, 42:5231–5249, 2022. 8

[15] G. Dev-Priya, M. Kushagra, N. Ngoc-Duy, S. Natesan, and P. Chee. Towards an efficient backbone for preserving features in speech emotion recognition: deep-shallow convolution with recurrent neural network. *Neural Computing and Applications*, 23:2457–2469, 2023. 6, 8

[16] Paul Ekman. Facial Expressions of Emotion: New Findings, New Questions. *Psychological Science*, S2CID 9274447, 1992. 2

[17] M. Farooq, F. Hussain, N. Baloch, F. Raja, H. Yu, and Y. Bin Zikria. Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. *Sensors*, 20(6008), 2020. 8

[18] C. Giovanni, I. Iacopo, and T. Massimiliano. Emovo corpus: an italian emotional speech database. *in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014. 11

[19] M. Greenwald, Cook E., and P. Lang. Affective judgment and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli. *J Psychophysiol*, 3:51–64, 1989. 4

[20] B. Gulmira, Y. Banu, S. Altynbek, and M. Assel. Emotional speech recognition method based on word transcription. *Sensor*, 22 (1937), 2022. 9

[21] P. Harar, R. Burget, and M. Dutta. Speech emotion recognition with deep learning. In *In Proceedings ofthe 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, India, 2017. 5

[22] M. Haytham, M. Fayek, and Lawrence C. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 2017. 6, 8

[23] 1. Izard and E. Carroll. Emotion theory and research: Highlights, unanswered questions, and emerging issues. *Review of Psychology*, 60(23):1–25, 2009. 1

[24] M. Joshi, B. Ghongade, M. Joshi, and V. Kulkarni. Deep bilstm neural network model for emotion detection using cross-dataset approach. *Biomedical Signal Processing and Control*, 22(7), 2022. 6, 7

[25] P. Kannadaguli and V. Bhat. A comparison of bayesian and hmm based approaches in machine learning for emotion detection in native kannada speaker. In *In Proceedings of the 2018 IEEMA Engineer Infinite Conference (eTechNxT)*, New Delhi, India,, 2018. 5

[26] P. Krishnan, A. Joseph, and V. Rajangam. Emotion classification from speech signal based on empirical mode decomposition and non-linear features. *Complex Intell. Syst.*, 7(4):1119–1934, 2021. 180, 187

[27] S. Lee, S. a A.nd Yildirim, A. Kazemzadeh, and S Narayanan. An articulatory study of emotional speech production. In *In 9th European Conference on Speech Communication and Technology*, Angkor Wat, Cambodia, 2005. 4

[28] S. R. Livingstone and F. A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess):a dynamic, multimodal set of facial and vocal expressions in north american english. *Article e0196391*, 13(5), 2018. 11

[29] S. Lugovic, I. Dunder, and M. Horvat. Techniques and applications of emotion recognition in speech. *Proceedings of 39th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO*, 13(6):1278–1283, 2016. 4

[30] F. Luis and R. Juan. Classification of emotions and evaluation of customer satisfaction from speech in real-world acoustic environments. *Digital Signal Processing*, 120 (103286), 2022. 1

[31] M. Maithri, U. Raghavendra, A. Gudigar, J. Samanth, D. Prabal, M. Murugappan, Y. Chakole, and U. Acharya. Auto-mated emotion recognition: Current trends and future perspectives. *Computer Methods and Programs in Biomedicine*, 215, 2022. 2

[32] X. Mao, L. Chen, and L. Fu. Recognizing emotion in speech. In *In Proceedings of the 2009 WRI World congress on Computer Science and Information Engineering, IEEE*, page 1970–1973, Los Angeles, CA, USA, 2009. 5

[33] H. Meng, T. Yan, F. Yuan, and H. We. Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE Access*, 120:125868–125881, 2019. 187

[34] Alam Monisha, Syeda Tamanna, and Sultana Sadia. A review of the advancement in speech emotion recognition for indo-aryan and dravidian languages. *Hindawi, Advances in Human-Computer Interaction*, 9602429(), 2022. 3, 4

[35] Mustaqeem and S. Kwon. Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network. *Int. J. Intell. Syst.*, 36:5116–5135, 2021. 187

[36] S. Mustaqeem Kwon. Clustering-based speech emotion recognition by incorporating learned features and deep bilstm. *IEEE Access*, 120:79861–79875, 2020. 187

[37] J. Nicholson, K. Takahashi, and R. Nakatsu. Emotion recognition in speech using neural networks. *Neural Computing and Applications*, 9(4):290–296, 2000. 4

[38] P. Oudeyer. The production and recognition of emotions in speech: features and algorithms. *Int J Hum Comput Interact Stud*, 59:157–183, 2003. 4

[39] M. K. Pichora-Fuller and K. Dupuis. Toronto emotional speech set (tess). *doi.org/10. 5683/SP2/E8H2MF. doi:10.5683/SP2/E8H2MF.*, 2020. 11

[40] V. Praseetha and S. Vadivel. Deep learning models for speech emotion recognition. *J. Comput. Sci.*, 14 (11):1577–1587, 2018. 180, 187

[41] R. Ramesh, V. B. Prahaladhan, P. Nithish, and K. Mohanaprasad. Speech emotion recognition using the novel swinemonet (shifted window transformer emotion network). *International Journal of Speech Technology*, 2024.

[42] Guiyoung S. and Kwon S. Spontaneous speech emotion recognition based on spectrogram with convolutional neural network. *The Transactions of the Korea Information Processing Society TKIPS*, 13(6), 2024. 3

[43] B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, QC, Canada, 2004. 5

[44] L. Siddique, Z. Aun, C. Heriberto, S. Fahad, S. Moazzam, and Q Junaid. Transformers in speech processing: A survey. *arXiv:2303.11607v1*, 22(7), 2023. 6, 8

[45] L. Siddique, R. Rajib, Y. Shahzad, Q. Junaid, and E. Julien. Information transfer learning for improving speech emotion classification accuracy. In *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, New Delhi, India,, 2020. 6, 8

[46] M. Soegaard and R. Friis Dam. *The Encyclopedia of Human-Computer Interaction.* Interaction Design Foundation, Switzerland, 2013. 2

[47] M. Swain, B. Maji, P. Kabisatpathy, and A. Routray. A dcrnn- based ensemble classifer for speech emotion recognition in odia language. *Complex Intelligent Systems*, 8(5):4237–4249, 2022. 5

[48] K. Venkataramanan and H Rajamohan. Emotion recognition from speech. *Audio and Speech Processing*, page 1–14, 2021. 180, 187

[49] T. Wani, T. Gunawan, S. Qadri, M. Kartiwi, and E. Ambikairajah. A comprehensive review of speech emotion recognition systems. *IEEE Access*, 9, 2021. 1, 2

[50] J. Williamson. Speech analyzer for analyzing pitch or frequency perturbations in individual speech patterns to determine the emotional state of the person. *US Patent*, 4(093), 1978. 4

[51] R. Yahia Cherif, A. Moussaouni, N. Frahta, and M. Berimi. Effective speech emotion recognition using deep learning approaches for algerian dialect. *In Proceedings of the International Conference of Women in Data Science at Taif University, WiDSTaif, Taif, Saudi Arabia*, pages 1–6, 2021. 187

[52] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan. Speech emotion recognition using fusion of three multi-task learning-based classifiers: Hsf-dnn, ms-cnn and lld-rnn. *Speech Communication*, 120:11–19, 2021. 187