Exploration of Ear Biometrics with Deep Learning

Author

Aimée Anne BOOYSENS

Supervisor

Prof. Serestina VIRIRI

A thesis submitted to the University of KwaZulu-Natal, College of Agriculture, Engineering and Science, in fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science

School of Mathematics, Statistics and Computer Science,

University of KwaZulu-Natal, Durban, South Africa.

Declaration of Authorship

I, Aimée BOOYSENS, declare that this thesis titled, 'Exploration of Ear Biometrics with Deep Learning' 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 mark, 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: Aimée Anne BOOYSENS

Signature: 

Date: 19/05/2024

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

Supervisor: Prof. Serestina VIRIRI

Signature: 

Date: 20/05/2024

Abstract

Biometrics is the recognition of a human using biometric characteristics for identification, which may be physiological or behavioural. Numerous models have been proposed to distinguish biometric traits used in multiple applications, such as forensic investigations and security systems. With the COVID-19 pandemic, facial recognition systems failed due to users wearing masks; however, human ear recognition proved more suitable as it is visible. This thesis explores efficient deep learning-based models for accurate ear biometrics recognition. The ears were extracted and identified from 2D profiles and facial images, focusing on both left and right ears. With the numerous datasets used, with particular mention of BEAR, EarVN1.0, IIT, ITWE and AWE databases. Many machine learning techniques were explored, such as Naïve Bayes, Decision Tree, K-Nearest Neighbor, and innovative deep learning techniques: Transformer Network Architecture, Lightweight Deep Learning with Model Compression and EfficientNet. The experimental results showed that the Transformer Network achieved a high accuracy of 92.60% and 92.56% with epochs of 50 and 90, respectively. The proposed ReducedFireNet Model reduces the input size and increases computation time, but it detects more robust ear features. The EfficientNet variant B8 achieved a classification accuracy of 98.45%. The results achieved are more significant than those of other works, with the highest achieved being 98.00%. The overall results showed that deep learning models can improve ear biometrics recognition when both ears are computed.

Declaration

The work described in the thesis was carried out in the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal from January 2018 to December 2023. This thesis was completed under the supervision of Professor Serestina Viriri.

This thesis 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.

Acknowledgments

I want to express my profound gratitude to my supervisor, Prof. Serestina Viriri, for his words of encouragement, dedication and commitment throughout the PhD degree program.

I am deeply grateful to my Daddy, Mommy, family and family friends for their love and support, and for always being there for me.

During this Doctoral degree program, so much changed in people's lives around me that words cannot express the amount of gratitude I have to the people in my life that kept me going and inspired me to keep me achieving my best.

I am extremely grateful to God for the abundant graces and blessings He has been showering on me.

List of Publications

I, Aimée Anne BOOYSENS, declare that the following are publications from this thesis:

1. **Booysens, A. and Viriri, S.**, Exploration of Ear Biometrics with Deep Learning, *Computer Vision and Graphics*, LNCS Springer, vol. 12334, pp. 25-35, (2020). DOI: `https://doi.org/10.1007/978-3-030-59006-2_3`

2. **Booysens, A. and Viriri, S.**, Exploration of Ear Biometrics Using EfficientNet, *Computational Intelligence and Neuroscience*, vol. 2022, (2022). DOI: `https://doi.org/10.1155/2022/3514807`

3. **Booysens, A. and Viriri, S.**, 2022. Ear biometrics using Deep Learning: a survey, *Applied Computational Intelligence and Soft Computing*, vol. 2022, (2022). DOI: `https://doi.org/10.1155/2022/9692690`

4. **Booysens, A. and Viriri, S.**, Transformation Network Model for Ear Recognition, *Machine Learning for Networking*, LNCS Springer, vol. 14525, pp. 250–266,(2024). DOI: `https://doi.org/10.1007/978-3-031-59933-0_17`

5. **Booysens, A. and Viriri, S.**, Lightweight Deep Learning with Model Compression for Ear Recognition, *Recent Challenges in Intelligent Information and Database Systems*, CCIS Springer. (under review - proof in Appendix A)

Contents

List of Figures

List of Tables

Abbreviations

ANN Artificial Neural Network

CNNs Convolutional Neural Networks

DCNNs Deep Convolutional Neural Networks

FCN Fully Convolutional Network

FCRN Fully Convolutional Residual Network

FN False Negative

FP False Positive

FPR False Positive Rate

K-NN K-Nearest Neighbor Algorithm

LBP Local Binary Pattern

NN Neural Networks

RBF Radial Basis Function

SVM Support Vector Machine

TPR True Positive Rate

TN True Negative

TP True Positive

U-Net U-shaped Network

VGGNet Visual Graphics Group Network

RESNET Residual Network

Chapter 1

General Introduction

1.1 Introduction

The ear develops in a foetus amid the fifth and seventh weeks of pregnancy, [3]. At this stage of the pregnancy, the face acquires a more distinguishable shape as the mouth, nostrils, and ears begin to form. There is still no exact timeline at which the outer ear is created during pregnancy, but it is accepted that a cluster of embryonic cells connect to establish the ear. These are called auricular hillocks, which begin growing in the lower portion of the neck. The auricular hillocks broaden and intertwine within the seventh week to deliver the ear's shape. Within the ninth week, the hillocks move to the ear canal and are more noticeable as the ear, [3]. The external anatomy of the ear can be seen in Figure 1.1. The growth of the ear in the first four months after birth is linear, and the ear is then stretched in development between the ages of four months and eight years. After this, the ear size and shape are constant until the age of seventy, increasing in size again.

Biometrics is the recognition of a human using their biometric characteristics, which may be physiological or behavioural. The physiological biometric features are the DNA, face, ear, facial, iris, fingerprint, hand geometry, hand vein, and palm print, with the behavioural biometrics being signatures, gait patterns, and keystrokes. Voice is considered a combination of biometric and physiological. Numerous systems have been developed to distinguish biometric traits, which have been used in several applications, from forensic investigations to security systems. With the COVID-19 pandemic, facial recognition systems failed due to users wearing masks;

however, human ear recognition proved more suitable as it is visible. In Table 1.1, the biometric characteristics and their ability to be unique are compared. The characteristics examined were distinctiveness, permanence, collectability, performance, and acceptability.

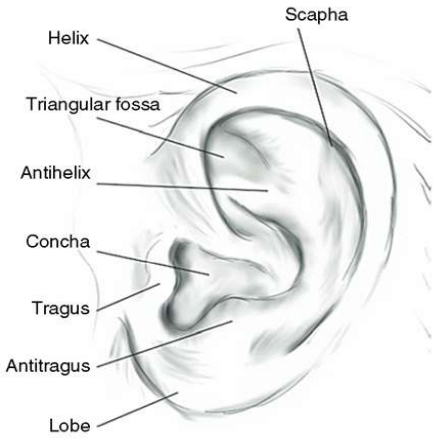


Figure 1.1: Diagram of the Outer Ear

In the different physiological biometric qualities, the ear has received much consideration of late as it tends to be said that it is a solid biometric for human acknowledgement, [19]. The ear biometrics framework is dependable as it does not change, has a uniform tone, and is fixed at the centre of the face's side. An individual's ear size is more critical than a unique finger impression. It simplifies capturing an image of the subject without needing to gain information from it, [19]. There are numerous difficulties in correctly gauging the details of the ear, including concealment of the ear by clothes, hair, ear ornaments, and jewellery. Another interference could be the different angle of the image, concealing essential characteristics of the ear's anatomy. These difficulties have made ear recognition a secondary role in identification systems and techniques commonly used for identification and verification.

Table 1.1: Summary of Biometric Characteristics

Biometric Identifier	**Biometric Type**	**Distinct-iveness**	**Perma-nence**	**Collect-ability**	**Perfor-mance**	**Accept-ability**
DNA	Physiological	High	High	Low	High	Low
Ear	Physiological	Medium	High	Medium	Medium	High
Face	Physiological	Low	Medium	High	Low	High
Facial	Physiological	High	Low	High	Medium	High
Fingerprint	Physiological	High	High	Medium	High	Medium
Gait	Behavioural	Low	Low	High	Low	High
Hand geometry	Physiological	Medium	Medium	High	Medium	Medium
Hand vein	Physiological	Medium	Medium	Medium	Medium	Medium
Iris	Physiological	High	High	Medium	High	Low
Keystroke	Behavioural	Low	Low	Medium	Low	Medium
Odour	Physiological	High	High	Low	Low	Medium
Palm print	Physiological	High	High	Medium	High	Medium
Retina	Physiological	High	Medium	Low	High	Low
Signature	Behavioural	Low	Low	High	Low	High
Voice	Combination of Physiological and Behavioural	Low	Low	Medium	Low	High

1.2 Motivation and Applications

This section discusses the motivation and application of the research work. The first part discusses the motivation, and the second explains the applications of ear identification.

1.2.1 Motivation

This thesis explores the efficiency of deep learning-based models for accurate ear biometrics recognition. Ear identification depends on specific criteria for grouping pixels into intensity values, gradient information, or textures. The domain of the applications is to determine the left and right ear using the facial region or 2D profile image.

The thesis will answer whether there could be room for improvement by using deep learning networks to determine both the left and right ears. It looks at whether there are state-of-the-art deep learning models that produce better results than currently used models. Otherwise, the current ear precision should be increased if the left and right ear cannot be determined.

1.2.2 Applications

There are several applications of ear recognition. These are surveillance systems, security, and general identity verification.

- **Surveillance Systems** - A human-computer interaction surveillance system can be built to identify human attributes such as gender, age, and ethnicity. There are many reasons that surveillance systems can be used, namely, terror-related crimes, law enforcement and security.
- **Security** - Ear identification could replace password logins on specific applications and computer systems.
- **General Identity Verification** - Ear identification using facial images has general uses, including electoral registration, banking, electronic commerce, identifying newborns, national IDs, passports and employee IDs.

1.3 Problem Statement

As stated before, biometrics is the recognition of a human using their biometric characteristics, which may be physiological or behavioural [44]. With the COVID-19 pandemic, facial identification has failed due to users wearing masks. However, the human ear has proven more suitable as it is visible. Several methods have been proposed in the literature for the ear identification of humans through facial and profile images [23, 68, 87, 122]. An example of the ear left and right image used for ear identification is shown below, Figure 1.2.

(a) Example ear extracted for a female

(b) Example ear extracted for a male

Figure 1.2: Examples of ear images

The literature discussed shows limited research regarding ear identification and detection. Most existing methods use machine learning algorithms, whose performance in correctly identifying the ear is inaccurate. The other issue observed is that most existing work only obtained the left or right ear [23, 68, 87, 122], not both. There are numerous difficulties in correctly gauging the details of the ear, including concealment of the ear by clothes, hair, ear ornaments, and jewellery. Another interference could be the different angle of the image, concealing essential characteristics of the ear's anatomy [23, 68, 87, 122]. These difficulties have made ear recognition a secondary role in identification systems used for identification and verification.

A deep learning model has faster identification rates because it iterates an extensive database with enough images for comparison. The threshold value and post-processing steps help eliminate false positives before detecting the ear. The images are analysed with those in the training set, and they won't have similar characteristics that may lead to incorrect identification.

1.4 Thesis Objectives

This thesis aims to design a model that accurately classifies profile and facial images to identify the ear by achieving the following objectives:

1. To conduct a critical survey of the state-of-the-art literature on ear identification in facial and profile images.
2. To model a deep learning-based framework for accurate ear biometrics identification.

3. To improve ear biometrics identification accuracy using enhanced deep learning techniques.

1.5 Contributions of the Thesis

The technical contributions of this research to the field of computer vision are summarised below:

- The thesis comprehensively reviews various traditional ear identification tests. It discusses a range of machine and deep learning techniques employed in biometric and ear identification, summarises prominent ear datasets and briefly discusses the steps involved in detection models from image pre-processing, feature extractions, and feature classifications. The thesis highlights state-of-the-art deep learning architectures, stating their strengths and weaknesses, evaluation metrics, and performance of the current ear identification classifiers.

- The thesis designed a robust image-processing algorithm that handles the pre-processing of images, such as contrast enhancement, feature extraction, and noise removal, before feeding it to the design machine learning and CNN model.

- Developed novel deep learning architectures and fine-tuned pre-trained CNN to identify ears and effectively classify them as the left or right ear.

- This thesis demonstrates that a model's performance accuracy and sensitivity can be improved through deep learning.

- Demonstrate that training a model on a dataset and testing it on a different dataset that does not originate from the training subset gives a better generalisation of the ear identification accuracy.

1.6 Organisation of work

The organisation of the thesis is as follows:

Chapter 2. The chapter presents a study of the state-of-the-art deep learning techniques for ear identification from images. It also discusses the popular convolutional neural network architectures used for ear identification and presents a critical analysis of the performance of some deep learning models. The chapter also presents a critical and comprehensive review of ear identification.

Chapter 3. The chapter describes the proposed ear identification methodology and presents the materials and methods used for the classification task.

Chapter 4. The chapter has the results and discussion, which covers the results of the various methods implemented in the previous chapters and compares those results to similar existing methods.

Chapter 5. In this Chapter, a conclusion with a general overview and contributions of the thesis is presented with a discussion of possible future work.

Chapter 2

Literature Review and Related Works

2.1 Ear Biometrics Using Deep Learning: A Survey

2.1.1 Brief Overview

This section presents a comparative analysis of previous research done for ear detection. The review paper discussed machine and deep learning architectures to detect ears and biometrics in images. It summarised prominent ear datasets and briefly discussed the steps involved in detection models from image pre-processing, feature extractions, and feature classifications. The thesis highlights state-of-the-art deep learning architectures, stating their strengths and weaknesses, evaluation metrics, and performance of the current ear classifiers.

The literature review is published in the *Applied Computational Intelligence and Soft Computing journal*.

Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2022, Article ID 9692690, 17 pages
https://doi.org/10.1155/2022/9692690

Review Article

Ear Biometrics Using Deep Learning: A Survey

Aimee Booysens and Serestina Viriri

School of Mathematics, Statistics & Computer Science, University of KwaZulu-Natal, Durban, South Africa

Correspondence should be addressed to Serestina Viriri; viriris@ukzn.ac.za

Received 17 February 2022; Accepted 11 July 2022; Published 17 August 2022

Academic Editor: Agostino Forestiero

This paper explores ear biometrics using a mixture of feature extraction techniques and classifies this feature vector using deep learning with convolutional neural network. This exploration of ear biometrics uses images from 2D facial profiles and facial images. The investigated feature techniques are Zernike Moments, local binary pattern, Gabor filter, and Haralick texture moments. The normalised feature vector is used to examine whether deep learning using convolutional neural network is better at identifying the ear than other commonly used machine learning techniques. The widely used machine learning techniques that were used to compare them are decision tree, naïve Bayes, K-nearest neighbors (KNN), and support vector machine (SVM). This paper proved that using a bag of feature techniques and the classification technique of deep learning using convolutional neural network was better than standard machine learning techniques. The result achieved by the deep learning using convolutional neural network was 92.00% average ear identification rate for both left and right ears.

1. Introduction

The ear begins to develop on a fetus amid the fifth and seventh weeks of pregnancy [1]. At this stage of the pregnancy, the face acquires a more distinguishable shape as the mouth, nostrils, and ears begin to form. There is still no exact timeline at which the outer ear is created during pregnancy, but it is accepted that a cluster of embryonic cells connect to establish the ear. These are called auricular hillocks, which begin to grow in the lower portion of the neck. The auricular hillocks broaden and intertwine within the seventh week to deliver the ear's shape. Within the ninth week, the hillocks move to the ear canal and are more noticeable as the ear [1]. The external anatomy of the ear can be seen in Figure 1. The growth of the ear in the first four months after birth is linear. The ear is then stretched in development between the ages of four months and eight years. After this, the ear size and shape are constant until the age of seventy, when they increase in size again.

Biometrics is the recognition of a human using their biometric characteristics, which may be physiological or behavioural. The physiological biometric features are the DNA, face, ear, facial, iris, fingerprint, hand geometry, hand vein, and palm print, with the behavioural biometrics being signatures, gait pattern, and keystrokes. Voice is considered as a combination of biometric and physiological. Numerous systems have been developed to distinguish biometric traits, which have been used in numerous applications such as forensic investigations and security systems. With the present worldwide pandemic, facial identification has failed due to users' wearing masks. However, the human ear has proven more suitable as it is visible. In Table 1, the characteristics that were looked at were the performance of the biometric if it is distinctive, permanence, ability to be collected, and acceptability.

In the different physiological biometric qualities, the ear has received much consideration of late as it tends to be said that it is a solid biometric for human acknowledgement [2]. The ear biometric framework is dependable as it does not change, it is of uniform tone, and its position is fixed at the centre of the face's side. The size of an individual's ear is more critical than a unique finger impression and makes it simpler to capture an image of the subject without necessarily needing to gain information from the subject [2]. There are numerous difficulties in correctly gauging the details of the ear. These are concealment of the ear by clothes,

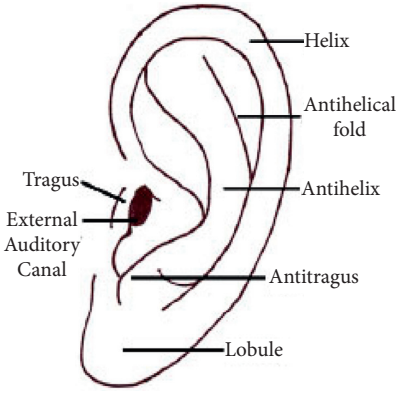


Figure 1: Diagram of the outer ear.

Table 1: Summary of biometric characteristics.

Biometric identifier	Biometric type	Distinctiveness	Permanence	Collectability	Performance	Acceptability
DNA	Physiological	High	High	Low	High	Low
Ear	Physiological	Medium	High	Medium	Medium	High
Face	Physiological	Low	Medium	High	Low	High
Facial	Physiological	High	Low	High	Medium	High
Fingerprint	Physiological	High	High	Medium	High	Medium
Gait	Behavioural	Low	Low	High	Low	High
Hand geometry	Physiological	Medium	Medium	High	Medium	Medium
Hand vein	Physiological	Medium	Medium	Medium	Medium	Medium
Iris	Physiological	High	High	Medium	High	Low
Keystroke	Behavioural	Low	Low	Medium	Low	Medium
Odor	Physiological	High	High	Low	Low	Medium
Palm print	Physiological	High	High	Medium	High	Medium
Retina	Physiological	High	Medium	Low	High	Low
Signature	Behavioural	Low	Low	High	Low	High
Voice	Combination of physiological and behavioural	Low	Low	Medium	Low	High

hair, ear ornaments, and jewellery. Another inference could be the different angle at which the image was taken, concealing essential characteristics of the ear's anatomy. These difficulties made ear recognition a secondary role in identification systems and techniques commonly used for identification and verification.

This paper's contributions are summarised below.

(1) A survey has been conducted with different deep learning architectures

(2) A study of the present ear bench-mark databases and their suitability for ear identification

(3) Different algorithms used for ear identification were outlined, highlighting the weaknesses and strengths

(4) A review of the present deep learning algorithms used for ear identification

The remainder of this work is organised as follows: Section 2 presents the foundation data on deep learning; Section 3 presents the vast majority of the ear information bases that are accessible for research; Section 4 presents a study of ear recognition calculations; the different profound learning strategies used to identify the ear are introduced in Section 5; and Section 6 presents the conclusion.

2. Review of Deep Learning

Deep learning is an AI model that utilises numerous layers to progressively understand the data. This paper will discuss the structures and contemporary strategies for deep learning designs in AI models that find the correct representation for the inputted information.

2.1. Neural Network (NN). A neural network (NN) is a type of machine learning algorithm that learns representations from data [3, 4]. A neutron may connect the processing unit from the directly linked network. Whenever there is a link, it has a weight that will be adjusted to assist the training process. The feed-forward neural network is when each neuron may be a function $f(x; \theta)$ which maps to an input, then to an output. The network learns the values of the parameters $\theta = w, b$, where w is a weight vector and b a scalar. This is often

performed through a backpropagation algorithm, as shown in the following equation:

$$F(x:\ \theta) = \sigma(w.x + b). \tag{1}$$

The first layer within the network is the input layer, and therefore, the last layer is the output layer. The middle layers within the algorithm are referred to as the hidden layers. When there are many hidden layers, this is often mentioned as a deep neural network; this is depicted in Figure 2.

2.2. Convolutional Neural Network (CNN). A convolutional neural network (CNN) is an NN that joins two or more layers together to produce one composite layer. The convolutional layer is able to learn features from the input data. By stacking many convolutional layers, the network is able to learn a hierarchy of increasingly complex features [3]. A pooling layer is usually added between successive convolutional layers to reinforce essential elements. In doing the CNN, it reduces the number of parameters that are passed to the lower layers. This is depicted in Figure 3.

2.3. Building Block for Convolutional Neural Networks

2.3.1. Convolutional Layer. This layer is a set of learnable filters or kernels used to slide over the entire input volume, performing a dot product between entries of the filter and the input layer [5]. The convolutional operation first extracts patches from its information in a sliding window fashion and then applies the same linear transformation to all the areas. The output of the convolutional operation is referred to as a feature map. The network will learn filters and then recognise the visual patterns that are in the input data. This is often shown asx_{ij}^{l}

$$x_{ij}^{l} = \sum_{a=0}^{m-1}\sum_{b=0}^{m-1} \omega_{ab} y_{(i+a)(j+b)}^{l-1}, \tag{2}$$

where x_{ij}^{l} is the computation of the input and is the sum of the contributions from the previous layer cells.

2.3.2. Pooling Layer. A pooling layer usually follows a single or multiple convolutional layers and is used to reduce the feature mapped dimensions keeping the essential elements [3]. A pooling layer is applied to a rectangular neighbourhood using a sliding window operation. Other pooling operations are maximum, depicted in Figure 4, average depicted in Figure 5, and weighted global pooling.

2.3.3. Nonlinearity Layer. The nonlinearity layer involves three steps. In step one, the layer performs the convolutional operation on the input feature map and produces a linear activation [3]. The second step would be to do the nonlinear transformation, and lastly, the pooling layer is used to modify the output. Nonlinear transformation can be carried out using activation functions; this gives the network the ability to learn a nontrivial representation, making the network resilient to slight modifications or noise in the input

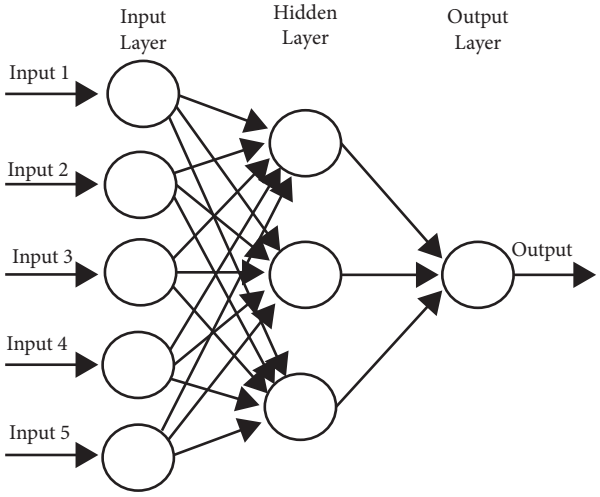


Figure 2: Diagram of neural networks.

data and improving the computational efficiency. This is often shown as$lY_i^{(l-1)}l-1$

$$Y_i^{(l)} = f\left(Y_i^{(l-1)}\right), \tag{3}$$

where l is the nonlinearity layer and the volume $Y_i^{(l-1)}$ is from the convolutional layer $l-1$.

2.3.4. Fully Connected Layer. The fully connected layer is used as a feature extractor. The features produced are then passed to the fully connected layers for classification. Each unit in the fully connected layer is connected to all the units in the previous layers. The last layer is usually a classifier that produces a probability map over the different classes. All the features are converted into one-dimensional feature vectors before passing into the fully connected layer. The reason that this is carried out is that spatial information in the image data is lost, has a high computational cost, and can only work with images that are of the same size [6]. This is often shown as

$$\begin{aligned} y_i^{(l)} &= f\left(z_i^{(l)}\right) \text{with } z_i^{(l)} \\ &= \sum_{j=1}^{m_1^{l-1}} w_{i,j}^{(l)} y_i^{(l-1)}. \end{aligned} \tag{4}$$

2.3.5. Optimisation. The performance of the deep CNN can be improved by training the network on a large data set. Training involves looking for the parameter of the model that reduces the cost function [3]. Gradient descent, shown in equation (5), is a widely used method for updating the network parameters through the backpropagation algorithm. The optimisation can be carried out at any stage in the process.

$$\Theta = \Theta - \alpha \cdot \nabla J(\Theta). \tag{5}$$

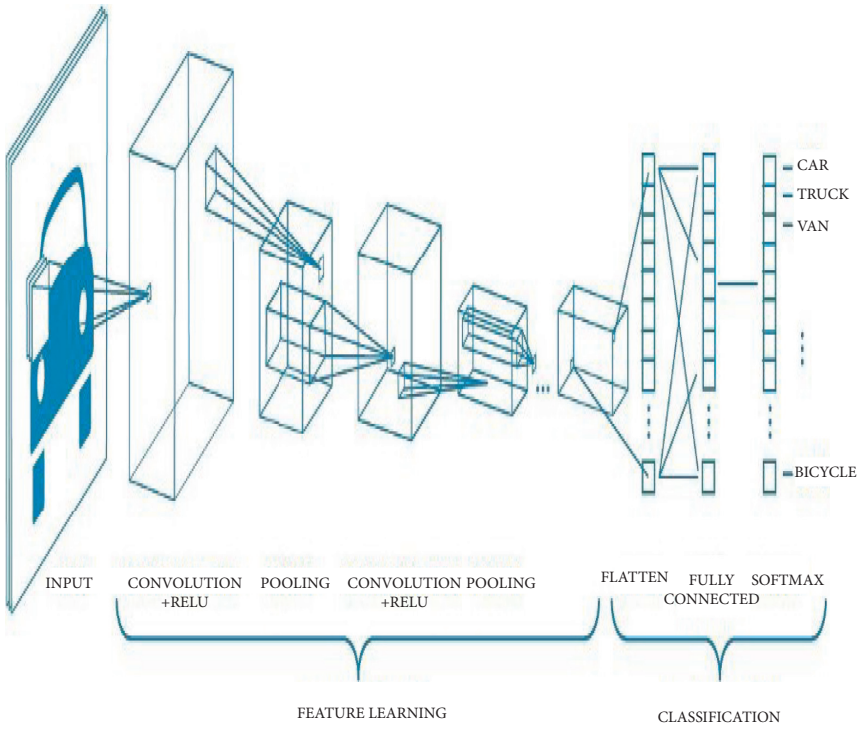


Figure 3: Diagram of convolutional neural networks.

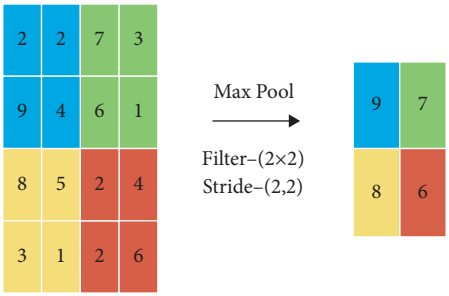


Figure 4: Diagram of maximum pooling layer.

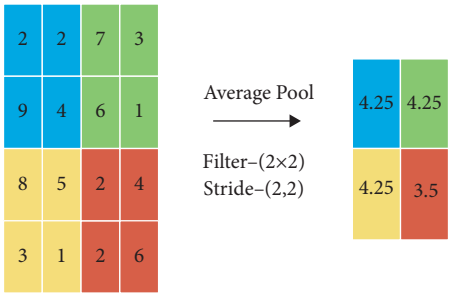


Figure 5: Diagram of average pooling layer.

2.3.6. Loss Function. Loss function is used in machine learning to evaluate how the specific algorithm model obtains data. The main goal of training an NN is to make sure that the loss is low. When the output is far from the actual value, the loss will be high and low when the prediction is close to the actual value [3]. The loss function used is mean-squared error, which is calculated by taking the mean of squared differences between actual and predicted values, and the binary cross entropy takes the output node to classify the data into two classes which are passed through a sigmoid function with an output of 0 or 1.

2.3.7. Parameter Initialisation. Parameter initialisation is a deep learning optimisation algorithm that is iterative and requires the user to state a starting point for the algorithm. The point at which the user chooses influences how fast learning can converge [3].

2.3.8. Hyperparameter Tuning. Hyperparameter tuning is the parameter that the user supplies to control the algorithm's behaviour before training starts, and this can be the learning rate, batch size, or image size [3].

2.3.9. Regularisation. Regularisation is a technique for improving the performance of machine learning algorithms on unseen data [3]. Regularisation is carried out to reduce the overfitting of the training set, and this happens when the gap between the training and test error is too large.

2.4. Deep Convolutional Neural Network Architectures

2.4.1. Single Pathway. A single pathway may be a primary network that resembles a feed-forward deep neural network [7]. Using one path, the data moves from the input layer to the classification layer. Kleesiek et al. [8] proposed a 3D

single-path CNN that has fully connected convolutional layers: the classification layer, which allows the network to classify multiple 3D pixels on just one occasion.

2.4.2. Cascaded Architecture. In the cascaded architecture, the output of the CNN is concatenated with another [9]. There are many variations with this architecture within the literature, but the input cascade is prominent. In this architecture, the output of the CNN becomes a direct input of another CNN. The input cascade is employed to concatenate the contextual information to the second CNN as additional image channels. Cascaded architecture is an improvement to the only pathway that performs multiscale label prediction separately. There are many other cascaded architectures: local pathway concatenation and hierarchical segmentation.

2.4.3. UNET. UNET improves a convolutional network that resembles an encoder and decoder network designed to do biomedical image segmentation [10]. The network consists of a contracting path and an expansive path, which provides it with the u-shaped architecture. The contracting path consists of the repeated application of two convolutional layers, followed by a rectified linear measure and a top pooling layer that goes along the trail to scale back the spatial information while feature information is increased. The expansive path consists of upsampling operations combined with high-resolution features from the contraction path through skip connections.

2.4.4. AlexNet Architecture. AlexNet architecture is an easy but powerful CNN architecture consisting of convolutional and pooling layers [11]. These layers are fully connected at the highest point, and the benefits of the AlexNet include the size with which it uses the GPU for training and performing the task. This architecture remains a starting point in applying deep neural networks, specifically for computer vision and speech recognition.

2.4.5. Visual Geometry Group Architecture. Visual geometry group architecture is a network created by Visual Graphics Group researchers at Oxford University [12]. It is characterised by a pyramidal shape because it comprises a group of convolutional layers followed by pooling layers; these pooling layers make the layers narrower in shape. The benefits include keeping a good architecture used for benchmarking for any task. The pretrained networks of the VGG are also primarily used for different applications but require numerous computational resources and are slow to coach, above all when training the dataset from scratch.

2.4.6. GoogLeNet Architecture. The GoogLeNet architecture is referred to as the inception network and was created by Google researchers [13]. It is made from twenty-two layers with two options that these layers can either convolute or pool the input. The architecture contains many beginning modules stacked over each other, allowing joint and parallel training, which helps with faster convergence. The benefits are that there is speedier training, which reduces the size. It , however, possesses an Xception network, which could increase the point for the divergence of the beginning module.

2.4.7. Residual Network (ResNet) Architecture. The residual network (ResNet) architecture is a 152-layer deep CNN architecture of the residual blocks. This is more profound than that of the AlexNet and VGG architectures as it is less computationally complex than these networks. It is referred to as a residual network [14], which is made up of numerous succeeding residual modules that are the essential building blocks of the architecture. These modules are stacked to produce an end-to-end network. The advantage of this architecture is that performance is improved due to its many residual layers and it is used for network training.

2.4.8. ResNeXt Architecture. ResNeXt architecture is the present state-of-the-art technique for visual perception, which is a hybridisation between inception and ResNeXt architectures [15]. ResNeXt is referred to as the aggregated residual transform network, but it is an improvement over the inception network. It splits the concept and transforms and merges in a commanding but easy way by bringing in cardinality. It uses residual learning, which will enhance the joining of the deep and wide networks. ResNeXt uses many transformations within a split, transform, and merge blocks; and the transformations in cardinality define these. ResNeXt used a mixture of VGG topology and GoogLeNet architecture to correct the spatial resolution using 3×3 filters within the split, transform, and merge blocks. The increase in cardinality improves the performance and produces a different and improved architecture.

2.4.9. Advance Inception Network. The advance inception network includes Inception-V3, Inception-V4, and Inception-ResNet. This is often an improved version of Inception-V1, Inception-V2, and GoogLeNet [16]. Inception-V3 reduces the computational cost of deep networks but does not affect generalisation. Szegedy et al. [17] replaced large-sized filters (5×5 and 7×7) with small and unequal filters (1×7 and 1×5) and used 1×1 convolution as a blockage before the vast filters. Inception-ResNet combines the strength of the residual learning and starting block.

2.4.10. DenseNet Architecture. The DenseNet architecture [16] is similar to ResNet but was created to fix the vanishing gradient problem. DenseNet utilises cross-layer connectivity by connecting each preceding layer to the next layer in a feed-forward manner. This was carried out to fix the ResNet by preserving identity transformations, which increased complexity. As it uses solid blocks, it allows to feature maps of all previous layers to be used as the inputs into the subsequent layers.

2.4.11. SqueezeNet Architecture. Hu et al. [18] proposed an auxiliary block for the choice to feature maps for object discrimination. The new block named SE-block overpowers the smaller feature maps and stimulates the category feature maps. It was created to be added into any CNN architecture before the convolution layer. It has two primary operations: squeeze and convolution. The convolution kernel captures local information but ignores features' contextual relations, while the squeeze operation captures global information of the feature maps. The network generates a feature map that is a more robust architecture and is helpful when there is low bandwidth.

2.4.12. Xception Architecture. Xception architecture is referred to as risky inception architecture that overdoes depthwise separable convolution [19]. The first inception block is modified by making it more complete and substituting different spatial dimensions (1×1, 5×5, and 3×3) with one dimension (3×3) followed by a 1×1 convolution to achieve computational complexity. It makes the network computationally efficient by uncoupling spatial and feature map channels.

2.4.13. Deep Reinforcement Learning. Deep reinforcement learning [20] may be a system trained entirely from scratch, ranging from random behaviour to an accurate knowledge domain from experience. It is a mixture of reinforcement and deep learning using fewer computation resources and data. The algorithm can learn from its environment and apply it to any sequential decision-making problems, including image analysis.

2.4.14. Fully Convolutional Network. A fully convolutional network [21] is a set of convolutional and pooling layers. Bi et al. [22] developed a multistage fully convolutional network with the parallel integration method for segmentation.

2.4.15. Deep Residual Network. Deep residual network [23] may be a particular sort of artificial neural network that builds on a pyramidal structure by utilising skip connections that skip some convolutional layers. It is composed mainly of multiple convolutional layers.

2.4.16. Convolutional and Deconvolutional Neural Networks. This architecture is formed from two significant parts: convolutional and deconvolutional networks [24]. Deconvolutional networks are CNNs that operate during a reversed process, and networks extract discriminated features. The deconvolutional layers are applied for smothering the segmentation maps to get the ultimate high-resolution output.

2.4.17. Residual Attention Neural. Zhou et al. [25] designed residual attention neural that improves CNNs feature representation by incorporating attention modules into CNN and forms a network capable of learning object-aware features. It employs a feed-forward CNN that stacks residual blocks with an attention module. It combines two different learning strategies into the eye module that permits fast feed-forward processing and top-down attention feedback during a single feed-forward process to supply dense features that infer each pixel. The bottom-up feed-forward structure produces low-resolution feature maps with reliable semantic information. The top-down learning strategy globally optimises the network such that it gradually outputs the maps to input during the training process. Table 2 shows a summary of the deep convolutional neural network architecture used for ear identification.

3. Overview of the Ear Dataset

Many factors can affect an ear detection system's performance. The ear images' datasets are easier to use than others. The more ear datasets are for researchers to use, the more this field can evolve and grow. It is always good to use high-quality images in research associated with soft biometrics. A brief description of a number of the available ear databases is highlighted in Table 3 and examples of images are shown in Figures 6 and 7.

3.1. Mathematical Analysis of Images (AMI) Ear Database. The AMI ear database was collected at the University of Las Palmas. The database comprises 700 ear images of 100 distinct Caucasian adult males and females between 19 and 65 years of age. All images within the database were taken under equivalent illumination and with a glued camera position. Both the left- and right-hand sides of the ears were captured. The pictures obtained are cropped to form the ear area, covering almost half of the image. The pose of the themes varies in yaw and surveying in pitch angles, and datasets are often found publicly.

3.2. The Indian Institute of Technology (IIT) Delhi Ear Database. The IIT database [26] was collected by the Indian Institute of Technology Delhi in New Delhi between October 2006 and June 2007. The database is formed from 421 images of 121 distinct adults of both males and females. All images were taken inside the environment, with no significant occlusions present, and only the right-hand side of the ear was captured. The pictures obtained in the dataset were both raw and normalised. The normalised images were in grey-scale with a size of 272×204 pixels.

3.3. The University of Beira Ear (UBEAR) Database. The University of Beira presented the UBEAR database [27]. The database comprises 4429 images of 126 subjects, and these were of both males and females. The images were taken under varying lighting conditions and angles, and partial occlusions were present. These images are of the ear, both the left- and right-hand side ear images were provided.

3.4. The Annotated Web Ear (AWE) Database. The AWE ear database [28] was a set of public figures from web images. The database was formed from 1000 images of 100

Table 2: Summary of the deep convolutional neural network architecture used for ear identification.

Deep convolutional neural network architecture	Summary of deep convolutional neural network used in ear identification	Accuracy (%)
AlexNet [11]	AlexNet is seen as a deep convolutional neural network architecture and applied to numerous ear recognition systems	53.6
DenseNet [16]	DenseNet connects each layer in the CNN to another and applied to ear image datasets, yielding positive results	62.0
ResNet [14]	ResNet is a class of extremely deep CNN architecture that addresses vanishing gradient by using skip connections that prevent information loss as the network goes deeper. As ResNet addressed the vanishing gradient issue, it has been applied to numerous ear image datasets yielding positive results	15.0
ResNeXt [15]	ResNeXt is a modularised CNN architecture, which has been applied to ear image datasets yielding positive results	95.8
Visual geometry group [12]	The visual geometry group is a very deep CNN and is one of the top performers. The VGG is used in recognition systems and has been applied to unconstrained ear image datasets, yielding positive results.	83.0

different subjects, whose sizes varied and were tightly cropped. Both the left- and right-hand sides of the ears were taken.

3.5. EarVN1.0. The EarVN1.0 database [29] comprises 28412 images of 164 Asian male and female subjects, and left- and right-hand sides of the ear were captured. It was collected during 2018 and is formed from unconstrained conditions, including camera systems and lighting conditions. The pictures are cropped from facial images to obtain the ears, and the pictures have significant variations in pose, scale, and illumination.

3.6. The Western Pomeranian University of Technology Ear (WPUTE) Database. The Western Pomeranian University of Technology Ear (WPUTE) database [32] was obtained in the year 2010 to gauge the ear recognition performance for images obtained in the wild. The database contains 2071 ear images belonging to 501 subjects. The images were of various sizes and held both the left- and right-hand sides of the ear and were taken under different indoor lighting conditions and rotations. There were some occlusions included in the database. These were the headset, earrings, and hearing aids.

3.7. The Unconstrained Ear Recognition Challenge (UERC). The Unconstrained Ear Recognition Challenge (UERC) database [14] was obtained in 2017, then extended in 2019, and is a mix of two databases that currently exist and a newly created one. The database contains 3706 subjects with 11804 ear images, and the database ears have both right- and left-hand side images.

3.8. In the Wild Ear (ITWE) Database. The In the Wild Ear (ITWE) database [33] was created for recognition evaluation and has 2058 total images, including 231 male and female subjects. A boundary box obtained these images of the ear. The coordinates of those boundary boxes were released with the gathering. The pictures contained cluttered backgrounds and were of variable size and determination. The database includes both the left- and right-hand sides of the ear, but no differentiation was given about the ears.

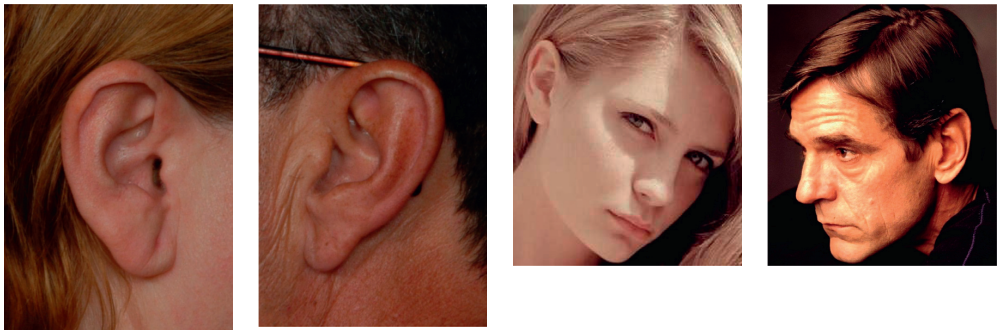
3.9. The University of Science and Technology, Beijing (USTB) Ear Database. The University of Science and Technology Beijing (USTB) Ear Database [30] contained cropped ear and head profile images of male and female subjects split into four sets. Dataset one includes 60 subjects and has 180 images of right-close-up ears during 2002. These images were taken under different lighting, experiencing some shearing and rotation. Dataset two contains 77 subjects and has 308 images of the right-hand side ear, approximately 2 m away from the ear, and the images were taken in 2004. These images were taken under different lighting conditions. Dataset three contains 103 subjects and has 1600 images. These images were taken during the year 2004. The images are on the proper and left rotation, and therefore, the images are of the dimensions 768×576. The dataset contains 25500 images of 500 subjects; these were obtained from 2007 to 2008; the subject was in the centre of the camera circle. The images were taken when the subject looked upwards, downwards, and at eye level. The images in this dataset contained different yaw and pitch poses. The databases are available on request and accessible for research.

3.10. The Carreira-Perpinan (CP) Ear Database. The Carreira-Perpinan (CP) [34] ear database is an early dataset of the ear utilised for ear recognition systems. It was created in 1995 and contained 102 images with 17 subjects. The images were captured in a controlled environment, and therefore, the images include variability in minor pose variation.

3.11. The Indian Institute of Technology, Kanpur (IITK) Ear Database. The Indian Institute of Technology Kanpur (IITK) is an ear database [35] that the Institute of Technology of Kanpur compiled. The database is split into three sets, the first set consists of 190 male and female subjects of profile images. The total number of images was 801. The second dataset also contained 801 total of 89 subjects, and

Table 3: Summary of datasets.

	Database	Year	Number of subjects	Number of images	Left ear count	Right ear count	Total ears	Image size	Country	Side
1	Institute of Technology Delhi Ear Database (IIT Delhi-I) [26]	2007	121	471		471	471	272 × 204	India	Right
	Institute of Technology Delhi Ear Database (IIT Delhi-II) [26]	NA	221	793		793	793	272 × 204	India	Right
2	The University of Science & Technology Beijing (USTB Ear I) [30]	2002	60	185		185	185	Varied	China	Right
	The University of Science & Technology Beijing (USTB Ear II) [30]	2004	77	308		308	308	Varied	China	Right
3	The Annotated Web Ears (AWE) database [28]	2016	100	1000	500	500	1000	Varied	Slovenia	Both
	The Annotated Web Ears database extended (AWE extend) [28]	2017	346	4104	2052	2052	4104	Varied	Slovenia	Both
4	Mathematical Analysis of Images Ear database (AMI) [31]	NA	106	700	420	280	700	492 × 702	Spain	Both
5	The West Pomeranian University of Technology Ear (WPUTE) database [32]	2010	501	2071	829	1242	2071	Varied	Poland	Both
6	Unconstrained Ear Recognition Challenge (UERC) database [14]	2017	3706	11804	5902	5902	11804	Varied	Slovenia	Both
7	EarVN1.0 [29]	2018	164	28412	14206	14206	28412	Varied and low-resolution	Vietnam	Both
8	The In-the Wild Ear (ITWE) database [33]	2015	55	605	424	181	605	Varied	Slovenia	Both
9	The Carreira-Perpinan (CP) [34]	1995	17	102	102		102	Varied	NA	Left
10	The University of Beira Ear (UBEAR) database [27]	2011	126	4430	2215	2215	4430	1280 × 960	Mozambique	Both
11	Indian Institute of Technology Kanpur (IITK) [35]	2011	801	190	95	95	190	Varied	India	Both
12	The forensic ear identification database (FEARID) [36]	2005	1229	1229	615	614	1229	Varied	United Kingdom, Italy and Netherlands	Both
13	University of Notre Dame (UND) [37]	2006	3480	952	952		952	Varied	France	Left
14	The Face Recognition Technology database (FERET) [38]	2010	9427	4745	3796	949	4745	Varied	Spain	Both
15	The Pose, Illumination, and Expression (PIE) [39]	2002	40000	68	34	34	68	Varied	USA	Both
16	The XM2VTS Ear Database [40]	NA	2360	295	89	206	295	720 × 576	UK	Both
17	The West Virginia University (WVU) [41]	2006	460	402	402		402	Varied	USA	Left



(a) (b) (c) (d)

Figure 6: Examples of original ear images. (a) Example of a 2D profile image of a female. (b) Example of a 2D profile image of a male. (c) Example of a facial image of a female. (d) Example of a facial image of a male.

(a) (b) (c) (d)

Figure 7: Examples of the extracted ear images. (a) Example ear extracted from 2D profile image of a female. (b) Example ear extracted from 2D profile image of a male. (c) Example ear extracted from facial image of a female. (d) Example ear extracted from facial image of a male.

these images had variations in pitch angle. The third dataset contains 1070 images of an equivalent of 89 subjects, but with a variation in yaw and angle.

3.12. The Forensic Ear Identification Database (FEARID). The Forensic Ear Identification Database (FEARID)[36] is different from other databases as it contains the ear prints. These contain no occlusions, variable angles, or illumination. Though there is no mention of any variables, other influences like the force the ear was pressed against the scanner and the scanner's cleanliness need to be considered. This database comprised 7364 images of 1229 subjects. This database was used for forensic application and not for biometric use.

3.13. The University of Notre Dame (UND) Database. The University of Notre Dame (UND) database contains [37] many subsets of 2D and 3D ear images. These images were appropriated for a period from 2003 to 2005. The database contains 3480 3D images from 952 male and female subjects and 464 2D images from 114 male and female subjects. These images were taken in different lighting conditions, yaw, pitch poses, and angles. The images are only of the left-hand side ear.

3.14. The Face Recognition Technology (FERET) Database. The Face Recognition Technology (FERET) database [38] is a sizeable facial image database and was obtained between the years 1995 and 1996. It contains 1564 subjects and has a total of 14126 images. These images were collected for face recognition and were of the left- and right-hand profile images, which made them perfect for 2D ear recognition.

3.15. The Pose, Illumination, and Expression (PIE). Carnegie Mellon University obtained the Pose, Illumination, and Expression database [39], which contains 40000 images and 68 subjects. The images are of the facial profile and have different poses, illuminations, and expressions.

3.16. The XM2VTS Ear Database. The XM2VTS ear database [40] is frontal and profiles face images from the University of Surrey; the database contains 295 subjects and 2360 images

captured during controlled conditions. These images were a set of cropped images of 720×576 size and were from video data.

3.17. The West Virginia University (WVU) Ear Database. The West Virginia University (WVU) Ear database [41] is a video database and is formed from 137 subjects. The system was an advanced capturing procedure that allowed them to capture the ear at different angles; these images included earrings and eyeglasses.

3.18. Summary. UBEAR, EarVN1.0, IIT, ITWE, and AWE databases are best suited for the ear identification due to their large data size. However, it shows that EarVN1.0 has the foremost prominent usage during age estimation using CNN techniques. It is an appropriate dataset where the ear images are taken in a controlled environment, while ITWE is compatible for classifying the ears in an uncontrolled environment, and examples of the extracted ears are shown in Figures 6 and 7.

4. Description of Ear Algorithms

This section presents different algorithms and techniques used for ear identification. It presents a description of these algorithms and suggests the most effective approach. A brief description of ear algorithms is highlighted in Table 4.

Ansari and Gupta [42] used outer helix curves of the ears as they moved parallel to at least one feature spot in the ear image. Helix curves were obtained using the Canny edge detector to remove the ear from the entire image. The obtained sides are then separated into a convex or concave edge, allowing the system to determine the helix edges. This technique was run on 700 side-ear images and had an accuracy of roughly 93%.

Abdel-Mottaleb and Zhou [43] segmented the ear from a facial profile image using supported template matching, where they modelled the ear by its external curve. Yuizono et al. [44] also used a template matching technique for detection, in which they used both hierarchical 2D images. In 3D ear detection, Chen and Bhanu [45] used a model-based (template matching) technique for ear detection. An averaged histogram of the shape index represents the model template. The detection is a four-step process: edge detection and threshold, image dilation, connected component labelling, and template matching. A test set of 30 subjects from the UCR database achieved a 91.5% detection rate with a 2.52% warming rate. Later, Chen and Bhanu [45] developed another shape-model-based technique for locating human ears inside face range images, where the ear shape model is represented by a group of discrete 3D vertices like the helix and antihelix parts. They started by locating the sting segments and grouping them into different clusters that are potential ear candidates. Arbab-Zavar and Nixon [46] developed an ear recognition system based on the ear's elliptical shape, employing a Hough transformation (HT). They achieved a 100% detection rate using the XM2VTS face profile database, consisting of 252 images from 63 subjects, and 91% using the UND, collection F, database.

Burge and Burger [47] have proposed a way to do ear recognition using geometric information about the ear. The ear has been represented by employing a neighbourhood graph obtained from a Voronoi diagram of the ear edge segments, whereas template comparison has been performed using subgraph matching. Choras [48] has used the ear's geometric properties to propose an ear recognition technique during which feature extraction is administered in two steps. In the initial step, global features are extracted. The second step extracts local features while matching local features. In another geometry-based technique proposed by Shailaja and Gupta [49], an ear is represented by two sets of features, global and native, obtained using outer and internal ear edges, respectively. Two ears during this technique are declared similar if they are matched to the feature sets. The method proposed has treated the ear as a planar surface and has created a homograph transform using SIFT feature points to register ears accurately. It has achieved robust results in background clutter, viewing angle, and occlusion. Cummings et al. [50] used the image ray transformation, based upon an analogy to light rays, to detect an image's ears. This transformation can highlight tubular structures like the helix of the ear and spectacle frames. By exploiting the elliptical shape of the helix, this method segmented the ear into regions and achieved a detection rate of 99.6% using the XM2VTS database.

Chen and Bhanu [45] fused complexion from colour images and edges from a range of images to perform ear detection. The images observed that the sting magnitude is more prominent around the helix and, therefore, the antihelix parts. They clustered the resulting edge segments and deleted the short irrelevant edges. Using the UCR database, they reported an accurate detection rate of 99.3% (896 out of 902). The UND databases (collections *F* and a subset of *G*) reported an accurate detection rate of 87.71% (614 out of 700). Hajsaid et al. [51] addressed the matter of an automated ear segmentation scheme by employing morphological operators. They used low computational cost appearance-based features for segmentation and a learning-based Bayesian classifier to determine whether the segmentation's output was incorrect. They achieved a 90% accuracy on 3750 facial images with 376 subjects within the WVU database.

Prakash and Gupta [52] used complexion and template-based techniques for automatic ear detection during a side profile face image. The technique first separates skin regions from nonskin regions and then searches for the ear within the skin regions employing a template matching approach. Finally, the ear region is validated using a moment-based shape descriptor. Experimentation on an assembled database of 150 side-profile face images yielded an accuracy of 94% . Basrur et al. [53] introduced the notion of "jet space similarity" for ear detection, which denotes the similarity between Gabor jets and reconstructed jets obtained via principal component analysis (PCA). They used the XM2VTS database for evaluation; however, they did not report their algorithm's accuracy.

Rahman et al. [54] used a cascaded AdaBoost technique, supported by Haar features for ear detection. This system is

Table 4: Summary of the ear algorithms.

Author	Algorithms used	Accuracy (%)	Summary
Ansari and Gupta [42]	Canny edge detector	93	Uses outer helix curves of the ears with Canny edge detector, and this only obtains the edges of the ear and is only used to determine the helix
Abdel-Mottaleb and Zhou [43]	Template matching	91.5	They used a segmented ear obtained from a facial profile and only modelled the ear's external curve
Arbab-Zavar and Nixon [46]	Hough transform	91	They only looked at the ear's elliptical shape, and they used a small sample of profile ears
Burge and Burger [47]	Geometric information	94	They did ear recognition using geometric information of the ear and used neighbourhood graphs obtained from a Voronoi diagram of the ear edge segments
Cummings et al. [50]	Image ray transform	99.6	Used ray transformation to detect an image of the ear and only obtained the helix of the ear and spectacle frames
Chen and Bhanu [45]	Fused complexion from colour images and edges from a range of images	87.71	Fused complexion from colour images and edges from a range of images to perform ear detection
Prakash and Gupta [52]	Complexion and template-based technique	94	Used complexions and template-based techniques for automatic ear detection
Basrur et al. [53]	Gabor jets and reconstructed jets obtained via principal component analysis	NA	Introduced the notion of "jet space similarity," but did not report their algorithm's accuracy
Rahman et al. [54]	Cascaded AdaBoost technique supported Haar features	100	This system is widely known within the domain of face detection because of the Viola–Jones method, and it is a speedy and comparatively robust face detection technique
Chang et al. [55]	Multimodal recognition system	90.9	This system supported both the face and ear recognition
Naseem et al. [56]	General classification algorithm	98	This system investigated two crucial issues: feature extraction and robustness to occlusion
Nanni and Lumini [57]	Multi-matcher-based technique	NA	This system considers overlapping subwindows to extract local features
Yan and Bowyer [58]	Contour extraction algorithm	21	This system only used the ear contour using the active outline
Minaee et al. [59]	Independent component analysis and a radial basis function	94.11	The original ear image database and decomposing it into linear combinations of many basic images
Abdel-Mottaleb and Zhou [43]	Support vector machine	100	This approach is used for 3D ear detection and then a sliding window approach and linear SVM classifier to identify the ear

widely known within the domain of face detection because of the Viola–Jones method. It is a speedy and comparatively robust face detection technique. They trained the AdaBoost classifier to detect the ear region even in the presence of occlusions and degradation in image quality. They reported a 100% detection performance on the cascaded detector tested against 203 profile images from the UND database, with a false detection rate of 5×10. A second experiment detected 54 ears out of 104 partially occluded images from the XM2VTS database.

Chang et al. [55] built a multimodal recognition system that supported face and ear recognition. The manually identified coordinates of the triangular fossa and the anti-tragus are used for ear detection for the ear images. Their ear recognition system was supported by Eigen-ears' concept, using principal component analysis (PCA). They reported performance of 72.7% for the ear in one experiment, compared to 90.9% for the multimodal system, using 114 subjects from the UND, collection *E*, database.

Naseem et al. [56] proposed a general classification algorithm for (image-based) visual perception, supported by a sparse representation computed by L1 minimisation. This framework provides new insights into ear recognition's two crucial issues: feature extraction and robustness to occlusion. The ear portion is manually cropped from each image, and no normalisation of the ear region is required. They conducted several experiments using the UND and USTB databases with session variability, various head rotations, and different lighting conditions. These experiments yielded a high recognition rate within the order of 98%.

Nanni and Lumini [57] have proposed a multi-matcher-based technique for ear recognition that obtains the ear's appearance-based local properties. It considers overlapping subwindows to extract local features using Gabor filters. Further, Laplacian Eigen Maps are accustomed to reduce the feature vectors' dimensionality. The ear is represented using the features obtained from a group of the most discriminative subwindows selected using the sequential forward floating selection (SFFS) algorithm. Matching during this technique is performed by combining the outputs of several 1-nearest neighbour classifiers constructed on different subwindows. Another technique that supports the fusion of colour spaces is proposed by Nanni and Lumini, where few colour spaces are selected using the SFFS algorithm, and Gabor features are extracted from them. Matching is

Table 5: Summary of the ear algorithms using CNN.

Author	Dataset	Accuracy	Summary
Emeršič et al. [60]	NA	30	It used handcrafted feature extraction methods such as LBP, POEM, and CNN to obtain the ear identification
Tian et al. [21]	AMI, WPUT, IITD, and UERC	70.58, 67.01, 81.98, and 57.75	This system used deep CNN to perform ear recognition. There were occlusions like no earrings, headsets, or similar occlusions
Raveane et al. [64]	NA	98	This system used variable conditions due to the odd shape human ear and changing lighting conditions
Zhang and Mu [65]	UND and UBEAR	100 and 98.22	This system contained large occlusions, scale, and pose variation
Kohlakala and Coetzer [66]	AMI and IIT-Delhi	99.2 and 96.06	It is used to classify ears either in the foreground or background of the image. The binary contour image applied the matching for feature extraction, and this was performed by implementing Euclidean distance measure, which had a ranking to verify for authentication
Tomczyk and Szczepaniak [67]	NA	NA	It shows the published experimental results that the approach did the rotation equivalence property to detect rotated structures
Alshazly et al. [68]	Three ear datasets but not stated	22	The paper took seven performing handcrafted descriptors to extract the discriminating ear image. Then took the extracted ear and trained it using SVM to learn a suitable model
Alkababji and Mohammed [69]	NA	97.8	It used the PCA and a genetic algorithm for feature reduction and selection
Jamil et al. [70]	Very underexposed or overexposed database	97	This work was the first to test the performance of CNN on very underexposed or overexposed images
Hansley et al. [71]	UERC challenge	NA	This was performed using handcrafted descriptors, which were fused to improve recognition

administered by combining several nearest neighbour classifiers constructed on different colour components.

Yan and Bowyer [58] developed an automatic ear contour extraction algorithm. This was carried out by detecting the ear pit based on the position of the nose and cutting the ear contour using the active outline starting around the ear tip. This paper's results showed that 21% of the images tested were incorrectly segmented, but if they changed it to use only depth information and not colour, only 15% of the images were incorrectly segmented. A hybrid system for ear recognition was investigated by Minaee et al. [59]. This system combines an independent component analysis (ICA) and a radial basis function (RBF) network. This was conducted by taking the original ear image database and decomposing it into linear combinations of many basic images. Then, the corresponding coefficients of these combinations are used in the RBF network. They achieved 94.11% using two databases of segmented ear images.

A 3D ear detection system was investigated by Abdel-Mottaleb and Zhou [43]. They showed a novel shape-based feature set called histograms of categorised shapes (HCS). This approach is used for 3D ear detection and then a sliding window approach and linear support vector machine (SVM) classifier to identify the ear. They reported a perfect detection rate, a 100% detection rate, and a 0% false-positive rate.

5. Review of Ear Algorithms Using CNN

This section presents different algorithms using CNN used for ear recognition. This paper presents a description of these algorithms and suggests the most effective approach. A brief description of the ear algorithms using CNN is highlighted in Table 5.

Emeršič et al. [60] organized the dataset of the UERC. It was introduced and used for the benchmark, training, and testing sets. In this study, it was seen that handcrafted feature extraction methods such as linear binary pattern (LBP) [61], patterns of oriented edge magnitudes (POEM) [62], and CNN-based feature extraction methods were used to obtain the ear identification. In this challenge, one method needs to figure out a way to remove occlusions like earrings, hair, other obstacles, and background from the ear image. The occlusion was carried out by creating a binary ear mask, and then the system recognition was conducted using the handcrafted features. Another proposed approach was to calculate the score of matrices from the CNN-based features and handcrafted features when they are fused. A 30% detection rate was produced.

Tian et al. [21] applied a deep convolutional neural network (CNN) to ear recognition in which they designed a CNN—it was made up of three convolutional layers, a fully connected layer, and a softmax classifier. The database used was USTB ear, which consisted of 79 subjects with various pose angles. There were occlusions like no earrings, headsets, or similar occlusions. Chowdhury et al. [63] proposed an ear biometric recognition system that uses local features of the ear and then uses a neural network to identify the ear. The method estimates where the ear could be in the input image and then gets the edge features from the identified ear. After identifying the ear, a neural network matches the extracted feature with a feature database. The databases used in this system were AMI, WPUT, IITD, and UERC, which achieved an accuracy of 70.58%, 67.01%, 81.98%, and 57.75%, respectively.

Table 6: Sources of the articles reviewed.

S no.	Article source	Quantity
Conference		
1	IEEE	30
2	MPDI Applied Science	6
3	IET	4
4	CiteSeerX	2
Journal		
1	Scientific reports	3
2	SAIEE Africa Research Journal	1
3	Indonesian Journal of Electrical Engineering and Computer Science	2
4	ArXiv	9
5	ACM	3
6	ScienceDirect	9
7	Springer	17
8	IJESC	2
Books		
1	Manning	1
Total		89

Table 7: Differences between this review article and the recent/existing review papers.

Author(s) and date of publication	Paper title	Aim/focus/objective	Paper coverage (year) and scope
(1) This paper	Ear Biometrics using Deep Learning: A Survey	This paper proved that using a bag of feature techniques and the classification technique of deep learning using convolutional neural network was better than standard machine learning techniques	Eighty-nine (89) application papers that are deep learning ear identification methods are reviewed in this paper
(2) Emeršič et al. [60] 29 June 2017	Training convolutional neural networks with limited training data for ear recognition in the wild	It was a handcrafted feature extraction method, such as LBP and patterns of oriented edge magnitudes (POEM), and CNN-based feature extraction methods were used to obtain the ear identification	Forty-one (41) application papers that are deep learning ear identification methods are reviewed in this paper
(3) Tian et al. [21] 16 February 2017	Ear recognition based on deep convolutional network	This system used deep convolutional neural network (CNN) to ear recognition. There were occlusions like no earrings, headsets, or similar occlusions	Fifteen (15) application papers that are deep learning ear identification methods are reviewed in this paper
(4) Raveane et al. [64] 18 June 2019	Ear detection and localization with convolutional neural networks in natural images and videos	This system used variable conditions, and this could also be because of the odd shape of the human ears and changing lighting conditions	Thirty-five (35) application papers that are deep learning ear identification methods are reviewed in this paper

Raveane et al. [64] presented that it is difficult to precisely detect and locate an ear within an image. This challenge increases when working with variable conditions, and this could also be because of the odd shape of the human ears and changing lighting conditions. The changing profile shape of an ear when photographed is displayed [64]. The ear detection system was a multiple convolutional neural network with a detection grouping algorithm to identify the ear's presence and location. The proposed method matches other methods' performance when analysed against clean and purpose-shot photographs, reaching an accuracy of upwards of 98%. It outperforms other works with a rate of over 86% when the system is subjected to noncooperative natural images where the subject appears in challenging orientations and photographic conditions.

Multiple scale faster region-based convolutional neural network (Faster R-CNN) to detect ears from 2D profile images was proposed by Zhang and Mu [65]. This method uses three regions of different scales to detect information from the ears' location within the context of the ear image. The system was tested with 200 web images and achieved an accuracy of 98% . Other experiments conducted were on the Collection J2 of the University of Notre Dame Biometrics Database (UND-J2) and the University of Beira Interior Ear (UBEAR) dataset; these achieved a detection rate of 100% and 98.22%, respectively, but these datasets contained large occlusions, scale, and pose variation.

Kohlakala and Coetzer [66] presented semiautomated and fully automated ear-based biometric verification systems. A convolutional neural network (CNN) and

Table 8: Cont. differences between this review article and the recent/existing review papers.

Author(s) and date of publication	Paper title	Aim/focus/objective	Paper coverage (year) and scope
(5) Zhang and Mu [65] 24 January 2017	Ear detection under uncontrolled conditions with multiple scale faster region-based convolutional neural networks	This system contained large occlusions, scale, and pose variation	Forty-one (41) application papers that are deep learning ear identification methods are reviewed in this paper
(6) Kohlakala and Coetzer [66] 1 June 2021	Ear-based biometric authentication through the detection of prominent contour	It is used to classify ears either in the foreground or background of the image. The binary contour image applied the matching for feature extraction, and this was carried out by implementing Euclidean distance measure, which had a ranking to verify for authentication	Twenty-one (21) application papers that are deep learning ear identification methods are reviewed in this paper
(7) Tomczyk and Szczepaniak [67] 13 December 2019	Ear detection using convolutional neural network on graphs with filter rotation	It shows the published experimental results that the approach performed the rotation equivalence property to detect rotated structures	Forty (40) application papers that are deep learning ear identification methods are reviewed in this paper
(8) Alshazly et al. [68] 8 December 2019	Handcrafted versus CNN features for ear recognition	The paper took seven performing handcrafted descriptors to extract the discriminating ear image. They then took the extracted ear and trained it using Support Vector Machines (SVM) to learn a suitable model	Seventy-three (73) application papers that are deep learning ear identification methods are reviewed in this paper

Table 9: Cont. differences between this review article and the recent/existing review papers.

Author(s) and date of publication	Paper title	Aim/focus/objective	Paper coverage (year) and scope
(9) Alkababji and Mohammed [69] 1 April 2021	Real-time ear recognition using deep learning	It used the principal component analysis (PCA) and a genetic algorithm for feature reduction and selection	Twenty-three (23) application papers that are deep learning ear identification methods are reviewed in this paper
(10) Jamil et al. [70] 1 August 2018	Can convolution neural network (CNN) triumph in ear recognition of uniform illumination invariant?	They considered that their work was the first to test the performance of CNN on very underexposed or overexposed images	Thirty-two (32) application papers that are deep learning ear identification methods are reviewed in this paper
(11) Hansley et al. [71] 24 October 2017	Employing fusion of learned and handcrafted features for unconstrained ear recognition	This was conducted using handcrafted descriptors, which were fused to improve recognition	Thirty-one (31) application papers that are deep learning ear identification methods are reviewed in this paper

morphological postprocessing were used to manually identify the ear region. They are used to classify ears either in the foreground or background of the image. The binary contour image applied the matching for feature extraction, and this was carried out by implementing a Euclidean distance measure, which had a ranking to verify for authentication. The Mathematical Analysis of Images ear database and the Indian Institute of Technology, Delhi, ear database were two databases, which achieved 99.20% and 96.06%, respectively.

Geometric deep learning (GDL) generalises convolutional neural network (CNN) to non-Euclidean domains, presented by [67] Tomczyk and Szczepaniak. It used convolutional filters with a mixture of Gaussian models. These filters were used so that the images could be easily rotated without interpolation. Their paper published experimental results on the approach of the rotation equivalence property to detect rotated structures. The result showed that it did not require labour-intensive training on all rotated and non-rotated images.

Alshazly et al. [68] presented and compared ear recognition models built with handcrafted and convolutional neural networks (CNN) features. The paper took seven handcrafted descriptors to extract the discriminating ear image. The extracted ear was trained using Support Vector Machines (SVM) to learn a suitable model, after which the CNN-based model used the AlexNet architecture. The results obtained on three ear datasets show the CNN-based models' performance by 22%. This paper also investigated if the left and right ears have symmetry. The results obtained by the two datasets indicate a high impact of balance between the ears.

Alkababji and Mohammed [69] presented the use of a deep learning item detector, which they called faster region-based convolutional neural networks (Faster R-CNN) for ear detection. This convolutional neural network (CNN) is used for feature extraction. It used Principal Component Analysis (PCA) and a genetic algorithm for feature reduction and selection. It also used a connected artificial neural network as the matcher. The results achieved an accuracy of 97.8% success.

Jamil et al. [70] built and trained a CNN model for ear biometrics in various uniform illuminations measured using lumens. They considered that their work was the first to test the performance of CNN on very underexposed or over-exposed images. The results showed that for images with uniform illumination and a luminance of above 25 lux, the results achieved were 100%. The CNN model had problems recognising images when the lux was below ten, but still obtained an accuracy of 97%. This result shows that the CNN architecture performs just as well as the other systems. It was found that the data set had rotations that affected the results.

Hansley et al. [71] presented an unconstrained ear recognition framework that was better than the current state-of-the-art systems using publicly available databases. They developed CNN-based solutions for ear normalisation and description. This was performed using handcrafted descriptors, which were fused to improve recognition, and was carried out in two stages. The first stage was to find the landmark detectors, which were untrained scenarios. The next step was to generate a geometric image normalisation to boost the performance. It was seen that the CNN descriptor was better than other CNN-based works in the literature. The obtained results were higher than different reported results for the UERC challenge.

6. Difference between Reviewed Articles

Tables 6 and 7 show the comparison of this review paper with recent/existing review papers to establish their differences. A critical analysis of Tables 6 and 7 reveals that the most recent and closest review paper to this article is the excellent review work. Tables 8 and 9 show the differences between the review article and the existing review papers.

7. Conclusion

This paper presented a comparative survey of various convolutional neural network architectures, with their strengths and weaknesses. A thorough analysis of the existing deep convolutional neural network methods used for ear identification was discussed. Furthermore, the paper discussed and investigated the success of using the ear as a primary biometric system for identification and verification. It was found that other works battled to identify the ear if pose and angle of the image were changed. This will be looked at in the future as to how this can be eliminated. Also, it was found that if clothes, hair, ear ornaments, and jewellery were not removed, it interfered with the identification of an ear. In addition, a study was performed on ear identification benchmarks and their performance on other CNN models measured by standard evaluating metrics.

Future work will be to investigate and implement EfficientNet models to automatically identify ears on the most prominent and publicly available datasets. EfficientNets that achieved state-of-the-art performance over other architectures to maximize accuracy and efficiency were explored and fine-tuned on profile images. The fine-tuning technique is valuable to utilize rich generic features learned from significant dataset sources such as ImageNet to complement the lack of annotated datasets affecting the ear domains.

Abbreviations

NN: Neural network
CNN: Convolutional neural network.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, and M. S. Nixon, "A survey on ear biometrics," *ACM Comput Surv*, vol. 45, no. 2, pp. 1–35, 2013.

[2] B. C. Bir, *Ear Biometrics*, Springer, Boston, MA, USA, 3D edition, 2009.

[3] J. Heaton, I. Goodfellow, B. Yoshua, and C. Aaron, *Deep Learning*, Springer, New York, NY, USA, 2018.

[4] F. Chollet, *Deep learning with Python*, Vol. 361, Manning, New York, NY, USA, 2018.

[5] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, Boston, MA, USA, June 2015.

[6] J. Dai, K. He, and J. Sun, "Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1635–1643, 2015.

[7] A. Lomuscio and L. Maganti, "An Approach to reachability Analysis for feed-forward relu neural networks," 2017, https://arxiv.org/abs/1706.07351?context=cs.

[8] J. Kleesiek, G. Urban, A. Hubert et al., "Deep MRI brain extraction: a 3D convolutional neural network for skull stripping," *NeuroImage*, vol. 129, pp. 460–469, 2016.

[9] F. Bonanno, G. Capizzi, G. L. Sciuto, C. Napoli, G. Pappalardo, and E. Tramontana, "A cascade neural network architecture investigating surface plasmon polaritons propagation for thin metals in openmp," in *International Conference on Artificial Intelligence and Soft Computing*-Springer, New York, NY, USA, 2014.

[10] Y. Weng, T. Zhou, Y. Li, and X. Qiu, "NAS-Unet: neural architecture search for medical image segmentation," *IEEE Access*, vol. 7, Article ID 44247, 2019.

[11] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, and M. S. Nasrin, "The history Began from Alexnet: A comprehensive survey on deep Learning Approaches," 2018, https://arxiv.org/abs/1803.01164.

[12] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Ensembles of deep learning models and transfer learning for ear recognition," *Sensors*, vol. 19, no. 19, p. 4139, 2019.

[13] R. R. Hallac, J. Lee, M. Pressler, J. R. Seaward, and A. A. Kane, "Identifying ear abnormality from 2D photographs using convolutional neural networks," *Scientific Reports*, vol. 9, no. 1, Article ID 18198, 2019.

[14] E. Ž, D. Štepec, V. Štruc, P. Peer, A. George, and A. Ahmad, "The unconstrained ear recognition challenge," in *Proceedings of the 2017 IEEE International joint Conference on Biometrics (IJCB)*, pp. 715–724, Denver, CO, USA, October 2017.

[15] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Deep convolutional neural networks for unconstrained ear recognition," *IEEE Access*, vol. 8, Article ID 170295, 2020.

[16] Y. Zhang, Z. Mu, L. Yuan, and C. Yu, "Ear verification under uncontrolled conditions with convolutional neural networks," *IET Biometrics*, vol. 7, no. 3, pp. 185–198, 2018.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, Las Vegas, NV, USA, June 2016.

[18] Z. Li, Z. Hu, J. Xu, T. Tan, H. Chen, and Z. Duan, "Computer-aided diagnosis of lung carcinoma using deep learning-a pilot study," 2018, https://arxiv.org/abs/1803.05471.

[19] K. Radhika, K. Devika, T. Aswathi, P. Sreevidya, V. Sowmya, and K. Soman, "Performance analysis of NASNet on unconstrained ear recognition," in *Nature Inspired Computing for Data Science*Springer, New York, NY, USA, 2020.

[20] Y. Li, "Deep reinforcement Learning: An overview," 2017, https://arxiv.org/abs/170107274.

[21] L. Tian and Z. Mu, "Ear recognition based on deep convolutional network," in *Proceedings of the 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 437–441, Datong, China, October 2016.

[22] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2065–2074, 2017.

[23] S. Dodge, J. Mounsef, and L. Karam, "Unconstrained ear recognition using deep neural networks," *IET Biometrics*, vol. 7, no. 3, pp. 207–214, 2018.

[24] R. Hussain, A. Lalande, K. B. Girum, C. Guigou, and A. Bozorg Grayeli, "Automatic segmentation of inner ear on CT-scan using auto-context convolutional neural network," *Scientific Reports*, vol. 11, no. 1, pp. 4406–4410, 2021.

[25] S. Zhou, F. Wang, Z. Huang, and J. Wang, "Discriminative feature learning with consistent attention regularization for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8040–8049, Seoul, Republic of Korea, October 2019.

[26] A. Kumar, "Iit delhi ear database version 1.0," 2007, http://webold.iitd.ac.in/biometrics/Database_Ear.htm.

[27] R. Raposo, E. Hoyle, A. Peixinho, and H. ProenĂ§a, "UBEAR: A dataset of ear images captured on-the-move in uncontrolled conditions," in *Proceedings of the 2011 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, Paris, France, April 2011.

[28] Ž Emeršič, V. Štruc, and P. Peer, "Ear recognition: more than a survey," *Neurocomputing*, vol. 255, pp. 26–39, 2017.

[29] V. T. Hoang, "EarVN1.0: a new large-scale ear images dataset in the wild," *Data in Brief*, vol. 27, Article ID 104630, 2019.

[30] Y. Zhang, Z. C. Mu, L. Yuan, C. Yu, and L Qing, "USTB-Helloear: a large database of ear images photographed under uncontrolled conditions," in *Image and Graphics*, Springer, New York, NY, USA, 2017.

[31] E. Gonzalez, L. Alvarez, and L. Mazorra, "Ami Ear Database," 2012, http://ctim.ulpgc.es/research_works/ami_ear_database/.

[32] D. Frejlichowski and N. Tyszkiewicz, "The west pomeranian university of technology ear database - a tool for testing biometric algorithms," *Image Analysis and Recognition*, Springer, Berlin, Germany, 2010.

[33] V. Emeršič and P. Peer, "Ear Biometric database in the wild," in *Proceedings of the 2015 4th International Work Conference on Bioinspired Intelligence (IWOBI)*, pp. 27–32, San Sebastian, Spain, June 2015.

[34] M. A. Carreira-Perpinan, "Compression neural networks for feature extraction: Application to human recognition from ear images," Master's thesis, Faculty of Informatics, Technical University of Madrid, Madrid, Spain, 1995.

[35] S. Prakash, U. Jayaraman, and P. Gupta, "Connected component based technique for automatic ear detection," in *Proceedings of the 2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 2741–2744, IEEE, Cairo, Egypt, November 2009.

[36] I. Alberink and A. Ruifrok, "Performance of the FearID earprint identification system," *Forensic Science International*, vol. 166, no. 2-3, pp. 145–154, 2007.

[37] P. Yan and K. Bowyer, "Empirical evaluation of advanced ear biometrics," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, p. 41, San Diego, CA, USA, September 2005.

[38] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

[39] T. Sim, S. Baker, and M. Bsat, *The CMU Pose, Illumination, and Expression (PIE) Database of Human Faces*, Carnegie Mellon University, Pittsburgh, PA, 2001.

[40] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: the extended M2VTS database," in *Proceedings of the Second International Conference on Audio and Video-Based Biometric Person Authentication*, pp. 965-966, 1999.

[41] A. Abaza, *High performance image processing techniques in Automated identification systems*, West Virginia University, Morgantown, WV, USA, 2008.

[42] S. Ansari and P. Gupta, "Localization of ear using outer helix curve of the ear," in *Proceedings of the 2007 International Conference on Computing: Theory and Applications (ICCTA'07)*, pp. 688–692, IEEE, Kolkata, India, March 2007.

[43] M. Abdel-Mottaleb and J. Zhou, "Human ear recognition from face profile images," in *International Conference on Biometrics*Springer, New York, NY, USA, 2006.

[44] T. Yuizono, Y. Wang, K. Satoh, and S. Nakayama, "Study on individual recognition for ear images by using genetic local search," in *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600)*, pp. 237–242, IEEE, Honolulu, HI, USA, May 2002.

[45] H. Chen and B. Bhanu, "Human ear recognition in 3D," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 718–737, 2007.

[46] B. Arbab-Zavar and M. S. Nixon, "On shape-mediated enrolment in ear biometrics," in *International Symposium on Visual Computing*, Springer, New York, NY, USA, 2007.

[47] M. Burge and W. Burger, *Ear Biometrics*, pp. 273–285, Biometrics Springer, New York, NY, USA, 1996.

[48] M. Choras, "Image feature extraction methods for ear biometrics–a survey," in *Proceedings of the 6th International Conference on Computer Information Systems and Industrial Management Applications (CISIM'07)*, June 2007.

[49] D. Shailaja and P. Gupta, "A simple geometric approach for ear recognition," in *Proceedings of the 9th International Conference on Information Technology (ICIT'06)*, pp. 164–167, IEEE, Bhubaneswar, India, December 2006.

[50] A. H. Cummings, M. S. Nixon, and J. N. Carter, "A novel ray analogy for enrolment of ear biometrics," in *Proceeding of the 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–6, Washington, DC, USA, September 2010.

[51] A. Abaza, C. Hebert, and M. A. F. Harrison, "Fast learning ear detection for real-time surveillance," in *Proceedings of the 2010 fourth IEEE international Conference on Biometrics: theory, Applications and systems (BTAS)*, pp. 1–6, Washington, DC, USA, September 2010.

[52] S. Prakash and P. Gupta, "An efficient ear localization technique," *Image and Vision Computing*, vol. 30, no. 1, pp. 38–50, 2012.

[53] V. Basrur, F. Yang, T. Kushimoto et al., "Proteomic analysis of early melanosomes: identification of novel melanosomal proteins," *Journal of Proteome Research*, vol. 2, no. 1, pp. 69–79, 2003.

[54] M. Rahman, M. R. Islam, N. I. Bhuiyan, B. Ahmed, and M. A. Islam, "Person identification using ear biometrics," *International Journal of The Computer, the Internet and Management*, vol. 15, no. 2, pp. 1–8, 2007.

[55] K. Chang, K. W. Bowyer, S. Sarkar, and B. Victor, "Comparison and combination of ear and face images in appearance-based biometrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1160–1165, 2003.

[56] I. Naseem, R. Togneri, and M. Bennamoun, "Sparse representation for ear biometrics," in *International Symposium on Visual Computing*, Springer, New York, NY, USA, 2008.

[57] L. Nanni and A. Lumini, "A multi-matcher for ear authentication," *Pattern Recognition Letters*, vol. 28, no. 16, pp. 2219–2226, 2007.

[58] P. Yan and K. W. Bowyer, "Biometric recognition using 3D ear shape," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1297–1308, 2007.

[59] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometric recognition using deep learning: a survey," 2019, https://arxiv.org/abs/1912.00271.

[60] Z. Emeršič, D. Štepec, V. Štruc, and P. Peer, "Training convolutional neural networks with limited training data for ear recognition in the wild," in *Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, USA, May 2017.

[61] Zq Wang and Xd Yan, "Multi-scale feature extraction algorithm of ear image," in *Proceedings of the 2011 International Conference on Electric Information and Control Engineering*, pp. 528–531, IEEE, Wuhan, China, April 2011.

[62] N. S. Vu, H. M. Dee, and A. Caplier, "Face recognition using the POEM descriptor," *Pattern Recognition*, vol. 45, no. 7, pp. 2478–2488, 2012.

[63] D. P. Chowdhury, S. Bakshi, G. Guo, and P. K. Sa, "On applicability of tunable filter bank based feature for ear biometrics: a study from constrained to unconstrained," *Journal of Medical Systems*, vol. 42, no. 1, pp. 11–20, 2018.

[64] W. Raveane, P. L. Galdamez, and M. A. Gonzalez Arrieta, "Ear detection and localization with convolutional neural networks in natural images and videos," *Processes*, vol. 7, no. 7, p. 457, 2019.

[65] Y. Zhang and Z. Mu, "Ear detection under uncontrolled conditions with multiple scale faster region-based convolutional neural networks," *Symmetry*, vol. 9, no. 4, p. 53, 2017.

[66] A. Kohlakala and J. Coetzer, "Ear-based biometric authentication through the detection of prominent contours," *SAIEE Africa Research Journal*, vol. 112, no. 2, pp. 89–98, 2021.

[67] A. Tomczyk and P. S. Szczepaniak, "Ear detection using convolutional neural network on graphs with filter rotation," *Sensors*, vol. 19, no. 24, p. 5510, 2019.

[68] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Handcrafted versus CNN features for ear recognition," *Symmetry*, vol. 11, no. 12, p. 1493, 2019.

[69] A. M. Alkababji and O. H. Mohammed, "Real time ear recognition using deep learning," *Telkomnika*, vol. 19, no. 2, pp. 523–530, 2021.

[70] N. Jamil, A. Almisreb, S. M. Z. S. Z. Ariffin, N. Md Din, and R. Hamzah, "Can Convolution neural network (CNN) triumph in ear recognition of uniform illumination invariant?" *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, 2018.

[71] E. E. Hansley, M. P. Segundo, and S. Sarkar, "Employing fusion of learned and handcrafted features for unconstrained ear recognition," *IET Biometrics*, vol. 7, no. 3, pp. 215–223, 2018.

Chapter 3

Detection of Ears from Images using Deep Learning

3.1 Exploration of Ear Biometrics with Deep Learning

3.1.1 Brief Overview

This section introduces a research paper that compares the accuracy of ear detection between machine and deep learning. The commonly used machine learning techniques reviewed were Naïve Bayes, Decision Tree and K-Nearest Neighbor, which were then compared to the classification technique of Deep Learning using Convolution Neural Networks. The work reviews the source of ear modelling, details the algorithms, methods and processing steps and finally tracks the input dataset's error and limitations for the final results obtained for ear identification.

The paper is published in the *Lecture Notes in Computer Science - Computer Vision and Graphics*, Springer, Cham.

Exploration of Ear Biometrics with Deep Learning

Aimee Booysens and Serestina Viriri( )

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

Abstract. Ear recognition has become a vital issue in image processing to identification and analysis for many geometric applications. This article reviews the source of ear modelling, details the algorithms, methods and processing steps and finally tracks the error and limitations for the input database for the final results obtain for ear identification. The commonly used machine-learning techniques used were Naïve Bayes, Decision Tree and K-Nearest Neighbor, which then compared to the classification technique of Deep Learning using Convolution Neural Networks. The results achieved in this article by the Deep Learning using Convolution Neural Network was 92.00% average ear identification rate for both left and right ear.

1 Introduction

Biometric technology is a topic that has had a growing interest over the years because there is an increasing need for security and authentication, among others. Recognition systems have centred mainly on biometric features, and these are faces or fingerprints. These traditional features are widely studied, and their behaviour understood, while other less known biometric features, such as ears, have the potential to become the better applications.

The ear has an advantage over the traditional biometric features, as they have a stable structure that does not change as a person ages. It is a known fact that the face changes continually based on expressions; this does not occur with ears. Also, the ears environment are always known as they located on the sides of the head. In contrast, facial recognition typically requires a controlled environment for accuracy; this type of situation is not always present. Lastly, the ear does not need proximity to achieve capture, whereas the traditional biometric features do, like the eyes and fingers. The qualities mentioned above make the ear a promising field to study for recognition.

It is an accepted fact that the shape and appearance of the ears are unique per individual and that they are of fixed form during the lifetime of a person. According to reports, the variation over time in a human ear is most noticeable from when a person is four-months-old to eight years old and after the age of 70 years-old. The ear growth that occurs between four-months-old to eight years

L. J. Chmielewski et al. (Eds.): ICCVG 2020, LNCS 12334, pp. 25–35, 2020.
https://doi.org/10.1007/978-3-030-59006-2_3

old is linear, after that it is constant until the person is around 70 years old. At this age, the ears begin to increases again [1]. While there are small changes that take place in the ear structure, these are restricted to the ear lobe and are not linear. Such predictability of the ear makes it an exciting realm for research as machine-learning can quickly identify the ear and obtain a result.

Machine-learning is taking over the world and is becoming part of everybody's life. One of the branches of machine learning is that of Deep Learning,[2]. This branch deals with different algorithms to the structure and function of the human brain with multiple layers of a neural network.

In this article, we will be taking a look at how commonly used machine-learning techniques compare to Deep Learning using Convolution Neural Networks (CNN) in the identification of the ear, and the left and right ear.

2 Related Work

Zhang and Mu, [3], investigated ways of detection ears from 2D profile images. They investigated if multiple-scale faster region-based Neural Networks would work to detect images automatically. Shortfalls of this article are the following are not taken into account: pose variation, occlusion and imaging conditions. Even with these shortfalls, the article managed to achieve for web images 98% detection accuracy rate for both ears. The other two databases have the same test done to them, and this was the collection of J2 of University of Notre Dame Biometrics which achieved 100% for ear detection rate and the University of Beira Interior Ear Database (UBEAR) reached 98.22% for the ear detection rate.

Galdámez et al. proposed a solution to do ear identification [4]. This solution involved a Convolution Neural Networks (CNN) for any image inputted into the system and will give an output of the ear. The results that obtained from the CNN compared to the results for the same dataset of other machine-learning techniques of Principal Component Analysis (PCA), Linear discriminant analysis (LDA) and Speeded-up robust feature (SURF). These tests are carried out on two different datasets Avila's Police School and Bisite Video Dataset. On the first dataset, the model for CNN achieved a detection accuracy rate of 84.82%. CNN compared to PCA, LDA and SURF, only achieved a detection accuracy rate of 66.37%, 68.36% and 76.75% respectfully. The second dataset achieved CNN achieved a detection accuracy rate of 40.12% and compared to PCA which 21.83% ear accuracy rate, LDA which 27.37% ear accuracy rate and SURF which achieved an ear accuracy rate of 30.68%. The reason that the second dataset achieved so low is that the images obtained from videos.

Multimodal biometric systems address numerous problems observed in single modal biometric systems which are proposed by Amirthalingam and Radhamani, [5]. The sophisticated methods employed to find the right combination of multiple biometric modalities and various level of fusion applied to get the best possible recognition result discussed in this article. The combination of face and ear modality suggested, and the proposed framework of the biometric system are

given. In this article, it claims that multi-biometrics improves over a single system, and uncorrelated modalities used to achieve performance in the multimodal system. This system produced an average ear detection rate of 72%.

3 Methods and Techniques

This research's depiction of the ear biometrics model, Fig. 1.

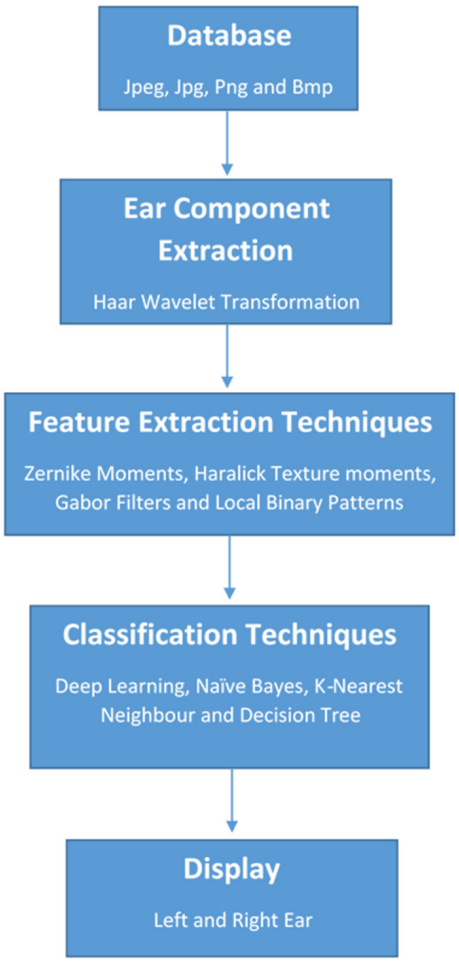


Fig. 1. Overview of ear identification system

3.1 Ear Components

The ear components are extracted using the Haar Wavelet Transformation (HWT) [6]. HWT is one of wavelet transformations. This transformation cross multiplies a function against the Haar Wavelet and is defined as Eq. (1).

$$yn = Hn.Xn \tag{1}$$

where yn is the Haar Transformation at an n-input at function Xn. Hn is the Haar Transformation which is a matrix which can be defined like Eq. (2).

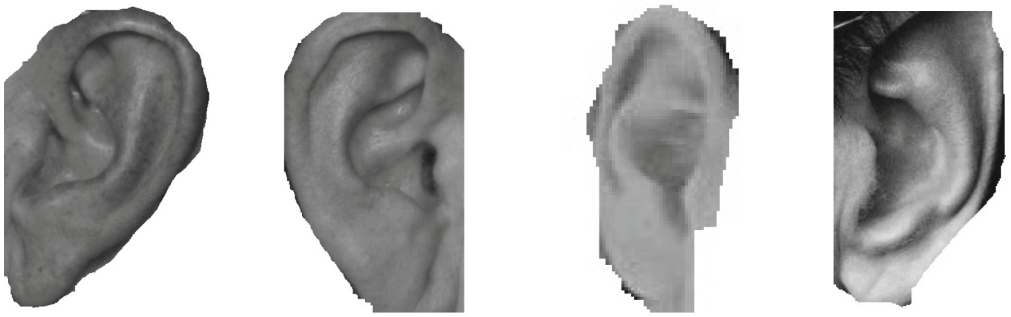
$$\frac{1}{2}\begin{Bmatrix} 1 & 1 & \sqrt{2} & 0 \\ 1 & 1 & -\sqrt{2} & 0 \\ 1 & -1 & 0 & \sqrt{2} \\ 1 & -1 & 0 & -\sqrt{2} \end{Bmatrix} \tag{2}$$

Some of the properties that the Haar Wavelet Transformation [6] has is that there is no need for multiplications and that the input and output matrix of the same length. The components that were extracted are the left ear and right ear.

The figures in 2 and 3 shows a sample of the images that were used in this article and how the Haar Wavelet Transformation extracts the ear from the original facial or 2D image.

(a) Example 1 of a 2D profile image (b) Example 2 of a 2D profile image (c) Example 1 of a facial image (d) Example 2 of a facial image

Fig. 2. Examples of original ear images



(a) Example 1 of ear extracted from 2D profile image (b) Example 2 of ear extracted from 2D profile image (c) Example 1 of ear extracted from facial image (d) Example 2 of ear extracted from facial image

Fig. 3. Examples of extracted ear images

3.2 Feature Extraction

Feature extraction techniques are used on each ear image in the dataset to obtain a feature vector. Analysis was done on different textural, structural or geometrical feature extractions. The feature extraction techniques that were used are Zernike Moments, Local Binary Pattern (LBP), Gabor Filter and Haralick texture Moments. Investigation that was done in order to obtain the correct feature extraction for all component.

The Zernike Moments are used to overcome redundancy in which certain geometric moments obtain [7]. They are a class of orthogonal moments which are rotational invariant and effective in image representation. Zernike moments are a set of complex, orthogonal polynomials defined as the interior of the unit circle. The general form of the Zernike Moments defined in Eq. (3).

$$Z_{nm}(x, y) = Z_{nm}(p, \theta) = R_{nm}(p)\epsilon^{jm\theta} \tag{3}$$

where x, y, p and θ correspond to Cartesian and Polar coordinates respectively, $n \in Z^+$ and $m \in Z$, constrained to $n - m$ even, $m \leq n$

$$R_{nm}(p) = \sum_{k=0}^{\frac{n-m}{2}} \frac{(-1)^k (n-k)!}{k!(\frac{n+m}{2} - k)!(\frac{n-m}{2} - k)!} p^{n-2k} \tag{4}$$

where $R_{nm}(p)$ is a radial polynomial and k is the order.

The Haralick Texture Moments are texture features that can analyse the spatial distribution of the image's texture features [7] with different spatial positions and angles. This research computer four of these Haralick Texture Moments – Energy, Entropy, Correlation and Homogeneity.

Entropy is the reflection of the disorder and the complexity of the texture of the images. This is defined using Eq. (5).

$$Entropy = \sum_{ij} \hat{f}(i, j) \log \hat{f}(i, j) \tag{5}$$

where $\hat{f}(i, j)$ is the $[i, j]$ entry if the grey level value of image matrix and i and j are points on the image matrix.

Energy is the measure of the local homogeneity and is the opposite of Entropy. It shows the uniformity of the texture of the images and computed using the Eq. (6).

$$Energy = \sum_{ij} \hat{f}(i, j)^2 \tag{6}$$

where $\hat{f}(i, j)$ is the $[i, j]$ entry if the grey level value of image matrix and i and j are points on the image matrix.

Homogeneity is the reflection of "equaliness" of the images' textures and scale of local changes in the texture of the images. If this result is high, then there is

no difference between the regions with regards to the texture of the images and is defined in Eq. (7).

$$Homogeneity = \sum_{i} \sum_{j} \frac{1}{1+(i-j)^2} \hat{f}_{i,j} \tag{7}$$

where $\hat{f}(i,j)$ is the $[i,j]$ entry if the grey level value of image matrix and i and j are points in the image matrix.

Correlation is the consistency of the texture of the images and described using equation (8).

$$Correlation = \sum_{ij} \frac{(i-\mu_i)(j-\mu_j)\hat{f}(i,j)}{\sigma_i \sigma_j} \tag{8}$$

in which μ_j, μ_i, σ_i and σ_j are described as:

$$\mu_i = \sum_{i=1}^{n} \sum_{j=1}^{n} i\hat{f}(i,j) \quad \mu_j = \sum_{i=1}^{n} \sum_{j=1}^{n} j\hat{f}(i,j) \tag{9}$$

$$\sigma_i = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} (i-\mu_i)^2 \hat{f}(i,j)} \quad \sigma_j = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} (i-\mu_j)^2 \hat{f}(i,j)} \tag{10}$$

The Gabor Filters are geometric moments which are the product between an elliptical Gaussian and a sinusoidal, [8]. Gabor elementary function is the product of the pulse with a harmonic oscillation of frequency.

$$g(t) = \epsilon^{-\alpha^2 (t-t_0)^2} \epsilon^{-i2\pi(f-f_o)+\phi} \tag{11}$$

where α is the time duration of the Gaussian envelope, t_0 denotes the centroid, f_0 is the frequency of the sinusoidal and ϕ indicates the phase shift.

Local Binary Pattern (LBP) is a geometric moment operator that described the surrounding of the pixels by obtaining a bit code of a pixel [9], this is defined using Eq. (12).

$$C = \sum_{k=0}^{k=7} (2^k b_k) \tag{12}$$

$$b_k = \begin{cases} 1, \sum_{k=0}^{k=7} (t_k \geq C) \\ \\ 0, \sum_{k=0}^{k=7} (t_k < C) \end{cases}$$

where t_k is the grayscale amount. b_k is the binary variables between 1 and 0 and C is a constant value of 0 and 1.

3.3 Classification

The main classification technique in this article is Deep Learning with the use of CNN, and these results are compared to numerous other classification techniques. These are different supervised and unsupervised machine-learning algorithms. The classification techniques used are Naïve Bayes, K-Nearest Neighbor and Decision Tree.

Deep Learning with Convolutional Neural Network (CNN) is a generalisation of feed forward neural networks to a sequence [2]. Given that a particular sequence has a certain number of inputs $(x_1, ..., x_t)$ this will then compute a sequence of outputs $(y_1, ..., y_t)$ which is done by using the below Eq. (15) and (14):

$$h_t = \sigma(W^{hx}x^t + W^{hh}h_{t-1}) \tag{13}$$

$$y_t = W^{yh}h_t \tag{14}$$

The RNN is easy to apply the map sequence when there is alignment between and input and an output.

Decision Tree uses recursive partitioning to separate the dataset by finding the best variable [10], and using the selected variable to split the data. Then using the entropy, defined in Eqs. (15) and (16), to calculate the difference that variable would make on the results if chosen. If the entropy is zero, then that variable is perfect to use, else a new variable needs to be selected.

$$H(D) = -\sum_{i=1}^{k} P(C_i|D)\log_k(P(C_i|D)) \tag{15}$$

where the entropy of a sample D concerns the target variable of k possible classes C_i.

$$P(C_i|D) = \frac{number\ of\ correct\ observation\ for\ that\ class}{total\ observation\ for\ that\ class} \tag{16}$$

where the probability of class C_i in D is obtained directly from the dataset.

Naïve Bayes classify an instance by assuming the presence or absence of a particular feature. Checks if it is unrelated to the presence or absence of another feature, given in the class variable [10,11]. Naïve Bayes calculation is done by using the probability for which it occurred, as defined in Eq. (17).

$$P(x_1, ..., x_n|y) = \frac{\prod_{i=1}^{n} P(y)P(x_n|y)}{P(x_1, ..., x_n)} \tag{17}$$

where case y is a class value, attributes are $x_1, ..., x_n$ and n is the sample size.

K-Nearest Neighbor (KNN) classifies by using a majority vote of its neighbours. The case is assigned to the class with the most common amongst

its dataset. A distance function measures the KNN, for example, Euclidean, as defined in Eq. (18).

$$d = \sqrt{\sum_{i=1}^{n} (x_i - q_i)^2} \tag{18}$$

where n is the size of the data, x_i is an element in the dataset, and q_i is a central point.

4 Results and Discussion

This research used a union of four different image databases. The total dataset contained 2997 facial and 2D profile images of approximately 360 subjects, which had been split the dataset into two groups of left and right ear. The datasets that are used were Annotated Web Ears (AWE) [12] and Annotated Web Ears additional database (AWEA) [12], AMI Ear (AMI) [13] and IIT Delhi Ear Database (IIT) [1].

The analysis completed was to obtain which feature extraction technique would achieve the most accurate True Positive Rate (TPR) for ear identification. The results were obtained by taking a combination of the feature vectors for each component and then classified using K-Nearest Neighbour to observe which combination of feature extraction technique achieved the highest True Positive Rate, results obtained are in Table 1 and Table 2.

Zernike Moments is a geometric feature extraction technique which is a useful feature extraction technique as it correctly obtains the edges of the ear. This technique is used in the normalised feature vector as it captures the shape and the proportion of the ear. Zernike Moments on its own achieved a TPR of 82.71% for the left ear and 84.49% for the right ear.

Local Binary Pattern is to a geometric feature extraction technique. This feature extraction technique correctly identified the inner lines of the ear. This feature extraction technique is used in the normalised feature vector as it can detect the shape of the ear. Local Binary Pattern on its own obtained a TPR of 71.8% for the left ear and 73.8% for right ear.

Haralick Texture is a texture extraction technique. It works well as a feature extraction texture as it picks up the course and the colour gradient of the ear. Haralick Texture achieved an average TPR of 99.7% for the left ear and 86.69% for the right ear.

Gabor filter was the worst achieving geometric feature of all the feature vector techniques. This feature vector technique works well with the other feature vector techniques which achieved a TPR of 92.87%. Whereas alone it only achieved a TPR of 66.32% for the left ear and 70.72% for the right ear.

Table 1 shows the average TPR of all the combinations of these feature extraction techniques. If there are more feature extraction techniques used, the better the results are. Hence this reason why all four feature extraction techniques need to be used to obtain the feature vector. This vector is then fused and normalised to get ear identification.

Table 1. Accuracy rates for the ears for the combination of feature extraction techniques

Feature techniques	Percentage (%)
Gabor Filter and Zernike Moments and Haralick Texture and Local Binary Pattern	92,87
Gabor Filter and Zernike Moments and Local Binary Pattern	90,80
Gabor Filter and Zernike Moments and Haralick Texture	91,75
Gabor Filter and Haralick Texture and Local Binary Pattern	90,45
Gabor Filter and Zernike Moments and Local Binary Pattern	87,56
Zernike Moments and Haralick Texture and Local Binary Pattern	88,46
Zernike Moments and Haralick Texture	86,78
Zernike Moments and Local Binary Pattern	88,96
Haralick Texture and Local Binary Pattern	87,86
Gabor Filter and Zernike Moments	85,42
Gabor Filter and Haralick Texture	83,26
Gabor Filter and Local Binary Pattern	82,69

Table 2. Accuracy rates per ear per feature extraction techniques

	Zernike Moments	LBP	Gabor Filter	Haralick Texture Moments
Left ear	82.71%	71.8%	66.32%	99.7%
Right ear	84.49%	73.8%	70.72%	86.69%

4.1 Results for both Left and Right Ear

Several empirical experiments were carried out to investigate if Deep Learning with CNN is a better machine-learning algorithm to determine ear. Deep Learning with CNN was compared to the results of that of the more commonly used machine-learning algorithms(Decision Trees, Naïve Bayes and K-Nearest Neighbour). The testing was done by filling the training datasets with randomly chosen images from the original dataset and then tested with the different machine-learning algorithms and Deep Learning using CNN.

The tests showed that the K-Nearest Neighbour machine-learning algorithm achieved 60.2% average ear detection rate. The worst machine-learning algorithm

was Decision Tree which achieved 53.4% ear accuracy identification rate which is a variation of 6.8% between the worst and best ear accuracy identification rate.

Deep Learning using CNN and this achieved 91% ear detection rate, it achieved a better result than that of K-Nearest Neighbour machine-learning algorithm because of this Deep Learning is a better machine-learning algorithm than the commonly used ones.

Table 3. Comparison of related works results to this research results for ear biometrics identification

	Left ear	Right ear	Both Ears
Zhang and Mu [3]	-	-	98.74%
Galdamez et al. [4]	-	-	62.74%
Amirthalingam and Radhamani [5]	-	-	72%
This research	**95.36%**	**88.64%**	**92%**

Table 3 shows the comparison between the results achieved by related works for ear identification and this research work for ear identification. As demonstrated, this research results obtained a lower True Positive Rate than that of Zhang and Mu [3], which could be as a result of the research using a larger dataset. Whereas a comparison to that of the other research works, this research achieved a higher True Positive Rate (Fig. 4).

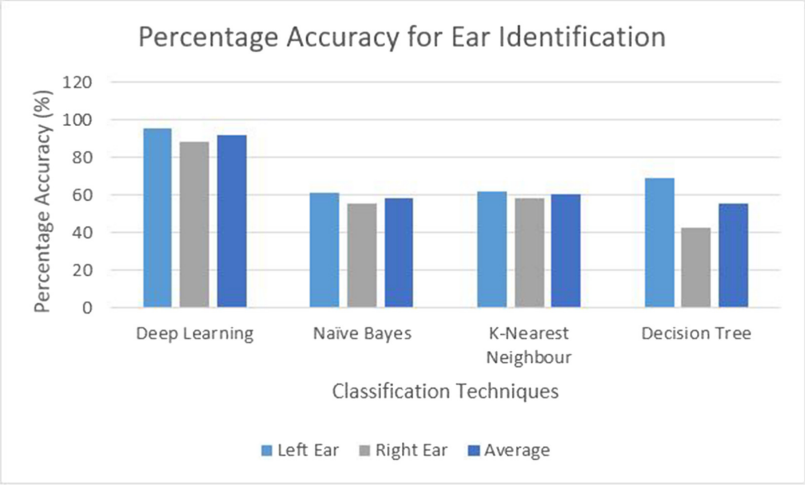


Fig. 4. Accuracy achieved for all classification techniques

5 Conclusion

This research presented an exploration of the ear biometric using Deep Learning Convolutional Neural Network. The ears were extracted and identified from 2D profile and facial images. The classification technique used was Deep Learning and was compared to more commonly used machine-learning algorithms. The identification that was done was left and right ear whereas other works only work on an average. The feature vector that was used is a combination of Haralick texture Moments, Zernike Moments, Gabor Filter and Local Binary Pattern. All these feature vectors are then fused and normalised to obtain a result. Naïve Bayes achieved 58.33% accuracy for the right ear, K-Nearest achieved 60.2% accuracy for the left ear, and Decision Tree achieved 55.72% accuracy for the left ear. Deep Learning achieved 95.36% accuracy for the left ear, 88.64% accuracy for the right ear. These results show that Deep Learning is a better way of obtaining results for the right and left ear. The total sum of the left and right ear identification rate of 92% was achieved.

References

1. Kumar, A., Wu, C.: Automated human identification using ear imaging. Pattern Recogn. **45**(3), 956–968 (2012)
2. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015)
3. Zhang, Y., Mu, Z.: Ear detection under uncontrolled conditions with multiple scale faster region-based convolutional neural networks. Symmetry **9**(4), 53 (2017)
4. Galdámez, P.L., Raveane, W., Arrieta, A.G.: A brief review of the ear recognition process using deep neural networks. J. Appl. Logic **24**, 62–70 (2017)
5. Amirthalingam, G., Radhamani, G.: A multimodal approach for face and ear biometric system. Int. J. Comput. Sci. Issues (IJCSI) **10**(5), 234 (2013)
6. Mulcahy, C.: Image compression using the Haar wavelet transform. Spelman Sci. Math. J. **1**(1), 22–31 (1997)
7. Teague, M.R.: Image analysis via the general theory of moments*. JOSA **70**(8), 920–930 (1980)
8. Berisha, S.: Image classification using Gabor filters and machine learning (2009)
9. Salah, S.H., Du, H., Al-Jawad, N.: Fusing local binary patterns with wavelet features for ethnicity identification. In: Proceedings of IEEE International Conference Signal Image Process, vol. 21, pp. 416–422 (2013)
10. Domingos, P.: A few useful things to know about machine learning. Commun. ACM **55**(10), 78–87 (2012)
11. Lowd, D., Domingos, P.: Naive bayes models for probability estimation. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 529–536. ACM (2005)
12. Emersic, Z., Struc, V., Peer, P.: Ear recognition: more than a survey. Neurocomputing (2017)
13. Esther Gonzalez, L.A., Mazorra, L.: AMI Ear Database (2018). Accessed 3 Feb 2014

3.2 Exploration of Ear Biometrics Using EfficientNet

3.2.1 Brief Overview

This section introduces a research paper whose main contribution is presenting the results of a CNN developed using EfficientNet. This paper presents the performance achieved in this research and shows the efficiency of EfficientNet on ear recognition. EfficientNet is a lightweight model based on the auto machine learning framework to develop a baseline EfficientNetB0 network and uniformly scaled up the depth, width and resolution using a simplified and effective compound coefficient to improve EfficientNet models B1-B8. The models performed effectively and attained superiority over the existing CNN models on the other CNN datasets.

The paper is published in the *Computational Intelligence and Neuroscience journal*.

Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 3514807, 14 pages
https://doi.org/10.1155/2022/3514807

Research Article

Exploration of Ear Biometrics Using EfficientNet

Aimee Booysens and Serestina Viriri

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

Correspondence should be addressed to Serestina Viriri; viriris@ukzn.ac.za

Received 23 May 2022; Revised 17 July 2022; Accepted 21 July 2022; Published 31 August 2022

Academic Editor: Muhammad Fazal Ijaz

Biometrics is the recognition of a human using biometric characteristics for identification, which may be physiological or behavioral. The physiological biometric features are the face, ear, iris, fingerprint, and handprint; behavioral biometrics are signatures, voice, gait pattern, and keystrokes. Numerous systems have been developed to distinguish biometric traits used in multiple applications, such as forensic investigations and security systems. With the current worldwide pandemic, facial identification has failed due to users wearing masks; however, the human ear has proven more suitable as it is visible. Therefore, the main contribution is to present the results of a CNN developed using EfficientNet. This paper presents the performance achieved in this research and shows the efficiency of EfficientNet on ear recognition. The nine variants of EfficientNets were fine-tuned and implemented on multiple publicly available ear datasets. The experiments showed that EfficientNet variant B8 achieved the best accuracy of 98.45%.

1. Introduction

The ear begins to develop in a fetus during the fifth and seventh weeks of pregnancy [1]. At this stage, the face acquires a more distinguishable shape as the mouth, nostrils, and ears begin to form. There is still no exact timeline at which the outer ear is created, but it is accepted that a cluster of embryonic cells connects to establish the ear. These are called auricular hillocks, which begin growing in the lower portion of the neck. The auricular hillocks broaden and intertwine within the seventh week to deliver the ear's shape. Within the ninth week, the hillocks move to the ear canal and are more noticeable as the ear [1]. The external anatomy of the ear can be seen in Figure 1. The growth of the ear in the first four months after birth is linear, and the ear is then stretched in development between the ages of four months and eight years. After this, the ear size and shape are constant until age seventy, increasing in size again.

Biometrics is the recognition of a human using their biometric characteristics, which may be physiological or behavioral. The physiological biometric features are the DNA, face, ear, facial, iris, fingerprint, hand geometry, hand vein, and palm print, and behavioral biometrics are signatures, gait patterns, and keystrokes. Voice is considered as a combination of biometric and physiological characteristics. Numerous systems have been developed to distinguish biometric traits, which have been used in multiple applications, such as forensic investigations and security systems. With the current worldwide pandemic, facial identification has failed due to users wearing masks. However, the human ear has proven more suitable as it is visible. In Table 1, an investigation was done to ascertain the performance, distinctiveness, permanence, collectability, and acceptability of the biometric.

In different physiological biometric qualities, the ear has received much consideration of late as it tends to be said that it is a solid biometric for human acknowledgment [2]. Ear biometric framework is dependable as it does not change and is of uniform tone, and its position is fixed at the center of the face's side. The size of an individual's ear is more critical than a unique finger impression and makes it simpler to capture an image of the subject without necessarily needing to gain information from the subject [2]. There are numerous difficulties in correctly gauging the details of the ear, and these are concealment of the ear by clothing, hair, ear ornaments, and jewelry. Another interference could be

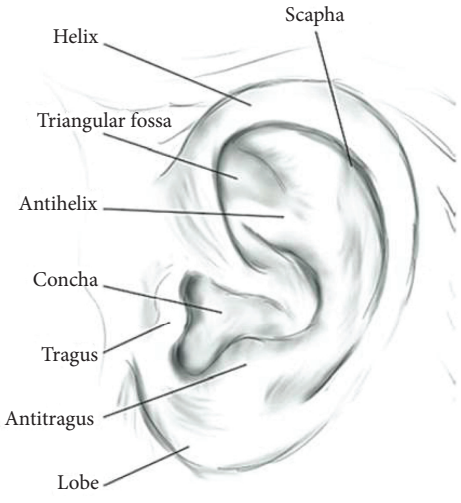


Figure 1: Diagram of the outer ear.

the different angles that the image was taken, concealing essential characteristics of the ear's anatomy. These difficulties have made ear recognition a secondary role in identification systems and techniques commonly used for identification and verification.

Although several computer-aided detection models have been developed to identify ears, low accuracy and sensitivity are still significant concerns that misidentify ears. Existing models are also computationally complex and expensive. The contributions of this work are summarized as follows:

(1) Implementation of state-of-the-art EfficientNets to develop an effective and inexpensive ear detection system. It is the first time the EfficientNet model is being applied to classify ears.

(2) The proposed model accuracy through EfficientNet.

(3) Finally, benchmark datasets were used to evaluate the performance of the model.

The remainder of the work is structured as follows: Section 2 presents related works, and Section 3 presents detailed data and methodology explored in this study. The experimental results and discussion are provided in Section 4, and Section 5 concludes the paper.

2. Related Work

This section presents different algorithms using the convolutional neural network (CNN) for ear identifications, and a summary of the related works is shown in Table 2.

Emeršič et al. [3] organized the dataset of the UERC which was used for the benchmark, training, and testing sets. In the completion, it was seen that handcrafted feature extraction methods, such as LBP [13] and patterns of oriented edge magnitudes (POEM) [14], and CNN-based feature extraction methods were used to obtain the ear identification. The challenges were to find methods to remove occlusions such as earrings, hair, other obstacles, and background from the ear image. The occlusion was done by creating a binary ear mask, and then the system recognition was done using the handcrafted features. Another proposed approach was to calculate the score of matrices from the CNN-based features and handcrafted features when they are fused, and a 30% detection rate was achieved.

Tian and Mu [4] applied a CNN to ear recognition in which they designed a CNN—it was made up of three convolutional layers, a fully connected layer, and a softmax classifier. The database used was USTB ear, which consisted of 79 subjects with various pose angles. The images utilized excluded earrings, headsets, or similar occlusions. Chowdhury et al. [15] proposed an ear biometric recognition system that uses local features of the ear and then uses a neural network to identify the ear. The method estimates where the ear could be in the input image and then takes the edge features from the identified ear. After identifying the ear, a neural network matches the extracted feature with a feature database. The databases used in this system were AMI, WPUT, IITD, and UERC, which achieved an accuracy of 70.58%, 67.01%, 81.98%, and 57.75%, respectively.

Raveane et al. [5] presented that it is difficult to precisely detect and locate an ear within an image, this challenge increases when working with the variable condition, and this could also be because of the odd shape of the human ears as well as lighting conditions and the changing profile shape of an ear when photographed [5]. The ear detection system used multiple CNNs, combined with a detection grouping algorithm, to identify an ear's presence and location. The proposed method matches other methods' performance when analyzed against clean and purpose-shot photographs, reaching an accuracy of upward of 98%. It outperforms them with a rate of over 86% when the system is subjected to non-cooperative natural images where the subject appears in challenging orientations and photographic conditions.

Multiple scale faster region-based convolutional neural network (Faster R-CNN) to detect ears from 2D profile images was proposed by Zhang and Mu [6]. This method was used by taking three regions of different scales that are detected to defer the information from the ear location within the context of the ear in the image, which was done to extract the ear correctly. The system was tested with 200 web images that achieved a 98% accuracy. Other experiments conducted were on the Collection J2 of the University of Notre Dame Biometrics Database (UND-J2) and University of Beira Interior Ear dataset (UBEAR); these achieved a detection rate of 100% and 98.22%, respectively, but these datasets contained large occlusions, scale, and pose variation.

Kohlakala and Coetzer [7] presented semi-automated and fully automated ear-based biometric verification systems. CNN and morphological postprocessing manually identify the ear region. It is used to classify ears in either the foreground or background of the image. The binary contour image applied the matching for feature extraction, and this was done by implementing a Euclidean distance measure, which had a ranking to verify for authentication. The Mathematical Analysis of Images Ear database and the Indian Institute of Technology Delhi Ear database were two databases, which achieved 99.20% and 96.06%, respectively.

Table 1: Summary of biometric characteristics.

Biometric identifier	Biometric type	Distinctiveness	Permanence	Collectability	Performance	Acceptability
DNA	Physiological	High	High	Low	High	Low
Ear	Physiological	Medium	High	Medium	Medium	High
Face	Physiological	Low	Medium	High	Low	High
Facial	Physiological	High	Low	High	Medium	High
Fingerprint	Physiological	High	High	Medium	High	Medium
Gait	Behavioral	Low	Low	High	Low	High
Hand geometry	Physiological	Medium	Medium	High	Medium	Medium
Hand vein	Physiological	Medium	Medium	Medium	Medium	Medium
Iris	Physiological	High	High	Medium	High	Low
Keystroke	Behavioral	Low	Low	Medium	Low	Medium
Odor	Physiological	High	High	Low	Low	Medium
Palm print	Physiological	High	High	Medium	High	Medium
Retina	Physiological	High	Medium	Low	High	Low
Signature	Behavioral	Low	Low	High	Low	High
Voice	Combination of physiological and behavioral	Low	Low	Medium	Low	High

Table 2: Summary of the related works.

Author	Dataset	Accuracy	Summary
Emeršič et al. [3]	NA	30	It was a handcrafted feature extraction method, such as LBP and patterns of oriented edge magnitudes (POEM), and CNN-based feature extraction methods were used to obtain the ear identification
Tian and Mu [4]	AMI, WPUT, IITD, and UERC	70.58, 67.01, 81.98, and 57.75	This system used deep convolutional neural network (CNN) to ear recognition. There were occlusions like no earrings, headsets, or similar occlusions
Raveane et al. [5]	NA	98	This system used variable conditions, and this could also be because of the odd shape of the human ears and changing lighting conditions
Zhang and Mu [6]	Notre Dame Biometrics database and University of Beira Interior Ear dataset	100 and 98.22	This system contained large occlusions, scale, and pose variation
Kohlakala and Coetzer [7]	Mathematical Analysis of Images Ear database and Indian Institute of Technology Delhi Ear database	99.2 and 96.06	It is used to classify ears in either the foreground or background of the image. The binary contour image applied the matching for feature extraction, and this was done by implementing a Euclidean distance measure, which had a ranking to verify for authentication
Tomczyk and Szczepaniak [8]	NA	NA	It shows the published experimental results that the approach did the rotation equivalence property to detect rotated structures
Hammam et al. [9]	Three ear datasets but not stated	22	The paper took seven performing handcrafted descriptors to extract the discriminating ear image. They then took the extracted ear and trained it using support vector machines (SVM) to learn a suitable model
Alkababji and Mohammed [10]	NA	97.8	It used the principal component analysis (PCA) and a genetic algorithm for feature reduction and selection
Jamil et al. [11]	Very underexposed or overexposed database	97	They considered that their work was the first to test the performance of CNN on very underexposed or overexposed images
Hansley et al. [12]	UERC challenge	NA	This was done using handcrafted descriptors, which were fused to improve recognition

Geometric deep learning (GDL) generalizes CNNs to non-Euclidean domains, presented by [8] Tomczyk and Szczepaniak. It used the convolutional filters with a mixture of Gaussian models. These filters were used so that the images could be easily rotated without interpolation. It shows the published experimental results that the approach

did the rotational equivalence property to detect rotated structures. Still, it does not need labor-intensive training on all rotated and nonrotated images.

Alshazly et al. [9] presented and compared ear recognition models built with handcrafted and CNN features. The paper took seven performing handcrafted descriptors to extract the discriminating ear image. They then took the extracted ear and trained it using support vector machine (SVM) to learn a suitable model. They then used CNN-based models, which used a variant of the AlexNet architecture. The results obtained on three ear datasets showed the CNN-based models' performance increased by 22%. This paper also investigated if the left and right ears have symmetry. The results obtained by the two datasets indicate a high impact of balance between the ears.

Alkababji and Mohammed [10] presented the use of a deep learning item detector called faster region-based convolutional neural network (Faster R-CNN) for ear detection. This CNN is used for feature extraction. It used the principal component analysis (PCA) and a genetic algorithm for feature reduction and selection. It also used a connected artificial neural network as the matcher. The results achieved an accuracy of 97.8% success rate.

Jamil et al. [11] build and train a CNN model for ear biometrics in various uniform illuminations measured using lumens. They considered that their work was the first to test the performance of CNN on underexposed or overexposed images. The results showed that for images with uniform illumination with a luminance of above 25 lux achieved a result of 100%. The CNN model had problems recognizing images when the lux was below ten, but produced an accuracy of 97%. This result shows that CNN architecture performs just as well as the other systems. It was found that the dataset had rotations which affected the results.

Hansley et al. [12] presented an unconstrained ear recognition framework that was better than the current state-of-the-art systems using publicly available databases. They developed CNN-based solutions for ear normalization and description. This was done using handcrafted descriptors, which were fused to improve recognition. This was done in two stages. The first stage was to find the landmark detectors, which were untrained scenarios. The next step was to generate a geometric image normalization to boost the performance. It was seen that the CNN descriptor was better than other CNN-based works in the literature. The obtained results were higher than different reported results for the UERC challenge.

3. Data and Methods

3.1. Dataset. In this study, all the experiments were performed with numerous public ear datasets; an explanation of these datasets is provided below. UBEAR, EarVN1.0, IIT, ITWE, and AWE databases are best suited for ear identification due to their large data size. However, it shows that EarVN1.0 has the foremost prominent usage during age estimation using CNN techniques. It is an appropriate dataset for ear images taken in a controlled environment, while ITWE is compatible for classifying ears in an uncontrolled environment, a summary of the datasets is shown in Table 3.

3.1.1. Mathematical Analysis of Images (AMI) Ear Database. The AMI Ear database [19] was collected at the University of Las Palmas. The database comprises 700 ear images of 100 distinct Caucasian male and female adults between the ages of 19 and 65. All images within the database were taken under an equivalent illumination and a glued camera position. Both the left- and right-hand sides of the ears were captured. The pictures obtained were cropped to form the ear area covering almost half the image. The pose of the images varies in yaw and servery in pitch angles, and this dataset is often found publicly.

3.1.2. The Indian Institute of Technology (IIT) Delhi Ear Database. The IIT database [16] was collected by the Indian Institute of Technology Delhi in New Delhi between October 2006 and June 2007. The database is formed from 421 images of 121 distinct adults of both male and female. All images were taken inside the environment, with no significant occlusions present, and only the right-hand side of the ear was captured. The pictures obtained in the dataset were both raw and normalized. The normalized images were in grayscale and of size 272×204 pixels.

3.1.3. The University of Beira Ear (UBEAR) Database. The University of Beira presented the UBEAR database [25]. The database comprises 4429 images of 126 subjects, and these were of both males and females. The images were taken under varying lighting conditions, and angles and partial occlusions were present. These images were of both the left- and right-hand sides of the ear.

3.1.4. The Annotated Web Ear (AWE) Database. The AWE database [18] is a set of public figures from web images. The database was formed from 1000 images of 100 different subjects whose sizes vary and were tightly cropped. Both the left- and right-hand sides of the ears were taken.

3.1.5. EarVN1.0. The EarVN1.0 database [22] comprises 28412 images of 164 Asian male and female subjects, and left- and right-hand sides of the ear were captured. Collection was during 2018 and is formed from unconstrained conditions, including camera systems and lighting conditions. The pictures are cropped from facial images to obtain the ears, and the pictures have significant variations in pose, scale, and illumination.

3.1.6. The Western Pomeranian University of Technology Ear (WPUTE) Database. The Western Pomeranian University of Technology Ear (WPUTE) database [20] was obtained within the year 2010 to gauge the ear recognition performance for images obtained within the wild. The database contains 2071 ear images belonging to 501 subjects. The images were of various sizes and were of both the left- and

Table 3: Summary of datasets.

	Database	Year	Number of subjects	Number of images	Left ear count	Right ear count	Total ears	Image size	Country	Sides
1	Institute of Technology Delhi Ear Database (IIT Delhi-I) [16]	2007	121	471		471	471	272 × 204	India	Right
	Institute of Technology Delhi Ear Database (IIT Delhi-II) [16]	NA	221	793		793	793	272 × 204	India	Right
2	The University of Science and Technology Beijing (USTB ear I) [17]	2002	60	185		185	185	Varied	China	Right
	The University of Science and Technology Beijing (USTB ear II) [17]	2004	77	308		308	308	Varied	China	Right
3	The Annotated Web Ears database (AWE) [18]	2016	100	1000	500	500	1000	Varied	Slovenia	Both
	The Annotated Web Ears database extended (AWE extend) [18]	2017	346	4104	2052	2052	4104	Varied	Slovenia	Both
4	Mathematical Analysis of Images Ear database (AMI) [19]	NA	106	700	420	280	700	492 × 702	Spain	Both
5	The West Pomeranian University of Technology Ear database (WPUTE) [20]	2010	501	2071	829	1242	2071	Varied	Poland	Both
6	Unconstrained Ear Recognition Challenge database (UERC) [21]	2017	3706	11804	5902	5902	11804	Varied	Slovenia	Both
7	EarVN1.0 [22]	2018	164	28412	14206	14206	28412	Varied and low resolution	Vietnam	Both
8	The In-the-Wild Ear database (ITWE) [23]	2015	55	605	424	181	605	Varied	Slovenia	Both
9	The Carreira-Perpinan (CP) [24]	1995	17	102	102		102	Varied	NA	Left
10	The University of Beira Ear Database (UBEAR) [25]	2011	126	4430	2215	2215	4430	1280 × 960	Mozambique	Both
11	Indian Institute of Technology Kanpur (IITK) [26]	2011	801	190	95	95	190	Varied	India	Both
12	The Forensic Ear Identification Database (FEARID) [27]	2005	1229	1229	615	614	1229	Varied	UK, Italy, and Netherlands	Both
13	University of Notre Dame (UND) [28]	2006	3480	952	952		952	Varied	France	Left
14	The Face Recognition Technology database $FERET) [29]	2010	9427	4745	3796	949	4745	Varied	Spain	Both
15	The Pose, Illumination and Expression (PIE) [30]	2002	40000	68	34	34	68	Varied	USA	Both
16	The XM2VTS Ear database [31]	NA	2360	295	89	206	295	720 × 576	UK	Both
17	The West Virginia University (WVU) [32]	2006	460	402	402		402	Varied	USA	Left

right-hand sides of the ear, and these were taken under different indoor lighting conditions and rotations. There were some occlusions included in the database, and these were the headset, earrings, and hearing aids.

3.1.7. The Unconstrained Ear Recognition Challenge (UERC). The Unconstrained Ear Recognition Challenge (UERC) database [21] was obtained in 2017, then extended in 2019, and is a mix of two databases that currently exist and a newly

created one. The database contains 3706 subjects with 11804 ear images, and the database ears have both right- and left-hand side images.

3.1.8. In-the-Wild Ear (ITWE) Database. The In-the-Wild Ear (ITWE) database [23] was created for recognition evaluation and has 2058 total images, and 231 male and female subjects. A boundary box obtained these images of the ear, and coordinates of those boundary boxes were released with the gathering. The pictures contained cluttering backgrounds and were of variable size and determination. The database includes both left- and right-hand sides of the ear, but no differentiation was given about the ears.

3.1.9. The University of Science and Technology Beijing (USTB) Ear Database. The University of Science and Technology Beijing (USTB) Ear database [17] contained cropped ear and head profile images of male and female subjects split into four sets. Dataset one includes 60 subjects and has 180 images of right close-up ears during 2002. These images were taken under different lightings and experienced some shearing and rotation. Dataset two contains 77 subjects, has 308 images of the right-hand side ear approximately 2 meters away from the ear, and these images were taken in 2004. These images were taken under different lighting conditions. Dataset three contains 103 subjects and has 1600 images, and these images were taken during the year 2004. The images are on the proper and left rotation, and therefore, the images are of the dimensions 768×576 pixels. The dataset contains 25500 images of 500 subjects; these were obtained from 2007 to 2008; the subject was in the center of the camera circle. The images were taken when the subject looked upward, downward, and at eye level. The images during this dataset contained different yaw and pitch poses. The databases are available on request and accessible for research.

3.1.10. The Carreira-Perpinan (CP) Ear Database. The Carreira-Perpinan (CP) [24] Ear database is an early dataset of the ear utilized for ear recognition systems. It was created in 1995 and contained 102 images with 17 subjects. The images were captured in a controlled environment, and therefore, the images include variability in minor pose variation.

3.1.11. The Indian Institute of Technology Kanpur (IITK) Ear Database. The Indian Institute of Technology Kanpur (IITK) is an ear database [26] that the Institute of Technology of Kanpur compiled. The database is split into three sets, and the first set consists of 190 male and female subjects of profile images. The total number of images was 801. The second dataset also contained 801, and with a total of 89 subjects, these images had variations in pitch angle. The third dataset contains 1070 images of an equivalent of 89 subjects, but with a variation in yaw and angle.

3.1.12. The Forensic Ear Identification Database (FEARID). The Forensic Ear Identification Database (FEARID) [27] is different from other databases as it only includes the ear prints. These contain no occlusions, variable angles, or illumination. Though there is no mention of any variables, other influences like the force the ear was pressed against the scanner and the scanner's cleanliness need to be considered. This database comprised 7364 images of 1229 subjects. This database was used for forensic application and not for biometric use.

3.1.13. The University of Notre Dame (UND) Database. The University of Notre Dame (UND) database contains [28] many subsets of 2D and 3D ear images. These images were appropriated over a period from 2003 to 2005. The database contains 3480 3D images from 952 male and female subjects and 464 2D images from 114 male and female subjects. These images were taken in different lighting conditions, yaw, pitch poses, and angles. The images are only of the left-hand side of the ear.

3.1.14. The Face Recognition Technology (FERET) Database. The Face Recognition Technology (FERET) database [29] is a sizeable facial image database and was obtained between the years 1995 to 1996. It contains 1564 subjects and has a total of 14126 images. These images were collected for face recognition and were of the left- and right-hand profile images, which made them perfect for 2D ear recognition.

3.1.15. The Pose, Illumination and Expression (PIE). Carnegie Mellon University obtained the Pose, Illumination and Expression database [30], which contains 40000 images and 68 subjects. The images are of the facial profile and have different poses, illuminations, and expressions.

3.1.16. The XM2VTS Ear Database. The XM2VTS Ear database [31] is frontal and profile facial images from the University of Surrey; the database contains 295 subjects and 2360 images captured during controlled conditions. These images were a set of cropped images of 720×576 pixel size and were from video data.

3.1.17. The West Virginia University (WVU) Ear Database. The West Virginia University (WVU) Ear database [32] is a video database and is formed from 137 subjects. The system was an advanced capturing procedure that allowed them to capture the ear at different angles; these images included earrings and eyeglasses.

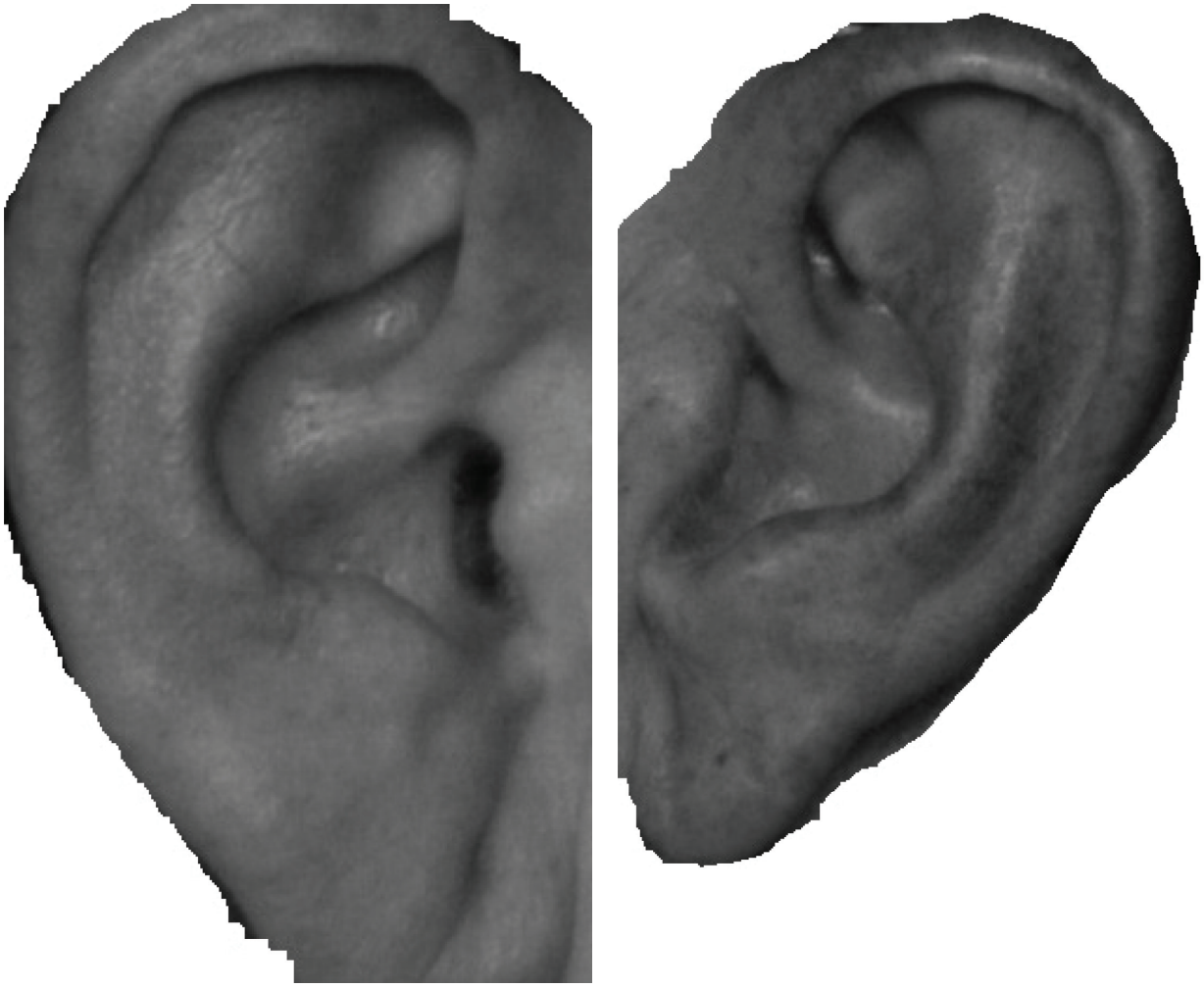
3.2. Preprocessing. Image preprocessing is a considerable part of the deep learning task. Most CNN models generally require a large dataset to learn to discriminate features suitably for making predictions and obtaining a good performance. As images in the datasets are of different sizes, the inputted images need to be resized to conform to all the other CNN models, but the features need to be preserved when resizing is performed. The examples of the original and the preprocessed images are shown in Figures 2 and 3.

3.3. Transfer Learning. In this study, the concept of transfer learning was adopted and helped with the pretrained CNN model for large datasets to learn features of the target (right

(a) (b)

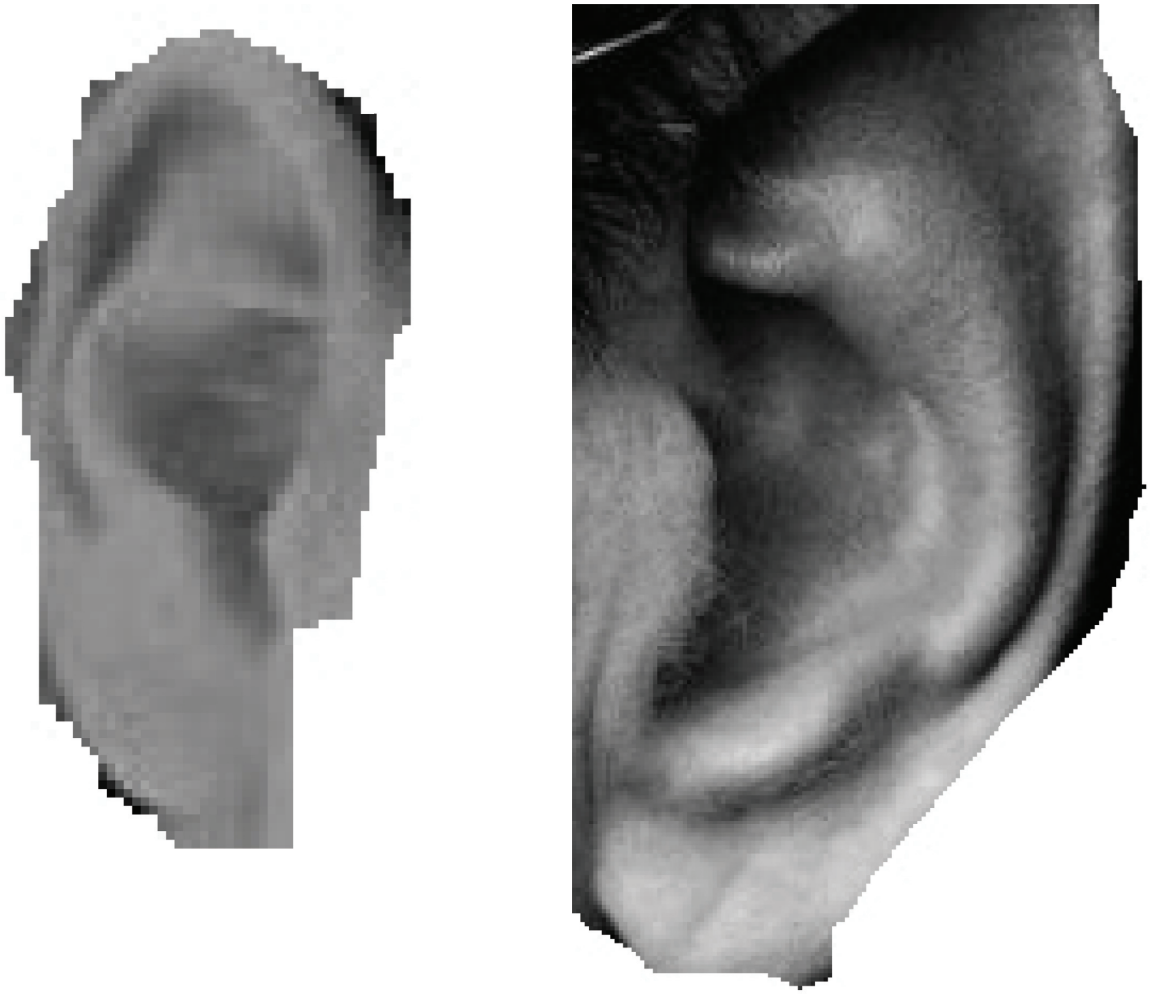
(c) (d)

Figure 2: Examples of original ear images. (a) Example of a 2D profile image for a female. (b) Example of a 2D profile image for a male. (c) Example of a facial image for a female. (d) Example of a facial image for a male.



(a) (b)

Figure 3: Continued.



(c) (d)

Figure 3: Examples of extracted ear images. (a) Example of ear extracted from 2D profile image for a female. (b) Example of ear extracted from 2D profile image for a male. (c) Example of ear extracted from facial image for a female. (d) Example of ear extracted from facial image for a male.

and left ears). It will transfer the features learned by the deep CNN models on other CNN models to this dataset. The number of deep CNN model parameters increases as the network gets deeper, which is used to achieve improved efficiency.

Hence, it requires many datasets for training, making it computationally complex and applying these models directly on small and new dataset results in feature extraction bias, overfitting, and poor generalization. The pretrained CNN modified and fine-tuned its structure to suit the dataset given. This concept of transfer learning is computationally expensive, has less training time, overcomes limitations of the dataset, improves performance, and is faster than training a model from the beginning. The pretraining CNN model fine-tuned in this work is the EfficientNets. The proposed structure is represented in Figure 4.

3.4. EfficientNet Architecture. EfficientNet is a lightweight model based on the auto machine learning framework to develop a baseline EfficientNet B0 network and uniformly scaled up the depth, width, and resolution using a simplified and effective compound coefficient to improve EfficientNet models B1–B8. The models performed efficiently and attained superiority over the existing CNN models on the other CNN datasets. EfficientNets are smaller and only require a few parameters, and they are faster and more generalizable to obtain higher accuracy on other datasets' poplar for the transfer learning task. The proposed study fine-tuned EfficientNet models B0–B8 on the dataset to detect the ears. In transferring the pretrained EfficientNets to the ear dataset, the models were fine-tuned by adding a global average pooling to reduce the number of parameters and fix overfitting. The dense layers follow the global average

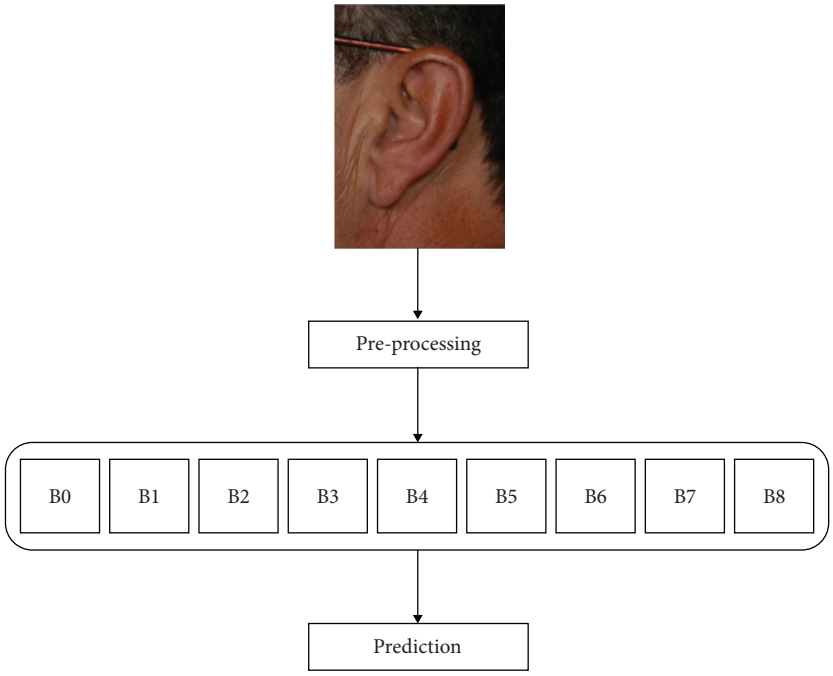


Figure 4: Block structure of the proposed model.

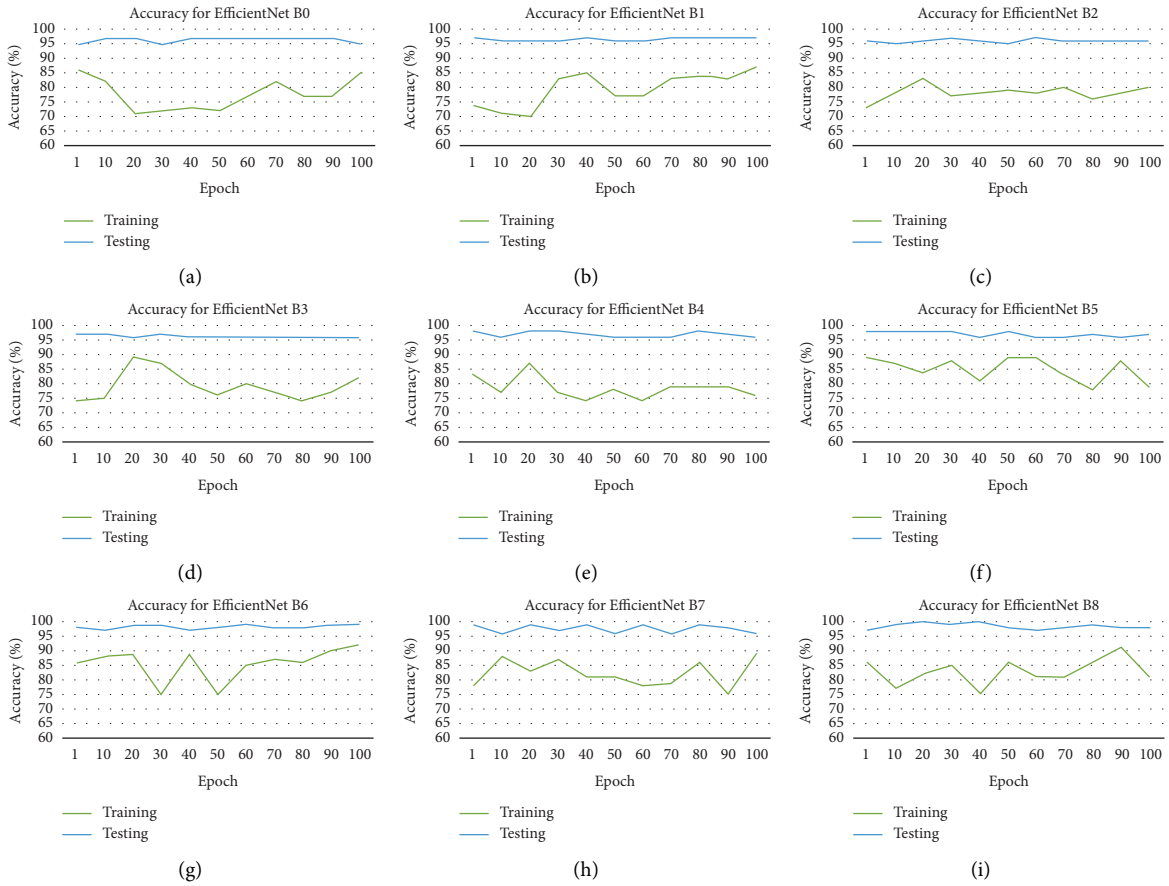


Figure 5: Accuracy for the ear dataset of each EfficientNet. (a) Accuracy for EfficientNet B0. (b) Accuracy for EfficientNet B1. (c) Accuracy for EfficientNet B2. (d) Accuracy for EfficientNet B3. (e) Accuracy for EfficientNet B4. (f) Accuracy for EfficientNet B5. (g) Accuracy for EfficientNet B6. (h) Accuracy for EfficientNet B7. (i) Accuracy for EfficientNet B8.

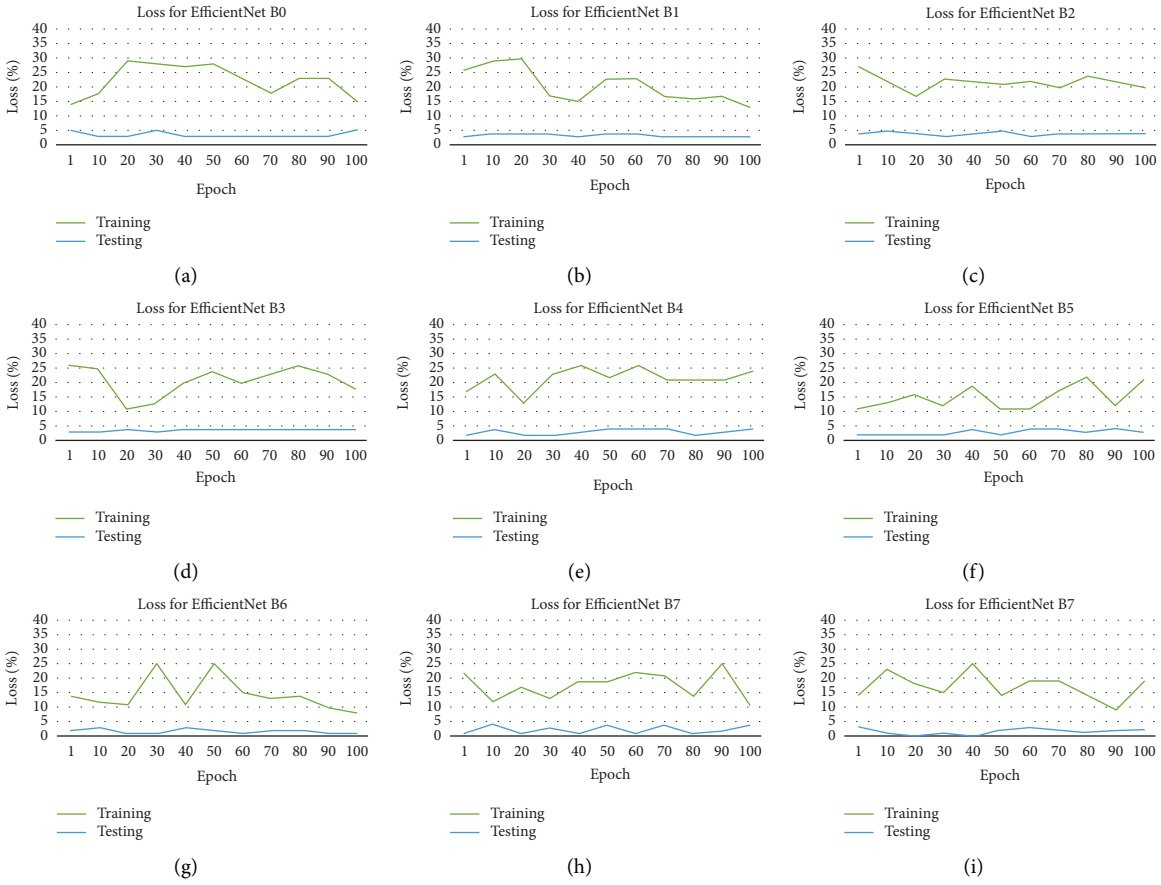


Figure 6: Loss for the ear dataset of each EfficientNet. (a) Loss for EfficientNet B0. (b) Loss for EfficientNet B1. (c) Loss for EfficientNet B2. (d) Loss for EfficientNet B3. (e) Loss for EfficientNet B4. (f) Loss for EfficientNet B5. (g) Loss for EfficientNet B6. (h) Loss for EfficientNet B7. (i) Loss for EfficientNet B8.

pooling with a ReLU activation function and a dropout rate of 0.4 before the output last layer [33]. This is done with the softmax activation function to determine the probabilities of the input data to represent the ears, and this can be seen in

$$\sigma(q)_i = \frac{e^{q_i}}{\sum_N^{y=1} e^{q_y}}, \tag{1}$$

where σ is the softmax activation function, q represents the input vector to the output layer, i is depicted from the exponential element e^{q_i}, N is the number of classes, and e^{q_y} represents the output vector of the exponential function.

It is known that many iterations could lead to model overfitting, while too few can cause model underfitting; this study used an early stopping strategy. It configured approximately 90 training iterations before terminating, this was to cater for early stopping to improve performance, and this was applied to control overfitting and used gradient descent. The EfficientNet B0-B8 models were trained with 100 iterations (epochs). The batch size for each iteration was 32, and the momentum equals 0.2 and was regulated. At the same time, categorical cross-entropy is the loss function used to update weights at each iteration. Hyperparameters used were evaluated and found to perform optimally, and this can be defined in

$$\alpha = \alpha - n \cdot \triangle_\alpha J\left(\alpha; x^i; y^i\right), \tag{2}$$

where $\triangle_\alpha J$ is the gradient of the loss with regard to α, n is the defined learning rate, α is the weight vector, while x and y are the respective training sample and label.

4. Results and Discussion

Various EfficientNet variants were fine-tuned on all the ear datasets to detect the ear. Each dataset is split into 20% training and 80% test sets. The experiments were entirely performed using Keras deep learning framework using the TensorFlow backend. The models were evaluated using the popular evaluation metrics, equation (3)–(7) (accuracy, sensitivity, specificity, and area under the curve). The performances of all experiments are evaluated by using a series of confusion matrix-based performance metrics.

The confusion matrices are used to evaluate the classifiers, with true positives (TPs) representing the ears that are correctly classified as positive, true negatives (TNs) representing the ears that are correctly classified as negative, false positives (FPs) representing the ears that are incorrectly classified as positive, and false negatives (FNs) representing the ears being incorrectly classified as negative.

Table 4: Performance of EfficientNet models.

Epoch	EfficientNet B0		EfficientNet B1		EfficientNet B2		EfficientNet B3		EfficientNet B4		EfficientNet B5		EfficientNet B6		EfficientNet B7		EfficientNet B8	
	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss
1	95	5	97	3	96	4	97	3	98	2	98	2	98	2	99	1	97	3
10	97	3	96	4	95	5	97	3	96	4	98	2	97	3	96	4	99	1
20	97	3	96	4	96	4	96	4	98	2	98	2	99	1	99	1	100	0
30	95	5	96	4	97	3	97	3	98	2	98	2	99	1	97	3	99	1
40	97	3	97	3	96	4	96	4	97	3	96	4	97	3	99	1	100	0
50	97	3	96	4	95	5	96	4	96	4	98	2	98	2	96	4	98	2
60	97	3	96	4	97	3	96	4	96	4	96	4	99	1	99	1	97	3
70	97	3	97	3	96	4	96	4	96	4	96	4	98	2	96	4	98	2
80	97	3	97	3	96	4	96	4	98	2	97	3	98	2	99	1	99	1
90	97	3	97	3	96	4	96	4	97	3	96	4	99	1	98	2	98	2
100	95	5	97	3	96	4	96	4	96	4	97	3	99	1	96	4	98	2

4.1. Specificity. It is the ratio of correctly classified negative instances by a model to the overall number of true-negative instances being tested, equation (5).

4.2. Accuracy. It is a measure that indicates the ratio of all the correctly recognized cases to the overall number of cases. While this metric generally gives a decent reflection of the classifier, it may not reflect a classifier's true performance in a scenario where there is an uneven class distribution. Accuracy can be computed using the following formula, equation (3).

4.3. Sensitivity. It is the ratio of all correctly classified positive instances by a model to the overall number of positive classifications by a model. A low precision indicates that a model suffers from high false positives. Precision can be computed using the following formula, equation (4).

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN'}, \tag{3}$$

$$\text{sensitivity} = \frac{TP}{TP + FN'}, \tag{4}$$

$$\text{specificity} = \frac{TN}{TN + FP'}, \tag{5}$$

$$\begin{aligned} TPR &= \text{sensitivity} \\ &= \frac{TP}{TP + FN'}, \end{aligned} \tag{6}$$

$$\begin{aligned} FPR &= 1 - \text{sensitivity} \\ &= \frac{FP}{FP + TN'}. \end{aligned} \tag{7}$$

The results obtained are presented in Figures 5 and 6 this is the accuracy and loss of these datasets. The various EfficientNet models average at the 100 epochs, and the accuracy is determined using the test set. The models performed at extracting and learning discriminative features from the dataset. EfficientNet B8 attains the best accuracy 98.45%, and the EfficientNet results are noted in Table 4.

An advantage of EfficientNets is that they are smaller with fewer parameters and faster, and obtain transfer learning successfully from the datasets. The worst performing EfficientNet is B2, as shown in Table 4. Even though it has minimal parameters, the reason that this performed poorly could have been because the images were downsampled. This was done to conform to the model's image input size. It can be seen that performance improves as the model gets deeper. EfficientNet B0 started poorly, beginning to converge from the 30 iteration, with little noise, until the 30 iteration and then stabilized until 50 iteration, when overfitting started. The best performing EfficientNet is B8, as shown in Table 4, and this is because of the large number of parameters. It began to converge from the 60 iteration and then stabilized until 90 iteration, when overfitting started. It is found that when the dataset is a large and equal number of classes, the results achieved were high. Determining the most suitable hyperparameters was one of the challenges faced and the overfitting, which was limited due to the data samples. The results of the proposed methods compared with related studies are presented in Figure 7.

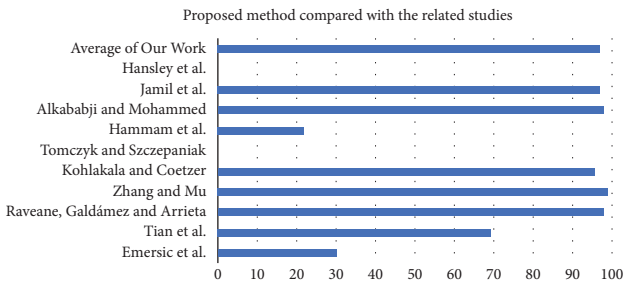


Figure 7: Proposed method compared with the related studies.

Table 5: Proposed method compared with the related studies.

Authors	Result
Emeršič et al. [3]	30
Tian and Mu [4]	69.33
Raveane et al. [5]	98
Zhang and Mu [6]	99.11
Kohlakala and Coetzer [7]	95.63
Tomczyk and Szczepaniak [8]	NA
Alshazly et al. [9]	22
Alkababji and Mohammed [10]	97.8
Jamil et al. [11]	97
Hansley et al. [12]	NA
Average of our work	97.07

5. Conclusion

This study investigated and implemented EfficientNet models to automatically identify ears on the most prominent and publicly available datasets. EfficientNets that achieved state-of-the-art performance over other architectures to maximize accuracy and efficiency were explored and fine-tuned on profile images. The fine-tuning technique is valuable to utilize rich generic features learned from significant dataset sources such as ImageNet to compliment the lack of annotated datasets affecting ear domains. The experimental results show the effectiveness of EfficientNets in extracting and learning distinctive features from the ear images and then classifying them into a left or right suitable class. Out of the nine EfficientNet variants explored in this study, the EfficientNet B8 outperformed the others, as evident in Table 5 and depicted in Figure 7. One of the significant downfalls of the proposed approach is training the model on small datasets and training on images with low resolutions. These limitations can easily result in significant overfitting. To overcome this, you need to have compelling image preprocessing techniques. Although the proposed methodology is specified to do ear detection, it could be extended to detect other parts of the face, given the right set of datasets.

Data Availability

Datasets used to support the findings of the study are publicly available.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, and M. S. Nixon, "A survey on ear biometrics," *ACM Computing Surveys*, vol. 45, no. 2, pp. 1–35, 2013.

[2] C. Bhanu, "Ear Biometrics," in *Advances in Intelligent Systems and Computing*Springer, Boston, MA, USA, 2009.

[3] Ž. Emeršič, D. Štepec, V. Štruc, and P. Peer, "Training Convolutional Neural Networks with Limited Training Data for Ear Recognition in the Wild," in *Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, USA, May 2017.

[4] L. Tian and Z. Mu, "Ear recognition based on deep convolutional network," in *Proceedings of the 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 437–441, IEEE, Datong, China, October 2016.

[5] W. Raveane, P. L. Galdamez, and M. A. Gonzalez Arrieta, "Ear detection and localization with convolutional neural networks in natural images and videos," *Processes*, vol. 7, no. 7, p. 457, 2019.

[6] Y. Zhang and Z. Mu, "Ear detection under uncontrolled conditions with multiple scale faster region-based convolutional neural networks," *Symmetry*, vol. 9, no. 4, p. 53, 2017.

[7] A. Kohlakala and J. Coetzer, "Ear-based biometric authentication through the detection of prominent contours," *SAIEE Africa Research Journal*, vol. 112, no. 2, pp. 89–98, 2021.

[8] A. Tomczyk and P. S. Szczepaniak, "Ear detection using convolutional neural network on graphs with filter rotation," *Sensors*, vol. 19, no. 24, p. 5510, 2019.

[9] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Handcrafted versus cnn features for ear recognition," *Symmetry*, vol. 11, no. 12, p. 1493, 2019.

[10] A. M. Alkababji and O. H. Mohammed, "Real time ear recognition using deep learning," *Telkomnika*, vol. 19, no. 2, pp. 523–530, 2021.

[11] N. Jamil, A. Almisreb, S. M. Z. S. Z. Ariffin, N. Md Din, and R. Hamzah, "Can Convolution Neural Network (Cnn) Triumph in Ear Recognition of Uniform Illumination Invariant?" *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, 2018.

[12] E. E. Hansley, M. P. Segundo, and S. Sarkar, "Employing fusion of learned and handcrafted features for unconstrained ear recognition," *IET Biometrics*, vol. 7, no. 3, pp. 215–223, 2018.

[13] Z.-q. Wang and X.-d. Yan, "Multi-scale feature extraction algorithm of ear image," in *Proceedings of the 2011 International Conference on Electric Information and Control Engineering*, pp. 528–531, IEEE, Wuhan, China, April 2011.

[14] N.-S. Vu, H. M. Dee, and A. Caplier, "Face recognition using the poem descriptor," *Pattern Recognition*, vol. 45, no. 7, pp. 2478–2488, 2012.

[15] D. P. Chowdhury, S. Bakshi, G. Guo, and P. K. Sa, "On applicability of tunable filter bank based feature for ear biometrics: a study from constrained to unconstrained," *Journal of Medical Systems*, vol. 42, no. 1, pp. 11–20, 2018.

[16] A. Kumar, "Iit delhi ear database version 1.0," 2007, https://webold.iitd.ac.in/biometrics/Database_Ear.htm.

[17] Y. Zhang, Z.-C. Mu, L. Yuan, C. Yu, and L. Qing, "USTB-Helloear: A Large Database of Ear Images Photographed Under Uncontrolled Conditions," *Image and Graphics*, vol. 12, Springer, New York, NY, USA, 2017.

[18] Ž. Emeršič, V. Štruc, and P. Peer, "Ear recognition: more than a survey," *Neurocomputing*, vol. 255, pp. 26–39, 2017.

[19] E. Gonzalez, L. Alvarez, and L. Mazorra, "Ami Ear Database," 2012, http://ctim.ulpgc.es/research_works/ami_ear_database/.

[20] D. Frejlichowski and N. Tyszkiewicz, "The West Pomeranian university of Technology Ear Database – a Tool for Testing Biometric Algorithms," *Image Analysis and Recognition*, Springer, Berlin, Germany, 2010.

[21] Ž. Emeršič, D. Štepec, V. Štruc et al., "The unconstrained ear recognition challenge," in *Proceedings of the 2017 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 715–724, IEEE, Denver, CO, USA, October 2017.

[22] V. T. Hoang, "Earvn1.0: a new large-scale ear images dataset in the wild," *Data in Brief*, vol. 27, Article ID 104630, 2019.

[23] v. Emeršič and P. Peer, "Ear biometric database in the wild," in *Proceedings of the 2015 4th International Work Conference on Bioinspired Intelligence (IWOBI)*, pp. 27–32, San Sebastian, Spain, June 2015.

[24] M. A. Carreira-Perpinan, "Compression Neural Networks for Feature Extraction: Application to Human Recognition from Ear Images," MSc thesis, Faculty of Informatics, Technical University of Madrid, Spain, , 1995.

[25] R. Raposo, E. Hoyle, A. Peixinho, and H. ProenĂsa, "Ubear: A Dataset of Ear Images Captured On-The-Move in Uncontrolled Conditions," in *Proceedings of the 2011 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, Paris, France, April 2011.

[26] S. Prakash, U. Jayaraman, and P. Gupta, "Connected component based technique for automatic ear detection," in *Proceedings of the 2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 2741–2744, IEEE, Cairo, Egypt, November 2009.

[27] I. Alberink and A. Ruifrok, "Performance of the fearid earprint identification system," *Forensic Science International*, vol. 166, no. 2-3, pp. 145–154, 2007.

[28] P. Yan and K. Bowyer, "Empirical evaluation of advanced ear biometrics," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, p. 41, September 2005.

[29] P. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

[30] T. Sim, S. Baker, and M. Bsat, "The Cmu Pose, Illumination, and Expression (Pie) Database of Human Faces," Tech. Rep. CMU-RI-TR-01-02, Carnegie Mellon University, Pittsburgh, PA, January 2001.

[31] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "Xm2vtsdb: the extended m2vts database," in *Proceedings of the Second international conference on audio and video-based biometric person authentication*, Citeseer, Washington, DC, USA, vol. 964, pp. 965-966, 1999.

[32] A. Abaza, *High Performance Image Processing Techniques in Automated Identification Systems*, West Virginia University, Morgantown, West Virginia, 2008.

[33] M. Oloko-Oba and S. Viriri, "Ensemble of efficientnets for the diagnosis of tuberculosis," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 9790894, 12 pages, 2021.

3.3 Transformation Network Model for Ear Recognition

3.3.1 Brief Overview

This section introduces a research paper whose main contribution is presenting the results of a CNN developed using Transfer Learning to pre-train the CNN before applying a Transformer Network. The performance achieved in this research shows the efficiency of the Transformer Network on ear recognition. The Transformer Network is an encoder-decoder architecture based on attention layers. The difference between a Convolutional Neural Network and a Transformer Network is that the data can be passed in parallel, which means that the GPU can be utilised effectively and efficiently.

The paper is published in the *Springer - Lecture Notes in Computer Science*.

Transformation Network Model for Ear Recognition

Aimee Booysens and Serestina Viriri( )

School of Mathematics, Statistics and Computer Science,
University of KwaZulu-Natal, Durban, South Africa
210501411@stu.ukzn.ac.za, viriris@ukzn.ac.za

Abstract. Biometrics is the recognition of a human using biometric characteristics for identification, which may be physiological or behavioural. The physiological biometric features are the face, ear, iris, fingerprint and handprint; behavioural biometrics are signatures, voice, gait pattern and keystrokes. Numerous systems have been developed to distinguish biometric traits used in multiple applications, such as forensic investigations and security systems. With the current worldwide pandemic, facial identification has failed due to users wearing masks; however, the human ear has proven more suitable as it is visible. This paper presents the main contribution to presenting the results of a CNN developed using Transfer Learning to pre-train the CNN before applying a Transformer Network. The performance achieved in this research shows the efficiency of the Transformer Network on ear recognition. The experiments showed that Transformer Network achieved the best accuracy of 92.60% and 92.56% with epochs of 50 and 90.

Keywords: Ear Biometrics · Ear Recognition · Transformer Network · Machine Learning

1 Introduction

The ear begins to develop in a fetus during the fifth and seventh weeks of pregnancy [2]. At this stage, the face acquires a more distinguishable shape as the mouth, nostrils, and ears begin to form. There is still no exact timeline at which the outer ear is created, but it is accepted that a cluster of embryonic cells connects to establish the ear. These are called auricular hillocks, which begin growing in the lower portion of the neck. The auricular hillocks broaden and inter and twine within the seventh week to deliver the ear's shape. Within the ninth week, the hillocks move to the ear canal and are more noticeable as the ear [2].

Biometrics is the recognition of a human using their biometric characteristics, which may be physiological or behavioural. The physiological biometric features are the DNA, face, ear, facial, iris, fingerprint, hand geometry, hand vein and palm print, and behavioural biometrics are signatures, and gait path-connected

É. Renault et al. (Eds.): MLN 2023, LNCS 14525, pp. 250–266, 2024.
https://doi.org/10.1007/978-3-031-59933-0_17

keystrokes. Voice is considered a combination of biometric and physiological characteristics. Numerous systems have been developed to distinguish biometric traits, which have been used in multiple applications, such as forensic investigations and security systems. With the current Worldwide pandemic, facial identification has failed due to users wearing masks. However, the human ear has proven more suitable as it is visible.

In different physiological biometric qualities, the ear has received much consideration of late as it tends to be said that it is a solid biometric for human acknowledgement [6] Ear biometric framework is dependable as it does not change, is of uniform tone, and its position is fixed at the centre of the face's side. The size of an individual's ear is more critical than a unique finger impression and makes it simpler to capture an image of the subject without necessarily needing to gain information from the subject, [6]. There are numerous difficulties in correctly gauging the details of the ear, these are concealment of the ear by clothing, hair, ear ornaments and jewellery. Another interference could be the different angle that the image was taken, concealing essential characteristics of the ear's anatomy. These difficulties have made ear recognition a secondary role in identification systems and techniques commonly used for identification and verification.

Although several computer-aided detection models have been developed to identify ears, low accuracy and sensitivity are still significant concerns that misidentify ears. Existing models are also computationally complex and expensive. In this paper, an ear recognition model based on Transformer Network is proposed.

The remaining work is structured as follows: Sect. 2 presents related works, and Sect. 3 presents detailed data and methodology explored in this study. The experimental results and discussion are provided in Sect. 4, and Sect. 5 concludes the paper.

2 Related Work

This section presents different algorithms using the Convolutional Neural Network (CNN) for ear identifications, a summary of the related works is shown in Table 1.

The competition Emeršič et al. [9] organised the dataset of the UERC, which was used for the bench-mark, training and testing sets. In the completion, it was seen that handcrafted feature extraction methods, such as LBP [29] and patterns of oriented edge magnitudes (POEM) [28], and CNN-based feature extraction methods were used to obtain the ear identification. The challenges were to find methods to remove occlusions such as earrings, hair, other obstacles, and background from the ear image. The occlusion was done by creating a binary ear mask, and then the system recognition was done using the handcrafted features. Another proposed approach was to calculate the score of matrices from the CNN-based features and handcrafted features when they are fused, a 30% detection rate was achieved.

Tian et al. [26] applied a CNN to ear recognition in which they designed a CNN - it was made up of three convolutional layers, a fully connected layer, and a softmax classifier. The database used was USTB ear, which consisted of 79 subjects with various pose angles. The images utilised excluded earrings, headsets, or similar occlusions. Chowdhury et al. [8] proposed an ear biometric recognition system that uses local features of the ear and then uses a neural network to identify the ear. The method estimates where the ear could be in the input image and then takes the edge features from the identified ear. After identifying the ear, a neural network matches the extracted feature with a feature database. The databases used in this system were AMI, WPUT, IITD, and UERC, which achieved an accuracy of 70.58%, 67.01%, 81.98%, and 57.75%.

Raveane, Galdámez and Arrieta [24] presented that it is difficult to precisely detect and locate an ear within an image, this challenge increases when working with the variable condition and this could also be because of the odd shape of the human ears as well as lighting conditions and the changing profile shape of an ear when photographed, [24]. The ear detection system used multiple CNN's, combined with a detection grouping algorithm, to identify an ear's presence and location. The proposed method matches other methods' performance when analysed against clean and purpose-shot photographs, reaching an accuracy of upwards of 98%. It outperforms them with a rate of over 86% when the system is subjected to non-cooperative natural images where the subject appears in challenging orientations and photographic conditions.

Multiple Scale Faster Region-based Convolutional Neural Networks (Faster R-CNN) to detect ears from 2D profile images was proposed by Zhang and Mu [32]. This method was used by taking three regions of different scales that are detected to defer the information from the ear location within the context of the ear in the image, which was done to extract the ear correctly. The system was tested with 200 web images that achieved a 98% accuracy. Other experiments conducted were on the Collection J2 of the University of Notre Dame Biometrics Database (UND-J2) and University of Beira Interior Ear dataset (UBEAR); these achieved a detection rate of 100% and 98.22%, respectively, but these datasets contained large occlusions, scale and pose variation.

Kohlakala and Coetzer [18] presented semi-automated and fully automated ear-based biometric verification systems. CNN and morphological postprocessing manually identify the ear region. It is used to classify ears in the image's foreground or background. The binary contour image applied the matching for feature extraction, and this was done by implementing a Euclidean distance measure, which had a ranking to verify for authentication. The Mathematical Analysis of Images ear database and the Indian Institute of Technology Delhi ear database were two databases, which achieved 99.20% and 96.06%, respectively.

Geometric deep learning (GDL) generalises CNNs to non-Euclidean domains, presented by [27] Tomczyk and Szczepaniak. It used convolutional filters with a mixture of Gaussian models. These filters were used so that the images could be easily rotated without interpolation. It shows the published experimental results that the approach did the rotational equivalence property to detect rotated

structures. Still, it does not need labour-intensive training on all rotated and non-rotated images.

Hammam et al. [5] presented and compared ear recognition models built with handcrafted and CNN features. The paper took seven, performing handcrafted descriptors to extract the discriminating ear image. They then took the extracted ear and trained it using Support Vector Machines (SVM) to learn a suitable model. They then used CNN-based models, which used a variant of the AlexNet architecture. The results obtained on three ear datasets showed the CNN-based models' performance increased by 22%. This paper also investigated if the left and right ears have symmetry. The results obtained by the two datasets indicate a high impact of balance between the ears.

Alkababji and Mohammed [4] presented the use of a Deep Learning item detector called faster region-based convolutional neural networks (Faster R-CNN) for ear detection. This CNN is used for feature extraction. It used the Principal Component Analysis (PCA) and a genetic algorithm for feature reduction and selection. It also used a connected artificial neural network as the matcher. The results achieved an accuracy of 97.8% success rate.

Jamil et al. [17] build and train a CNN model for ear biometrics in various uniform illuminations measured using lumens. They considered that their work was the first to test the performance of CNN on underexposed or overexposed images. The results showed that images with uniform illumination with a luminance of above 25 lux, achieved a result of 100%. The CNN model had problems recognising images when the lux was below ten, but produced an accuracy of 97%. This result shows that CNN architecture performs just as well as the other systems. It was found that the dataset had rotations which affected the results.

Hansley et al. [15] presented an unconstrained ear recognition framework that was better than the current state-of-the-art systems using publicly available databases. They developed CNN-based solutions for ear normalisation and description. This was done using a handcrafted descriptor. The published experimental results show This was done in two stages. The first stage was to find the landmark detectors, which were untrained scenarios. The next step was to generate a geometric image normalisation to boost the performance. It was seen that the CNN descriptor was better than other CNN-based works in the literature. The obtained results were higher than different reported results for the UERC challenge.

Table 1. Summary of the related works

Author	Dataset	Accuracy	Summary
Zhang and Mu [32]	Notre Dame Biometrics Database and University of Beira Interior Ear dataset	100 & 98.22	This system contained large occlusions, scale and pose variation.
Kohlakala and Coetzer [18]	Mathematical Analysis of Images ear database and Indian Institute of Technology Delhi ear database	99.2 & 96.06	It is used to classify ears either in the foreground or background of the image. The binary contour image applied the matching for feature extraction, and this was done by implementing a Euclidean distance measure, which had a ranking to verify for authentication.
Tomczyk and Szczepaniak [27]	NA	NA	It shows the published experimental results that the approach did the rotation equivalence property to detect rotated structures.
Hammam et al. [5]	Three ear datasets but not stated	22	The paper took seven performing handcrafted descriptors to extract the discriminating ear image. They then took the extracted ear and trained it using Support Vector Machines (SVM) to learn a suitable model.
Alkababji and Mohammed [4]	NA	97.8	It used the Principal Component Analysis (PCA) and a genetic algorithm for feature reduction and selection.
Jamil et al. [17]	Very underexposed or overexposed database	97	They considered that their work was the first to test the performance of CNN on very underexposed or overexposed images.
Hansley et al. [15]	UERC challenge	NA	This was done using handcrafted descriptors, which were fused to improve recognition.
Tian et al. [26]	AMI, WPUT, IITD, and UERC	70.58, 67.01, 81.98, & 57.75	This system used deep convolutional neural network (CNN) to ear recognition. There were occlusions like no earrings, headsets, or similar occlusions.
Raveane, Galdámez and Arrieta [24]	NA	98	This system used variable conditions and this could also be because of the odd shape of the human ears and changing lighting conditions.
Emeršič et al. [9]	NA	30	It was a handcrafted feature extraction methods, such as LBP and patterns of oriented edge magnitudes (POEM), and CNN-based feature extraction methods were used to obtain the ear identification.

3 Data and Methods

3.1 Dataset

In this study, all the experiments were performed with numerous public ear datasets an explanation of these datasets is provided below. UBEAR, EarVN1.0, IIT, ITWE and AWE databases are best suited for ear identification due to their large data size. However, it shows that EarVN1.0 has the foremost prominent usage during age estimation using CNN techniques. It is an appropriate dataset for ear images taken in a controlled environment, while ITWE is compatible with classifying ears in an uncontrolled environment, a summary of the datasets is shown in Table 3.

Mathematical Analysis of Images (AMI) Ear Database. The AMI ear database [14] was collected at the University of Las Palmas. The database comprises 700 ear images of 100 distinct Caucasian male and female adults between the ages of 19 and 65. All images within the database were taken under an equivalent illumination and a glued camera position. Both the left- and right-hand sides of the ears were captured. The pictures obtained were cropped to form the ear area covering almost half the image. The pose of the images varies in yaw and servery in pitch angles, and this dataset is often found publicly.

The Indian Institute of Technology (IIT) Delhi Ear Database. The IIT database [19] was collected by the Indian Institute of Technology Delhi in New Delhi between October 2006 and June 2007. The database is formed from 421 images of 121 distinct adults of both male and female. All images were taken inside the environment, with no significant occlusions present, and only the right-hand side of the ear was captured. The pictures obtained in the dataset were both raw and normalised. The normalised images were in greyscale and of size 272×204 pixels.

The University of Beira Ear Database (UBEAR). The University of Beira presented the UBEAR database [23]. The database comprises 4429 images of 126 subjects, and these were of both males and females. The images were taken under varying lighting conditions, angles and partial occlusions were present. These images were of both the left- and right-hand side of the ear.

The Annotated Web Ear Database (AWE). The AWE ear database [11] is a set of public figures from web images. The database was formed from 1000 images of 100 different subjects whose sizes vary and were tightly cropped. Both the left- and right-hand sides of the ears were taken.

EarVN1.0. The EarVN1.0 database [16] comprises, 28412 images of 164 Asian male and female subjects, left- and right-hand sides of the ear were captured. Collection was during 2018 and is formed from unconstrained conditions, including camera systems and lighting conditions. The pictures are cropped from facial images to obtain the ears, and the pictures have significant variations in pose, scale and illumination.

The Western Pomeranian University of Technology Ear Database (WPUTE). The Western Pomeranian University of Technology Ear Database WPUTE [13] was obtained in the year 2010 to gauge the ear recognition performance for images obtained within the wild. The database contains 2071 ear images belonging to 501 subjects. The images were of various sizes and were of both the left- and right-hand sides of the ear, these were taken under different indoor lighting conditions and rotations. There were some occlusions included in the database, these were the headset, earrings and hearing aids.

The Unconstrained Ear Recognition Challenge (UERC). The Unconstrained Ear Recognition Challenge (UERC) database [10] was obtained in 2017, then extended in 2019 and is a mix of two databases that currently exist and a newly created one. The database contains 3706 subjects with, 11804 ear images, and the database ears have both right- and left-hand side images.

In the Wild Ear Database (ITWE). The In the Wild Ear Database (ITWE) [12] was created for recognition evaluation and has, 2058 total images, 231 male and female subjects. A boundary box obtained these images of the ear, and the coordinates of those boundary boxes were released with the gathering. The pictures contained cluttering backgrounds and were of variable size and determination. The database includes both left- and right-hand sides of the ear, but no differentiation was given about the ears.

The University of Science and Technology, Beijing (USTB) Ear Database. The University of Science and Technology Beijing (USTB) Ear Database [31] contained cropped ear and head profile images of male and female subjects split into four sets. Dataset one includes 60 subjects and has 180 images of right close-up ears during 2002. These images were taken under different lighting and experienced some shearing and rotation. Data set two contains 77 subjects and has 308 images of the right-hand side ear approximately 2 m away from the ear and were taken in 2004. These images were taken under different lighting conditions. Dataset three contains 103 subjects and has 1600 images, these images were taken during the year 2004. The images are on the proper and left rotation, and therefore the images are of the dimensions 768×576 pixels. The dataset contains, 25500 images of 500 subjects; these were obtained from 2007 to 2008; the subject was in the centre of the camera circle. The images were taken when the subject looked upwards, downwards and at eye level. The images

during this dataset contained different yaw and pitch poses. The databases are available on request and accessible for research.

The Carreira-Perpinan (CP) Ear Database. The Carreira-Perpinan (CP) [7] ear database is an early dataset of the ear utilised for ear recognition systems. It was created in 1995 and contained 102 images with 17 subjects. The images were captured in a controlled environment, and therefore the images include variability in minor pose variation.

The Indian Institute of Technology Kanpur (IITK) Ear Database. The Indian Institute of Technology Kanpur (IITK) is an ear database [22] that the Institute of Technology of Kanpur compiled. The database is split into three sets, the first set consists of 190 male and female subjects of profile images. The total number of images was 801. The second dataset also contained 801, with a total of 89 subjects, these images had variations in pitch angle. The third dataset contains 1070 images of an equivalent of 89 subjects, but with a variation in yaw and angle.

The Forensic Ear Identification Database (FEARID). The Forensic Ear Identification Database (FEARID) database [3] is different from other databases as it only includes ear prints. These contain no occlusions, variable angles, or illumination. Though there is no mention of any variables, other influences like the force the ear was pressed against the scanner and the scanner's cleanliness need to be considered. This database comprised, 7364 images of 1229 subjects. This database was used for forensic applications and not for biometric use.

The University of Notre Dame (UND) Database. The University of Notre Dame (UND) database contains [30] many subsets of 2D and 3D ear images. These images were appropriated over a period from 2003 to 2005. The database contains, 3480 3D images from 952 male and female subjects and 464, 2D images from 114 male and female subjects. These images were taken in different lighting conditions, yaw, pitch poses and angles. The images are only of the left-hand side of the ear.

The Face Recognition Technology Database (FERET). The Face Recognition Technology Database (FERET) [21] is a sizeable facial image database, and was obtained between the years 1995 to 1996. It contains 1564 subjects and has a total of 14126 images. These images were collected for face recognition and were of the left- and right-hand profile images, which made them perfect for 2D ear recognition.

The Pose, Illumination and Expression (PIE). Carnegie Mellon University obtained The Pose, Illumination and Expression database [25], which contains,

Table 2. Summary of Datasets

	Database	Year	Number of subjects	Number of Images	Left Ear Count	Right Ear Count	Total Ears	Image Size	Country	Sides
1	Institute of Technology Delhi Ear Database (IIT Delhi-I) [19]	2007	121	471		471	471	272×204	India	Right
	Institute of Technology Delhi Ear Database (IIT Delhi-II) [19]	NA	221	793		793	793	272×204	India	Right
2	The University of Science & Technology Beijing (USTB Ear I) [31]	2002	60	185		185	185	Varied	China	Right
	The University of Science & Technology Beijing (USTB Ear II) [31]	2004	77	308		308	308	Varied	China	Right
3	The Annotated Web Ears database (AWE) [11]	2016	100	1000	500	500	1000	Varied	Slovenia	Both
	The Annotated Web Ears database extended (AWE extend) [11]	2017	346	4104	2052	2052	4104	Varied	Slovenia	Both
4	Mathematical Analysis of Images Ear Database (AMI) [14]	NA	106	700	420	280	700	492×702	Spain	Both
5	The West Pomeranian University of Technology Ear Database (WPUTE) [13]	2010	501	2071	829	1242	2071	Varied	Poland	Both
6	Unconstrained Ear Recognition Challenge database (UERC) [10]	2017	3706	11804	5902	5902	11804	Varied	Solvenia	Both
7	EarVN1.0 [16]	2018	164	28412	14206	14206	28412	Varied and low resolution	Vietnam	Both
8	The In-the wild Ear Database (ITWE) [12]	2015	55	605	424	181	605	Varied	Solvenia	Both
9	The Carreira-Perpinan (CP) [7]	1995	17	102	102		102	Varied	NA	Left
10	The University of Beira Ear Database (UBEAR) [23]	2011	126	4430	2215	2215	4430	1280×960	Mozambique	Both
11	Indian Institute of Technology Kanpur (IITK) [22]	2011	801	190	95	95	190	Varied	India	Both
12	The Forensic Ear Identification Database (FEARID) [3]	2005	1229	1229	615	614	1229	Varied	United Kingdom, Italy and Netherlands.	Both
13	University of Notre Dame (UND) [30]	2006	3480	952	952		952	Varied	France	Left

Table 3. Summary of Datasets

	Database	Year	Number of subjects	Number of Images	Left Ear Count	Right Ear Count	Total Ears	Image Size	Country	Sides
14	The Face Recognition Technology Database (FERET) [21]	2010	9427	4745	3796	949	4745	Varied	Spain	Both
15	The Pose, Illumination and Expression (PIE) [25]	2002	40000	68	34	34	68	Varied	USA	Both
16	The XM2VTS Ear Database [20]	NA	2360	295	89	206	295	720×576	UK	Both
17	The West Virginia University (WVU) [1]	2006	460	402	402		402	Varied	USA	Left

40000 images and 68 subjects. The images are of the facial profile and have different poses, illuminations and expressions.

The XM2VTS Ear Database. The XM2VTS ear database [20] is frontal and profiles facial images from the University of Surrey; the database contains 295 subjects and, 2360 images captured during controlled conditions. These images were a set of cropped images 720×576 pixel size and were from video data.

The West Virginia University (WVU) Ear Database. The West Virginia University (WVU) Ear Database [1] is a video database and is formed from 137 subjects. The system was an advanced capturing procedure that allowed them to capture the ear at different angles; these images included earrings and eyeglasses.

3.2 Pre-processing

Image pre-processing is a considerable part of the deep-learning task. Most CNN models generally require a large dataset to learn to discriminate features suitably for making predictions and obtaining a good performance. As images in the datasets are of different sizes, the inputted images need to be resized to conform to all the other CNN models, but the features need to be preserved when resizing is performed.

3.3 Transfer Learning

In this study, the concept of transfer learning was adopted and helped with the pre-trained CNN model for large datasets to learn features of the target (right and left ears). It will transfer the features learned by the deep CNN models on other CNN models to this dataset. The number of deep CNN model parameters increases as the network gets deeper, which is used to achieve improved efficiency.

Hence, it requires many datasets for training, making it computationally complex and applying these models directly on small and new dataset results in feature extraction bias, overfitting, and poor generalisation. The pre-trained

CNN modified and fine-tuned its structure to suit the dataset given. This concept of transfer learning is computationally expensive, has less training time, overcomes limitations of the dataset, improves performance, and is faster than training a model from the beginning. The pretraining CNN model fine-tuned in this work is the Transformer Network. The proposed structure is represented in Fig. 1.

3.4 Transformer Network Architecture

The Transformer Network is an encoder-decoder architecture based on attention layers. The difference between a Convolutional Neural Network and the Transformer Network is that the data can be passed in parallel, this means that the GPU can be utilised effectively and efficiently. The speed of the training is also increased by processing it in parallel. It is seen that the Transformation Network is based on a multi-headed attention layer and by doing this the vanishing gradient issue is overcome.

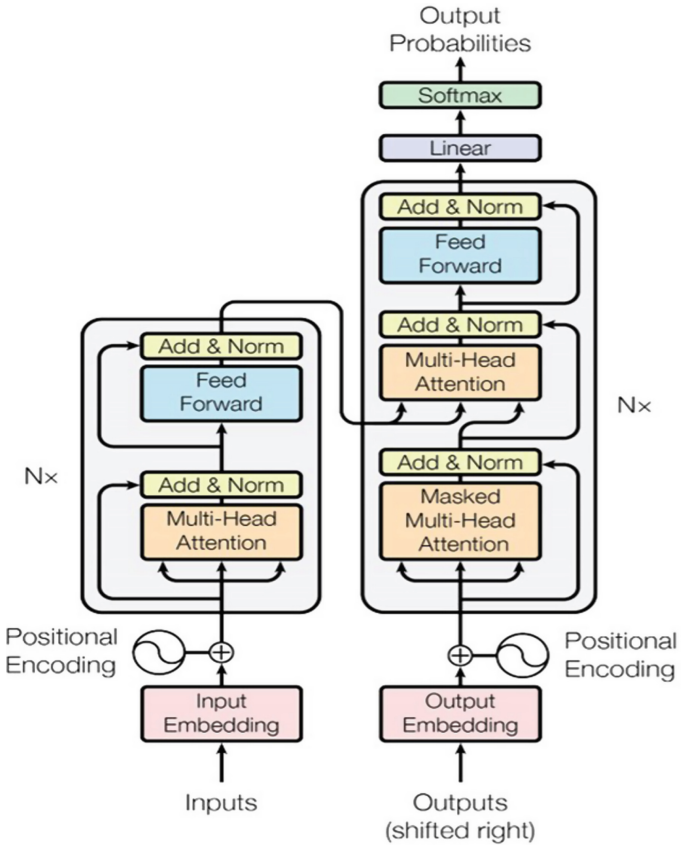
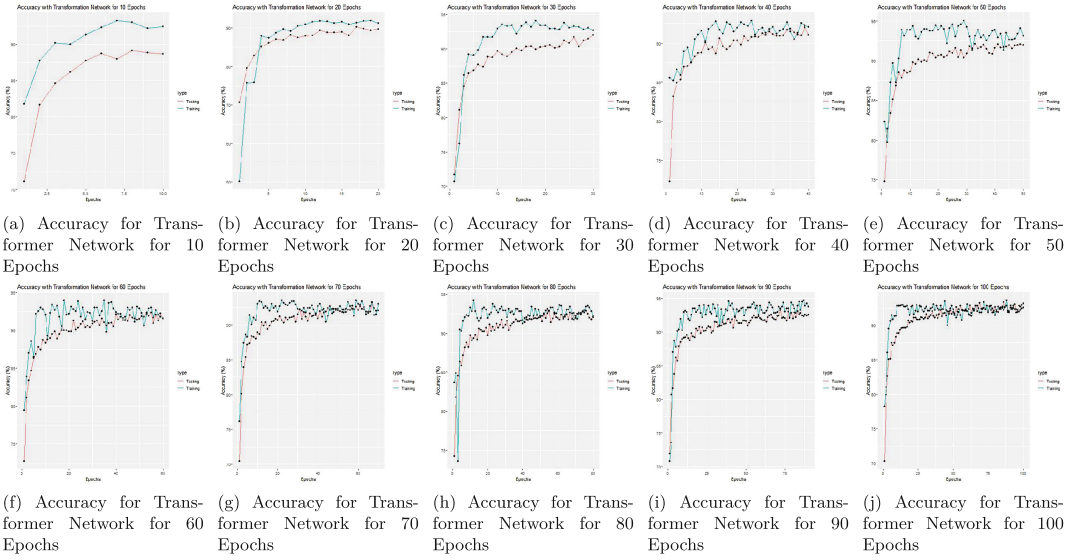


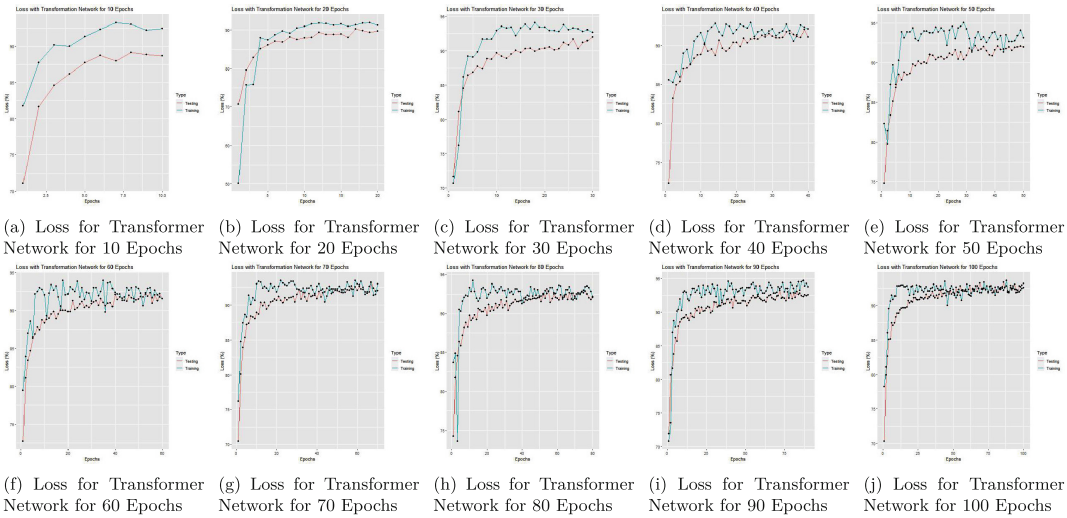
Fig. 1. Block structure of the proposed model



(a) Accuracy for Transformer Network for 10 Epochs (b) Accuracy for Transformer Network for 20 Epochs (c) Accuracy for Transformer Network for 30 Epochs (d) Accuracy for Transformer Network for 40 Epochs (e) Accuracy for Transformer Network for 50 Epochs

(f) Accuracy for Transformer Network for 60 Epochs (g) Accuracy for Transformer Network for 70 Epochs (h) Accuracy for Transformer Network for 80 Epochs (i) Accuracy for Transformer Network for 90 Epochs (j) Accuracy for Transformer Network for 100 Epochs

Fig. 2. Accuracy for the ear dataset of each Transformer Network



(a) Loss for Transformer Network for 10 Epochs (b) Loss for Transformer Network for 20 Epochs (c) Loss for Transformer Network for 30 Epochs (d) Loss for Transformer Network for 40 Epochs (e) Loss for Transformer Network for 50 Epochs

(f) Loss for Transformer Network for 60 Epochs (g) Loss for Transformer Network for 70 Epochs (h) Loss for Transformer Network for 80 Epochs (i) Loss for Transformer Network for 90 Epochs (j) Loss for Transformer Network for 100 Epochs

Fig. 3. Loss for the ear dataset of each Transformer Network

Table 4. Proposed method compared with the related studies

Author	Dataset	Accuracy	Summary
Emeršič et al. [9]	NA	30	It used handcrafted feature extraction methods, such as LBP, POEM and CNN-based feature extraction methods were used to obtain the ear identification.
Tian et al. [26]	AMI, WPUT, IITD, and UERC	70.58, 67.01, 81.98, & 57.75	This system used deep CNN to do ear recognition.There were occlusions like no earrings, headsets, or similar occlusions.
Raveane, Galdámez and Arrieta [24]	NA	98	This system used variable conditions due to the odd shape human ear and changing lighting conditions.
Zhang and Mu [32]	UND and UBEAR	100 & 98.22	This system contained large occlusions, scale and pose variation.
Kohlakala and Coetzer [18]	AMI and IIT-Delhi	99.2 & 96.06	It is used to classify ears either in the foreground or background of the image. The binary contour image applied the matching for feature extraction, and this was done by implementing a Euclidean distance measure, which had a ranking to verify for authentication.
Tomczyk and Szczepaniak [27]	NA	NA	It shows the published experimental results that the approach did the rotation equivalence property to detect rotated structures.
Hammam et al. [5]	Three ear datasets but not stated	22	The paper took seven, performing handcrafted descriptors to extract the discriminating ear image. Then took the extracted ear and trained it using SVM to learn a suitable model.
Alkababji and Mohammed [4]	NA	97.8	It used the PCA and a genetic algorithm for feature reduction and selection.
Jamil et al. [17]	Very underexposed or overexposed database	97	This work was the first to test the performance of CNN on very underexposed or overexposed images.
Hansley et al. [15]	UERC challenge	NA	This was done using handcrafted descriptors, which were fused to improve recognition.
Our Work	**AWE, AMI and IIT**	**92**	

4 Results and Discussion

Transformer Network variants were fine-tuned on all the ear datasets to detect the ear. Each dataset is split into 20% training and 80% test sets. The experiments were entirely performed using Keras deep learning framework using the TensorFlow backend. The models were evaluated using the popular evaluation metrics, Eq. 1 -5, (accuracy, recall and precision. The performances of all experiments are evaluated by using a series of confusion matrix-based performance metrics.

The confusion matrices used to evaluate the classifiers, with true positives (TP) representing the ears that are correctly classified as positive, true negatives (TN) representing the ears that are correctly classified as negative, false positives (FP) representing the ears that are incorrectly classified as positive, and false negatives (FN) representing the ears being incorrectly classified as negative.

Precision. It is the ratio of correctly classified negative instances by a model to the overall number of true negative instances being tested, Eq. 3.

Accuracy. It is a measure that indicates the ratio of all the correctly recognized cases to the overall number of cases. While this metric generally gives a decent reflection of the classifier, it may not reflect a classifier's true performance in a swhichrio where there is an uneven class distribution. Accuracy can be computed using the following formula, Eq. 1.

Recall. It is the ratio of all correctly classified positive instances by a model to the overall number of positive classifications by a model. A low precision indicates that a model suffers from high false positives. Precision can be computed using the following formula, Eq. 2.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TN}{TN + FP} \tag{3}$$

$$TPR = Recall = \frac{TP}{TP + FN} \tag{4}$$

$$FPR = 1 - Precision = \frac{FP}{FP + TN} \tag{5}$$

The results obtained are presented in Figs. 2 and 3 this is the accuracy and loss of these datasets. In the Transformer Network at different epochs the accuracy is determined using the test set. The models performed at extracting and

learning discriminative features from the dataset. Transformer Network with 50 and 90 Epochs attains the best accuracy 92.60 and 92.56%, and the Transformer Network results are noted in Table 5.

An advantage of Transformer Networks is that they are smaller with fewer parameters, faster, and obtain transfer learning successfully from the datasets. The worst performing was 20 epochs, as shown in Table 5. The reason that this performed poorly could have been because it did not have enough data to learn from. This was done to conform to the model's image input size. It can be seen that performance improves as the model gets deeper. On average it was seen that overfitting occurred at 30 iterations and stabilised at around 50. The best performing Global Transformer Network is at epochs 50 and 90, as shown in Table 5 and, this is because of the large number of parameters. It began to converge from the 30 iterations and then stabilised until 50 iterations when overfitting started. Determining the most suitable hyperparameters was one of the challenges faced as the overfitting, which was limited due to the data samples. The results of the proposed methods compared with related studies are presented in Fig. 4.

Table 5. Performance of Transformer Network

Epochs	Accuracy (%)	Loss (%)
10	90.42	47.78
20	87.06	37.47
30	91.10	31.30
40	90.97	30.94
50	92.60	27.96
60	91.74	28.03
70	91.81	26.80
80	92.18	26.67
90	92.56	25.42
100	91.91	26.91

5 Conclusion

This study investigated and implemented did pre-process by fine-tuning and pre-training the CNN before applying the Transformer Network to automatically identify ears on the most prominent and publicly available datasets. Transformer Networks that achieved state-of-the-art performance over other architectures to maximise accuracy and efficiency were explored and fine-tuned on profile images. The fine-tuning technique is valuable to utilise rich generic features learned from significant datasets sources such as ImageNet to complement the lack of annotated datasets affecting ear domains. The experimental results show the effectiveness of Transformer Network in extracting and learning distinctive

features from the ear images and then classifying them into a left or right-suitable class. Out of the ten Transformer Network variants explored in this study, the Transformer Network with 90 Epochs outperformed the others, as evident in Table 4. One of the limitations found was that it is easily over-fitted. To overcome this, you need to have compelling image preprocessing techniques. Although the proposed methodology is specified to do ear detection, it could be extended to detect other parts of the face, given the right set of datasets.

References

1. Abaza, A.: High Performance Image Processing Techniques in Automated Identification Systems. West Virginia University, Morgantown (2008)
2. Abaza, A., Ross, A., Hebert, C., Harrison, M.A.F., Nixon, M.S.: A survey on ear biometrics. ACM Comput. Surv. **45**(2) (2013). https://doi.org/10.1145/2431211.2431221
3. Alberink, I., Ruifrok, A.: Performance of the fearid earprint identification system. Forensic Sci. Int. **166**(2–3), 145–154 (2007)
4. Alkababji, A.M., Mohammed, O.H.: Real time ear recognition using deep learning. Telkomnika **19**(2), 523–530 (2021)
5. Alshazly, H., Linse, C., Barth, E., Martinetz, T.: Handcrafted versus CNN features for ear recognition. Symmetry **11**(12), 1493 (2019)
6. Chen, H., Bhanu, B.: Ear Biometrics 3D, pp. 241–248. Springer, US, Boston, MA (2009). https://doi.org/10.1145/2431211.2431221
7. Carreira-Perpinan, M.A.: Compression neural networks for feature extraction: application to human recognition from ear images (1995)
8. Chowdhury, D.P., Bakshi, S., Guo, G., Sa, P.K.: On applicability of tunable filter bank based feature for ear biometrics: a study from constrained to unconstrained. J. Med. Syst. **42**(1), 1–20 (2018)
9. Emeršič, Ž., Štepec, D., Štruc, V., Peer, P.: Training convolutional neural networks with limited training data for ear recognition in the wild. arXiv preprint arXiv:1711.09952 (2017)
10. Emeršič, Ž., et al.: The unconstrained ear recognition challenge. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), pp. 715–724. IEEE (2017)
11. Emeršič, Ž, Štruc, V., Peer, P.: Ear recognition: more than a survey. Neurocomputing **255**, 26–39 (2017)
12. Emeršič, V., Peer, P.: Ear biometric database in the wild, pp. 27–32 (2015). https://doi.org/10.1109/IWOBI.2015.7160139
13. Frejlichowski, D., Tyszkiewicz, N.: The west Pomeranian university of technology ear database - a tool for testing biometric algorithms
14. Gonzalez, E., Alvarez, L., Mazorra, L.: Ami ear database (2012). http://ctim.ulpgc.es/research_works/ami_ear_database/
15. Hansley, E.E., Segundo, M.P., Sarkar, S.: Employing fusion of learned and handcrafted features for unconstrained ear recognition. IET Biometrics **7**(3), 215–223 (2018)
16. Hoang, V.T.: Earvn1.0: a new large-scale ear images dataset in the wild. Data Brief **27**, 104630 (2019).https://doi.org/10.1016/j.dib.2019.104630, https://www.sciencedirect.com/science/article/pii/S2352340919309850

17. Jamil, N., Almisreb, A., Ariffin, S.M.Z.S.Z., Din, N.M., Hamzah, R.: Can convolution neural network (CNN) triumph in ear recognition of uniform illumination invariant? (2018)
18. Kohlakala, A., Coetzer, J.: Ear-based biometric authentication through the detection of prominent contours. SAIEE Africa Res. J. **112**(2), 89–98 (2021)
19. Kumar, A.: Iit delhi ear database version 1.0. (2007). http://webold.iitd.ac.in/~biometrics/Database_Ear.htm
20. Messer, K., et al.: XM2VTSDB: the extended M2VTS database. In: Second International Conference on Audio and Video-based Biometric Person Authentication, vol. 964, pp. 965–966. Citeseer (1999)
21. Phillips, P., Wechsler, H., Huang, J., Rauss, P.J.: The FERET database and evaluation procedure for face-recognition algorithms. Image Vis. Comput. **16**(5), 295–306 (1998)
22. Prakash, S., Jayaraman, U., Gupta, P.: Connected component based technique for automatic ear detection. In: 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 2741–2744. IEEE (2009)
23. Raposo, R., Hoyle, E., Peixinho, A., ProenĂsa, H.: Ubear: A dataset of ear images captured on-the-move in uncontrolled conditions (2011).https://doi.org/10.1109/CIBIM.2011.5949208
24. Raveane, W., Galdamez, P.L., Gonzalez Arrieta, M.A.: Ear detection and localization with convolutional neural networks in natural images and videos. Processes **7**(7), 457 (2019)
25. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression (pie) database of human faces. Technical report CMU-RI-TR-01-02, Carnegie Mellon University, Pittsburgh, PA (2001)
26. Tian, L., Mu, Z.: Ear recognition based on deep convolutional network. In: 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 437–441. IEEE (2016)
27. Tomczyk, A., Szczepaniak, P.S.: Ear detection using convolutional neural network on graphs with filter rotation. Sensors **19**(24), 5510 (2019)
28. Vu, N.S., Dee, H.M., Caplier, A.: Face recognition using the poem descriptor. Pattern Recogn. **45**(7), 2478–2488 (2012)
29. Wang, Z.Q., Yan, X.D.: Multi-scale feature extraction algorithm of ear image. In: 2011 International Conference on Electric Information and Control Engineering, pp. 528–531. IEEE (2011)
30. Yan, P., Bowyer, K.: Empirical evaluation of advanced ear biometrics. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)-Workshops, pp. 41–41. IEEE (2005)
31. Zhang, Y., Mu, Z.C., Yuan, L., Yu, C., Qing, L.: USTB-Helloear: a large database of Ear Images Photographed Under Uncontrolled Conditions, pp. 405–416 (2017). https://doi.org/10.1007/978-3-319-71589-6_35
32. Zhang, Y., Mu, Z.: Ear detection under uncontrolled conditions with multiple scale faster region-based convolutional neural networks. Symmetry **9**(4), 53 (2017)

3.4 Lightweight Deep Learning with Model Compression for Ear Recognition

3.4.1 Brief Overview

This section introduces a research paper whose main contribution is presenting the results of a CNN developed with the ReducedFireNet model. The model reduces the input size and increases computation time but detects more robust features such as ears. The ReducedFireNet Model consists of four Fire modules consisting of two layers: a 1 x 1 convolutional layer and a concatenation of 1 x 1 and 3 x 3 convolutional layers. A Max-pooling layer is applied to the first three Fire modules. The reason for the Max-pooling layer is to reduce the size of the input and increase computation time. It is also seen that it helps detect more robust features such as ears. The last layer of the matrix has a GlobalAveragePooling layer applied. A dropout layer is applied once the Max pooling layer has been applied on the second layer. The worst performance was with the maximum number of images and an image size of 1 x 1 due to insufficient learning data. However, the model performance improves as it gets deeper.

The paper is under review by the *Communications in Computer and Information Science - Intelligent Information and Database System*, Springer, Cham.

Lightweight Deep Learning with Model Compression for Ear Recognition

Aimee Booysens[1] and Serestina Viriri[1]

School of Mathematics, Statistics and Computer Science
University of KwaZulu-Natal
Durban, South Africa
210501411@stu.ukzn.ac.za and viriris@ukzn.ac.za

Abstract. The ReducedFireNet model reduces the size of the input and increases computation time, but detects more robust features such as ears. The ReducedFireNet Model consists of four Fire modules consisting of two layers: a 1 x 1 convolutional layer and a concatenation of 1 x 1 and 3 x 3 convolutional layers. A Max-pooling layer is applied to the first three Fire modules. The reason for the Max-pooling layer is to reduce the size of the input and increase computation time. It is also seen that it helps detect more robust features such as ears. The last layer of the matrix has a GlobalAveragePooling layer applied. A dropout layer is applied once the Max pooling layer has been applied on the second layer.The worst performance was with the maximum number of images and an image size of 1 x 1, due to insufficient learning data. However, the model performance improves as it gets deeper. On average, overfitting occurred at the max number of images and image size 16 x 16 iterations and stabilised at around the max number of images and image size 128 x 128 because of the large number of parameters. These experiments showed that the ReducedFireNet model achieved a high accuracy of 87.91%.

Keywords: Ear Recognition · Ear Biometrics · Deep Learning · Lightweight

1 Introduction

The ear develops in a foetus during the fifth and seventh weeks of pregnancy [1]. At this stage, the face acquires a more distinguishable shape as the mouth, nostrils, and ears begin to form. There is still no exact timeline at which the outer ear is created, but it is accepted that a cluster of embryonic cells connects to establish the ear. These are called auricular hillocks, which begin growing in the lower portion of the neck. The auricular hillocks broaden and intertwine within the seventh week to deliver the ear's shape. Within the ninth week, the hillocks move to the ear canal and are more noticeable as the ear [1]. The growth of the ear in the first four months after birth is linear, and the ear is then stretched in development between the ages of four months and eight years. After this, the

ear size and shape are constant until age seventy, increasing in size again.

Biometrics is the recognition of a human using their biometric characteristics, which may be physiological or behavioural. The physiological biometric features are the DNA, face, ear, facial, iris, fingerprint, hand geometry, hand vein and palm print, and behavioural biometrics are signatures, gait patterns and keystrokes. Voice is considered a combination of biometric and physiological characteristics. Numerous systems have been developed to distinguish biometric traits, which have been used in multiple applications, such as forensic investigations and security systems. With the current Worldwide pandemic, facial identification has failed due to users wearing masks. However, the human ear has proven more suitable as it is visible.

In different physiological biometric qualities, the ear has received much consideration of late as it tends to be said that it is a solid biometric for human acknowledgement [4]. Ear biometric framework is dependable as it does not change, is of uniform tone, and is fixed at the centre of the face's side. An individual's ear size is as unique as a finger impression. It simplifies capturing an image of the subject without necessarily needing to gain information from it, [4]. There are numerous difficulties in correctly gauging the details of the ear, and these are the concealment of the ear by clothing, hair, ear ornaments and jewellery. Another interference could be the angle of the image, concealing essential characteristics of the ear's anatomy. These difficulties have made ear recognition a secondary role in identification systems and techniques commonly used for identification and verification.

Although several computer-aided detection models have been developed to identify ears, low accuracy and sensitivity are still significant concerns that cause ears to be misidentified. Existing models are also computationally complex and expensive. The contributions of this work are summarised as follows:

1. Implement state-of-the-art Lightweight Deep Learning with Model Compression to develop an effective and inexpensive ear detection system. It is the first time this model has been applied to classify ears.

2. The proposed model accuracy through Lightweight Deep Learning with Model Compression.

3. Finally, benchmark datasets were used to evaluate the performance of the model.

The remaining work is structured as follows: Section 2 presents related works, and Section 3 presents detailed data and methodology explored in this study. The experimental results and discussion are provided in Section 4, and Section 5 concludes the paper.

2 Literature Review and Related Work

This section presents different algorithms using the Convolutional Neural Network (CNN) for ear identifications. The competition Emeršič et al. [6] organized the dataset of the UERC, which was used for the benchmark, training and testing sets. In the completion, it was seen that handcrafted feature extraction methods, such as LBP [14] and patterns of oriented edge magnitudes (POEM) [13], and CNN-based feature extraction methods were used to obtain the ear identification. The challenges were to find methods to remove occlusions such as earrings, hair, other obstacles, and background from the ear image. The occlusion was done by creating a binary ear mask, and then the system recognition was done using the handcrafted features. Another proposed approach was calculating the score of matrices from the CNN-based features and handcrafted features when fused; a 30% detection rate was achieved.

Tian et al. [11] applied a CNN to ear recognition in which they designed a CNN – it was made up of three convolutional layers, a fully connected layer, and a softmax classifier. The database used was USTB ear, which consisted of 79 subjects with various pose angles. The images utilized excluded earrings, headsets, or similar occlusions. Chowdhury et al. [5] proposed an ear biometric recognition system that uses local features and a neural network to identify the ear. The method estimates where the ear could be in the input image and then takes the edge features from the identified ear. After identifying the ear, a neural network matches the extracted feature with a feature database. The databases used in this system were AMI, WPUT, IITD, and UERC, which achieved an accuracy of 70.58%, 67.01%, 81.98%, and 57.75%.

Raveane, Galdámez and Arrieta [10] presented that detecting and locating an ear within an image is difficult. They found that the challenge increases when working with the variable condition, and this could also be because of the odd shape of the human ears as well as lighting conditions and the changing profile shape of an ear when photographed, [10]. The ear detection system used multiple CNNs and a detection grouping algorithm to identify an ear's presence and location. The proposed method matches other methods' performance when analyzed against clean and purpose-shot photographs, reaching an accuracy of upwards of 98%. It outperforms them with a rate of over 86% when the system is subjected to non-cooperative natural images where the subject appears in challenging orientations and photographic conditions.

Multiple Scale Faster Region-based Convolutional Neural Networks (Faster R-CNN) to detect ears from 2D profile images was proposed by Zhang and Mu [15]. They used methods by taking three regions of different scales that were detected to defer the information from the ear location within the context of the ear in the image, which was done to extract the ear correctly. The system was tested with 200 web images that achieved a 98% accuracy. Other experiments conducted were on the Collection J2 of the University of Notre Dame

Biometrics Database (UND-J2) and the University of Beira Interior Ear dataset (UBEAR); these achieved a detection rate of 100% and 98.22%, respectively, but these datasets contained large occlusions, scale and pose variation.

Kohlakala and Coetzer [9] presented semi-automated and fully automated ear-based biometric verification systems. CNN and morphological postprocessing manually identify the ear region. It is used to classify ears in the image's foreground or background. The binary contour image applied the matching for feature extraction, and this was done by implementing a Euclidean distance measure, which had a ranking to verify for authentication. The Mathematical Analysis of Images ear database and the Indian Institute of Technology Delhi ear database were two databases which achieved 99.20% and 96.06%, respectively.

Geometric deep learning (GDL) generalizes CNNs to non-Euclidean domains, presented by [12] Tomczyk and Szczepaniak. It used convolutional filters with a mixture of Gaussian models. These filters were used to rotate the images without interpolation easily. The published experimental results show that the approach did the rotational equivalence property to detect rotated structures. Still, it does not need labour-intensive training on all rotated and non-rotated images.

Hammam et al. [3] presented and compared ear recognition models built with handcrafted and CNN features. The paper took seven, performing handcrafted descriptors to extract the discriminating ear image. They then trained the extracted ear using Support Vector Machines (SVM) to learn a suitable model. They then used CNN-based models, which used a variant of the AlexNet architecture. The results obtained on three ear datasets showed the CNN-based models' performance increased by 22%. This paper also investigated if the left and right ears have symmetry. The results obtained by the two datasets indicate a high impact of balance between the ears.

Alkababji and Mohammed [2] presented using a Deep Learning item detector called faster region-based convolutional neural networks (Faster R-CNN) for ear detection. This CNN is used for feature extraction. It used the Principal Component Analysis (PCA) and a genetic algorithm for feature reduction and selection. It also used a connected artificial neural network as the matcher. The results achieved an accuracy of 97.8% success rate.

Jamil et al. [8] build and train a CNN model for ear biometrics in various uniform illuminations measured using lumens. They considered that their work was the first to test the performance of CNN on underexposed or overexposed images. The results showed that images with uniform illumination with a luminance of above 25 lux achieved a result of 100%. The CNN model had problems recognizing images when the lux was below ten but produced an accuracy of 97%. This result shows that CNN architecture performs just as well as the other systems.

It was found that the dataset had rotations which affected the results.

Hansley et al. [7] presented an unconstrained ear recognition framework better than the current state-of-the-art systems using publicly available databases. They developed CNN-based solutions for ear normalization and description. It was done using handcrafted descriptors fused to improve recognition in two stages. The first stage was to find the landmark detectors and untrained scenarios. The next step was to generate a geometric image normalization to boost the performance. It was seen that the CNN descriptor was better than other CNN-based works in the literature. The obtained results were higher than different reported results for the UERC challenge.

3 Methods and Techniques

3.1 Dataset

This study performed experiments with numerous public ears; UBEAR, EarVN1.0, IIT, ITWE and AWE databases are best suited for ear identification due to their large data size. However, it shows that EarVN1.0 has the foremost prominent usage during age estimation using CNN techniques. It is an appropriate dataset for ear images taken in a controlled environment, while ITWE is compatible with classifying ears in an uncontrolled environment.

3.2 Pre-processing

Image pre-processing is a considerable part of the deep-learning task. Most CNN models require a large dataset to learn to discriminate features suitably for making predictions and obtaining a good performance. As images in the datasets are of different sizes, the inputted images must be resized to conform to all the other CNN models. However, the features need to be preserved when resizing is performed.

3.3 ReducedFireNet Model Architecture

The ReducedFireNet Model is a convolutional neural network (CNN) type that uses multiple fire modules. The ReducedFireNet Model consists of four Fire modules consisting of two layers: a 1 x 1 convolutional layer and a concatenation of 1 x 1 and 3 x 3 convolutional layers. A Max-pooling layer is applied to the first three Fire modules. The reason for the Max-pooling layer is to reduce the size of the input and increase computation time. It is also seen that it helps detect more robust features such as ears. The last layer of the matrix has a GlobalAveragePooling layer applied. A dropout layer is applied once the Max pooling layer has been applied on the second layer. The reason that this is done is to reduce overfitting and to make it more robust. Rectified Linear Unit (ReLU) is the activation function applied to each convolutional layer's output. Once the

dense layer is created, a softmax activation function, shown in equation 1, is applied to classify the ear. The depiction of how the ReducedFireNet Model works is shown in figure, 1.

$$\sigma\left(q\right)_i = \frac{e^{q_i}}{\sum_{N}^{y=1} e^{q_y}} \tag{1}$$

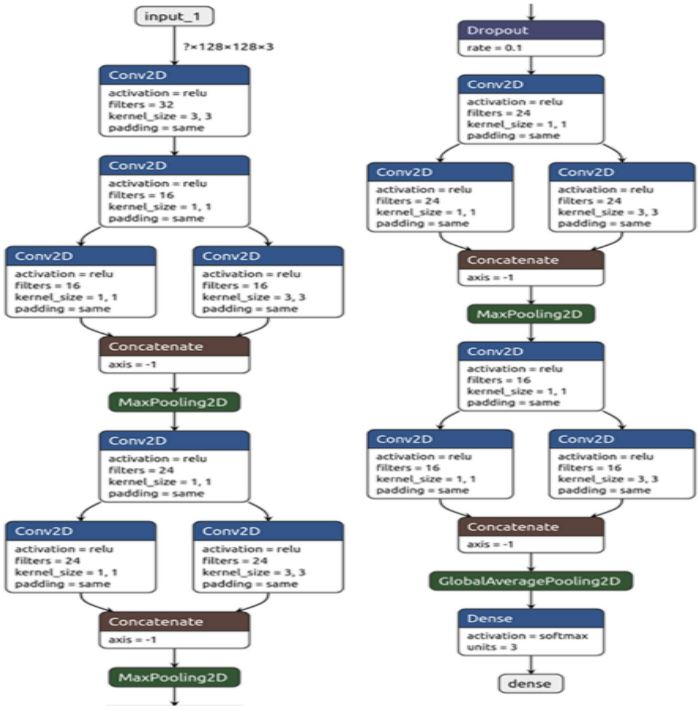


Fig. 1: ReducedFireNet Model Configuration

4 Results and Discussion

ReducedFireNet Model variants were fine-tuned on all the ear datasets to detect the ear. Each dataset is split into 20% training and 80% test sets. The experiments used the Keras deep learning framework using the TensorFlow backend. The models were evaluated using the popular evaluation metrics, equation 2 - 6 (accuracy, recall and precision. All experiments' performances are evaluated using confusion matrix-based performance metrics.

Table 1: Performance of ReducedFireNet Model

Size of image	Accuracy (%)	Loss (%)
1 x 1	44.15	55.85
3 x 3	55.44	44.56
8 x 8	53.00	47.00
16 x 16	44.98	55.02
32 x 32	52.56	47.44
64 x 64	73.87	26.13
128 x 128	87.91	12.09

Precision is the ratio of correctly classified negative instances by a model to the overall number of true negative instances being tested, equation 4.

Accuracy is a measure that indicates the ratio of all the correctly recognized cases to the overall number of cases. While this metric generally gives a decent reflection of the classifier, it may not reflect a classifier's true performance in an uneven class distribution scenario. Accuracy can be computed using the following formula, equation 2.

Recall is the ratio of all correctly classified positive instances by a model to the overall number of positive classifications by a model. A low precision indicates that a model suffers from high false positives. Precision can be computed using the following formula, equation 3.

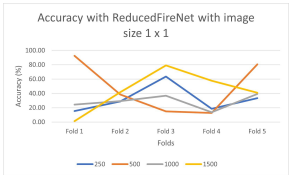
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

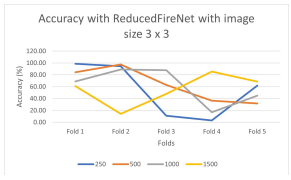
$$Precision = \frac{TN}{TN + FP} \tag{4}$$

$$TPR = Recall = \frac{TP}{TP + FN} \tag{5}$$

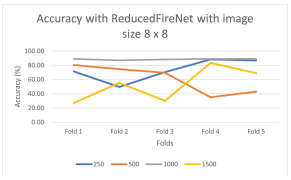
$$FPR = 1 - Precision = \frac{FP}{FP + TN} \tag{6}$$



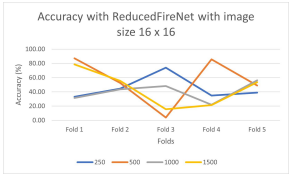
(a) Accuracy for ReducedFireNet Model with image size 1 x 1



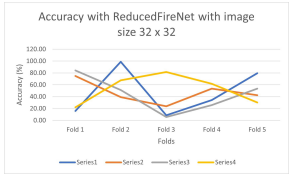
(b) Accuracy for ReducedFireNet Model with image size 3 x 3



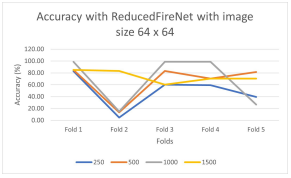
(c) Accuracy for ReducedFireNet Model with image size 8 x 8



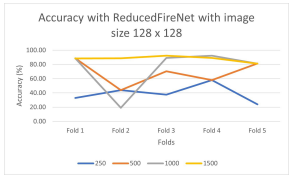
(d) Accuracy for ReducedFireNet Model with image size 16 x 16



(e) Accuracy for ReducedFireNet Model with image size 32 x 32



(f) Accuracy for ReducedFireNet Model with image size 64 x 64



(g) Accuracy for ReducedFireNet Model with image size 128 x 128

Fig. 2: Accuracy for the ear dataset of each ReducedFireNet Model

Table 2: Proposed method compared with the related studies

Author	Dataset	Accuracy	Summary
Emeršič et al. [6]	NA	30	It used handcrafted feature extraction methods, such as LBP, POEM and CNN-based feature extraction methods were used to obtain the ear identification.
Tian et al.[11]	AMI, WPUT, IITD, and UERC	70.58, 67.01, 81.98, & 57.75	This system used deep CNN to do ear recognition.There were occlusions like no earrings, headsets, or similar occlusions.
Raveane, Galdámez and Arrieta [10]	NA	98	This system used variable conditions due to the odd shape human ear and changing lighting conditions.
Zhang and Mu [15]	UND and UBEAR	100 & 98.22	This system contained large occlusions, scale and pose variation.
Kohlakala and Coetzer [9]	AMI and IIT-Delhi	99.2 & 96.06	It is used to classify ears either in the foreground or background of the image. The binary contour image applied the matching for feature extraction, and this was done by implementing a Euclidean distance measure, which had a ranking to verify for authentication.
Tomczyk and Szczepaniak[12]	NA	NA	It shows the published experimental results that the approach did the rotation equivalence property to detect rotated structures.
Hammam et al. [3]	Three ear datasets but not stated	22	The paper took seven, performing handcrafted descriptors to extract the discriminating ear image. Then, the extracted ear was trained using SVM to learn a suitable model.
Alkababji and Mohammed [2]	NA	97.8	It used the PCA and a genetic algorithm for feature reduction and selection.
Jamil et al. [8]	Very underexposed or overexposed database	97	This work was the first to test the performance of CNN on very underexposed or overexposed images.
Hansley et al. [7]	UERC challenge	NA	This was done using handcrafted descriptors, which were fused to improve recognition.
Proposed Work	**AWE, AMI and IIT**	**87.91**	

The results are presented in Figures 2; this is the accuracy and loss of these datasets. The accuracy of the ReducedFireNet Model at a different number of images is used in the testing set. The models performed at extracting and learning discriminative features from the dataset. ReducedFireNet Model with the max number of images and image size 128 x 128 was the best accuracy at 87.91%.

An advantage of the ReducedFireNet Model is that it reduces the input size and increases computation time. The worst performance was with the max number of images and an image size 1 x 1, as shown in figure 2. The reason that this performed poorly could have been because it did not have enough data to learn from. It was done to conform to the model's image input size. It can be seen that performance improves as the model gets deeper. On average, it was seen that overfitting occurred at the max number of images and image size 16 x 16 iterations and stabilised at around the max number of images and image size 128 x 128, as shown in Table 1 because of the large number of parameters.

5 Conclusion

This study investigated and implemented pre-processing by fine-tuning and pre-training the CNN before applying the Lightweight Deep Learning with Model Compression to automatically identify ears on the most prominent and publicly available datasets. Lightweight Deep Learning with Model Compression that achieved state-of-the-art performance over other architectures to maximize accuracy and efficiency was explored and fine-tuned on profile images. The fine-tuning technique is valuable to utilise rich generic features learned from significant datasets sources such as ImageNet to complement the lack of annotated datasets affecting ear domains. The experimental results show the effectiveness of Lightweight Deep Learning with Model Compression in extracting and learning distinctive features from the ear images and then classifying them into a left or right-suitable class. Out of the seven Lightweight Deep Learning with Model Compression variants explored in this study, the Lightweight Deep Learning with Model Compression with image size 128 x 128 outperformed the others, as evident in Table 2. One of the limitations found was that it is easily over-fitted. To overcome this, you need to have compelling image pre-processing techniques. Although the proposed methodology is specified for ear detection, it could be extended to detect other parts of the face, given the right set of datasets.

References

1. Abaza, A., Ross, A., Hebert, C., Harrison, M.A.F., Nixon, M.S.: A survey on ear biometrics. ACM Comput. Surv. **45**(2) (Mar 2013). https://doi.org/10.1145/2431211.2431221, `https://doi.org/10.1145/2431211.2431221`

2. Alkababji, A.M., Mohammed, O.H.: Real time ear recognition using deep learning. Telkomnika **19**(2), 523–530 (2021)
3. Alshazly, H., Linse, C., Barth, E., Martinetz, T.: Handcrafted versus cnn features for ear recognition. Symmetry **11**(12), 1493 (2019)
4. Bhanu, Bir, C., Hui: Ear Biometrics, 3D, pp. 241–248. Springer US, Boston, MA (2009). https://doi.org/10.1145/2431211.2431221, `https://doi.org/10.1145/2431211.2431221`
5. Chowdhury, D.P., Bakshi, S., Guo, G., Sa, P.K.: On applicability of tunable filter bank based feature for ear biometrics: a study from constrained to unconstrained. Journal of medical systems **42**(1), 1–20 (2018)
6. Emeršič, Ž., Štepec, D., Štruc, V., Peer, P.: Training convolutional neural networks with limited training data for ear recognition in the wild. arXiv preprint arXiv:1711.09952 (2017)
7. Hansley, E.E., Segundo, M.P., Sarkar, S.: Employing fusion of learned and handcrafted features for unconstrained ear recognition. IET Biometrics **7**(3), 215–223 (2018)
8. Jamil, N., Almisreb, A., Ariffin, S.M.Z.S.Z., Md Din, N., Hamzah, R.: Can convolution neural network (cnn) triumph in ear recognition of uniform illumination invariant? (2018)
9. Kohlakala, A., Coetzer, J.: Ear-based biometric authentication through the detection of prominent contours. SAIEE Africa Research Journal **112**(2), 89–98 (2021)
10. Raveane, W., Galdamez, P.L., Gonzalez Arrieta, M.A.: Ear detection and localization with convolutional neural networks in natural images and videos. Processes **7**(7), 457 (2019)
11. Tian, L., Mu, Z.: Ear recognition based on deep convolutional network. In: 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). pp. 437–441. IEEE (2016)
12. Tomczyk, A., Szczepaniak, P.S.: Ear detection using convolutional neural network on graphs with filter rotation. Sensors **19**(24), 5510 (2019)
13. Vu, N.S., Dee, H.M., Caplier, A.: Face recognition using the poem descriptor. Pattern Recognition **45**(7), 2478–2488 (2012)
14. Wang, Z.q., Yan, X.d.: Multi-scale feature extraction algorithm of ear image. In: 2011 International Conference on Electric Information and Control Engineering. pp. 528–531. IEEE (2011)
15. Zhang, Y., Mu, Z.: Ear detection under uncontrolled conditions with multiple scale faster region-based convolutional neural networks. Symmetry **9**(4), 53 (2017)

Chapter 4

Results and Discussions

4.1 Introduction

This chapter analyses the performance of identifying the ear images with different machine and deep learning techniques. The results achieved in this thesis project are also shown and discussed.

Section 4.2 explains the programming environment used in the thesis; section 4.3 discusses the datasets used. Sections 4.4 and 4.5 discuss the results of the machine learning classification techniques and deep learning for ear identification.

4.2 Programming Environment

The system implemented in the thesis was created on a computer with an Intel Core(TM) i7-4770S @ 3.10 GHz and 8.00GB RAM. The system was implemented in Rstudio using different plugins: Imager, Keras, Tensorflow and Class. All these plugins are in the public domain and are used for Rstudio image processing. Imager procedures process the images to greyscale and allow the images to be resized. Class processes all the required machine learning procedures, Decision Tree, Naïve Bayes and K-Nearest Neighbor, feature extraction techniques, Gabor Filters, Zernike Moments and many more. The Tensorflow and Keras procedures are used to process deep leaarning techniques.

4.3 Overview of Ear Datasets

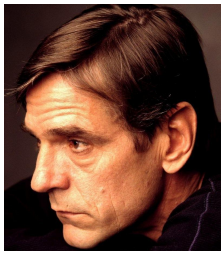
Many factors can affect an ear detection system's performance. Ear image datasets are easier to use than others. The more ear datasets available for researchers to use, the more this field can evolve and grow. Using high-quality images in research associated with soft biometrics produces the best results. These datasets have been described in sections 2.1, 3.3 and 3.2. A brief description of the available ear datasets is highlighted in Table 4.1, and examples of images are shown in Figure 4.1 and 4.2.

Due to their extensive data size, UBEAR, EarVN1.0, IIT, ITWE and AWE databases are best suited for ear identification. However, it shows that EarVN1.0 has the foremost prominent usage during age estimation using CNN techniques. It is an appropriate dataset while ear images are taken in a controlled environment, while ITWE is compatible with classifying ears in an uncontrolled environment. Examples of the extracted ears are shown in 4.1 and 4.2

(a) Example a 2D profile image for a female

(b) Example a 2D profile image for a male

(c) Example a facial image for a female



(d) Example a facial image for a male

Figure 4.1: Examples of original ear images

(a) Example ear extracted from 2D profile image for a female

(b) Example ear extracted from 2D profile image for a male

(c) Example ear extracted from facial image for a female

(d) Example ear extracted from facial image for a male

Figure 4.2: Examples of extracted ear images

Table 4.1: Summary of Databases

	Database	Year	Number of subjects	Number of Images	Left Ear Count	Right Ear Count	Total Ears	Image Size	Country	Sides
1	Institute of Technology Delhi Ear Database (IIT Delhi-I) [67]	2007	121	471		471	471	272 x 204	India	Right
	Institute of Technology Delhi Ear Database (IIT Delhi-II) [67]	NA	221	793		793	793	272 x 204	India	Right
2	The University of Science & Technology Beijing (USTB Ear I) [121]	2002	60	185		185	185	Varied	China	Right
	The University of Science & Technology Beijing (USTB Ear II) [121]	2004	77	308		308	308	Varied	China	Right
3	The Annotated Web Ears database (AWE) [45]	2016	100	1000	500	500	1000	Varied	Slovenia	Both
	The Annotated Web Ears database extended (AWE extend) [45]	2017	346	4104	2052	2052	4104	Varied	Slovenia	Both
4	Mathematical Analysis of Images Ear Database (AMI) [55]	NA	106	700	420	280	700	492 x 702	Spain	Both
5	The West Pomeranian University of Technology Ear Database (WPUTE) [51]	2010	501	2071	829	1242	2071	Varied	Poland	Both
6	Unconstrained Ear Recognition Challenge database (UERC) [43]	2017	3706	11804	5902	5902	11804	Varied	Solvenia	Both
7	EarVN1.0 [61]	2018	164	28412	14206	14206	28412	Varied and low resolution	Vietnam	Both
8	The In-the Wild Ear Database (ITWE) [48]	2015	55	605	424	181	605	Varied	Solvenia	Both
9	The Carreira-Perpinan (CP) [24]	1995	17	102	102		102	Varied	NA	Left

	Database	Year	Number of subjects	Number of Images	Left Ear Count	Right Ear Count	Total Ears	Image Size	Country	Sides
10	The University of Beira Ear Database (UBEAR) [95]	2011	126	4430	2215	2215	4430	1280x960	Mozambique	Both
11	Indian Institute of Technology Kanpur (IITK) [91]	2011	801	190	95	95	190	Varied	India	Both
12	The Forensic Ear Identification Database (FEARID) [7]	2005	1229	1229	615	614	1229	Varied	United Kingdom, Italy and Netherlands.	Both
13	University of Notre Dame (UND) [119]	2006	3480	952	952		952	Varied	France	Left
14	The Face Recognition Technology Database (FERET) [88]	2010	9427	4745	3796	949	4745	Varied	Spain	Both
15	The Pose, Illumination and Expression (PIE) [104]	2002	40000	68	34	34	68	Varied	USA	Both
16	The XM2VTS Ear Database [77]	NA	2360	295	89	206	295	720 x 576	UK	Both
17	The West Virginia University (WVU) [1]	2006	460	402	402		402	Varied	USA	Left

4.4 Results from Machine Learning Techniques

4.4.1 Composition of the Extraction Features

Analysis was done to ascertain which feature extraction technique would achieve the most accurate True Positive Rate (TPR) for ear identification. The feature extraction techniques used in this thesis are Local Binary Pattern, Zernike Moments, Gabor Filter, and Haralick Texture Moments. All these feature extraction techniques were classified using the K-Nearest Neighbor classifier.

Local Binary Pattern (LBP)

This section presents results for the feature extraction technique described in the paper in section 3.1.

The Linear Binary Pattern is a Textural Feature Extraction Technique embedded in the OpenImageR toolbox [56] and was used to implement the Linear Binary Pattern. This thesis applied the Linear Binary Pattern Feature Extraction Technique to the same ear images used to train the other feature extraction techniques. The same dataset was used for training, validation and testing for the Linear Binary Pattern. The results of the experiment are shown in Table 4.2:

Table 4.2: Results achieved by Linear Binary Pattern

Ear Side	True Identification	False Identification	True Percentage	False Percentage
Right Ear	2212	785	73.80%	26.2%
Left Ear	2152	845	71.80%	28.2%

The Linear Binary Pattern was tested using the entire dataset, which consisted of 2997 images. Each image contained the left and right ears. From the results in Table 4.2. Linear Binary Pattern has an Average True Positive classification rate of 72.8%.

Zernike Moments

This section presents results for the feature extraction technique described in the paper in section 3.1.

The Zernike Moments is a geometric Feature Extraction Technique embedded in the OpenImageR toolbox [56] and was used to implement the Zernike Moments. This thesis applied the Zernike Moments Feature Extraction Technique to the same ear images used to train the other feature extraction techniques. The same dataset was used for training, validation and testing for the Zernike Moments. The results of the experiment are shown in Table 4.3:

Table 4.3: Results achieved by Zernike Moments

Ear Side	**True Identification**	**False Identification**	**True Percentage**	**False Percentage**
Right Ear	2533	464	84.49%	15.51%
Left Ear	2479	518	82.71%	17.29%

The Zernike Moments was tested using the dataset, which consisted of 2997 images. Each image contained the left and right ears. From the results in Table 4.3. Zernike Moments has an Average True Positive classification rate of 83.6%.

Haralick Texture Moments

This section presents results for the feature extraction technique described in the paper in section 3.1.

The Haralick Texture Moments is a Textural Feature Extraction Technique embedded in the OpenImageR toolbox [56] and was used to implement the Haralick Texture Moments. This thesis applied the Haralick Texture Moments Feature Extraction Technique to the same ear images used to train the other feature extraction techniques. The same dataset was used for training, validation and testing for the Haralick Texture Moment. The results of the experiment are shown in Table 4.4:

Table 4.4: Results achieved by Haralick Texture Moments

Ear Side	True Identification	False Identification	True Percentage	False Percentage
Right Ear	2599	398	86.69%	13.31%
Left Ear	2989	8	99.70%	0.30%

The Haralick Texture Moments were tested using the whole dataset, which consisted of 2997 images. Each image contained the left and right ears. From the results in Table 4.4. Haralick Texture Moments has an Average True Positive classification rate of 93.20%.

Gabor Filter

This section presents results for the feature extraction technique described in the paper in section 3.1.

The Gabor Filter is a Textural Feature Extraction Technique embedded in the OpenImageR toolbox [56] and was used to implement the Gabor Filter. This thesis applied the Gabor Filter Feature Extraction Technique to the same ear images used to train the other feature extraction techniques. The same dataset was used for training, validation and testing for the Gabor Filter. The results of the experiment are shown in Table 4.5:

Table 4.5: Results achieved by Gabor Filter

Ear Side	True Identification	False Identification	True Percentage	False Percentage
Right Ear	2120	877	70.72%	29.28%
Left Ear	1988	1009	66.32%	33.68%

The Gabor Filter was tested using the entire dataset, which consisted of 2997 images. Each image contained the left and right ears. From the results in Table 4.5. Gabor Filter has an Average True Positive classification rate of 68.52%.

General Discussion of the Extraction Features

The results of the True Positive Rate that each combination of feature extraction techniques achieved showed that specific feature extraction techniques achieved a higher result in distinguishing the ears of profile and facial images. The results for each ear with combinations of Feature Extraction Techniques have been summarised in Table 4.6.

The analysis completed was to determine which feature extraction technique would achieve the most accurate True Positive Rate (TPR) for ear identification. The results were obtained by combining the feature vectors for each component and then classifying using K-Nearest Neighbor to observe which combination of feature extraction techniques achieved the highest True Positive Rate; results obtained are in Table 4.6.

Zernike Moments is a valuable geometric feature extraction technique as it correctly obtains the edges of the ear. This technique is used in the normalised feature vector to capture the ear's shape and proportion. Zernike Moments achieved a TPR of 82.71% for the left ear and 84.49% for the right ear.

Local Binary Pattern is a geometric feature extraction technique. This feature extraction technique is used in the normalised feature vector as it can detect the shape of the ear. This feature extraction technique correctly identified the inner lines of the ear. Local Binary Pattern obtained a TPR of 71.8% for the left ear and 73.8% for the right ear.

Haralick Texture is a texture extraction technique. It works well as a feature extraction texture as it picks up the course and the colour gradient of the ear. Haralick Texture achieved an average TPR of 99.7% for the left ear and 86.69% for the right ear.

The Gabor filter was the worst-achieving geometric feature of all the feature vector techniques. It only achieved a TPR of 66.32% for the left ear and 70.72% for the right ear. This feature vector technique works well with the other feature vector techniques, achieving a TPR of 92.87%.

Table 4.6 shows the average TPR of all the combinations of these feature extraction techniques. Better results are obtained by using more feature extraction techniques. Hence, all four feature extraction techniques must be used to obtain the feature vector. This vector is then fused and normalised to get ear identification.

Table 4.6: True Positive Rates per Ear per Combination of Feature Extraction Technique

Feature Techniques	**Percentage (%)**
Gabor Filter, Zernike Moments, Haralick Texture & Local Binary Pattern	92.87
Gabor Filter, Zernike Moments & Local Binary Pattern	90.80
Gabor Filter, Zernike Moments & Haralick Texture	91.75
Gabor Filter, Haralick Texture & Local Binary Pattern	90.45
Gabor Filter, Zernike Moments & Local Binary Pattern	87.56
Zernike Moments, Haralick Texture & Local Binary Pattern	88.46
Zernike Moments & Haralick Texture	86.78
Zernike Moments & Local Binary Pattern	88.96
Haralick Texture & Local Binary Pattern	87.86
Gabor Filter & Zernike Moments	85.42
Gabor Filter and Haralick Texture	83.26

4.4.2 Classification and Identification

Once the composite fused and normalised feature vector is obtained, it is used to determine how it receives the ear from an image. It was done using different machine learning algorithms to get the ear from the profile and facial image. These machine learning algorithms used were K-Nearest Neighbor, Decision Tree and Naïve Bayes.

This section presents results for the Classification and Identification technique described in the paper in section 3.1.

Decision Tree

The Decision Tree is a supervised machine learning algorithm embedded in the OpenImageR toolbox [56] and was used to implement the Decision Tree. The thesis took the same dataset as the other machine learning algorithm to test and train this machine learning algorithm. The results that are obtained from this experiment are shown in Table 4.7:

Table 4.7: Accuracy Achieved for Decision Tree

Ear Side	True Identification	False Identification	True Percentage	False Percentage
Left Ear	1 600	1 397	53.4%	46.60%
Right Ear	1 925	1 072	64.22%	35.78%

The Decision Tree was tested using the entire dataset, which shows that the Decision Tree performed the best for the right ear with 64.22%, whereas the left ear only achieved 53.4% accuracy.

Naïve Bayes

The Naïve Bayes is a supervised machine learning algorithm which was implemented by using the OpenImageR toolbox [56]. The thesis took the same dataset as the other machine learning algorithm to test and train this machine learning algorithm. The results that are obtained from this experiment are shown in Table 4.8:

The Naïve Bayes was tested using the full dataset, which shows that the Naïve Bayes performed the best for the left ear with 59.88% whereas the left ear only achieved 58.33% accuracy.

Table 4.8: Accuracy Achieved for Naïve Bayes

Ear Side	True Identification	False Identification	True Percentage	False Percentage
Left Ear	1 795	1 202	59.88%	40.12%
Right Ear	1 748	1 249	58.33%	41.67%

K-Nearest Neighbor (KNN)

The KNN is a supervised machine learning algorithm which was embedded in the OpenImageR toolbox [56] and was used to implement the KNN. The thesis took the same dataset as the other machine learning algorithm to test and train this machine learning algorithm. The results that are obtained from this experiment are shown in Table 4.9:

Table 4.9: Accuracy Achieved for K-Nearest Neighbor (KNN)

Ear Side	True Identification	False Identification	True Percentage	False Percentage
Left Ear	1 804	1 193	60.20%	39.80%
Right Ear	1 761	1 236	58.75%	41.25%

The KNN was tested using the entire dataset, showing that the KNN performed best for the left ear with 60.20%, whereas the right ear only achieved 58.75% accuracy.

General Discussion of the Classification

As discussed in section 3.1, several empirical experiments were conducted to investigate whether the machine learning algorithm can determine ears from facial and profile images. Decision Trees, Naïve Bayes and K-Nearest Neighbor have commonly used machine learning algorithms. The testing was done by filling the training datasets with randomly chosen images from the original dataset and testing with the different machine learning algorithms.

The tests showed that the K-Nearest Neighbor machine learning algorithm achieved a 60.2% average ear detection rate. The worst machine learning algorithm was Decision Tree, which reached 53.4% ear accuracy identification rate, a variation of 6.8% between the worst and best ear accuracy identification rate, as seen in Table 4.10.

Table 4.10: Accuracy Achieved for All Machine Learning

Ear Side	**Naïve Bayes**	**K-Nearest Neighbor**	**Decision Tree**
Left Ear	59.88%	60.20%	53.4%
Right Ear	58.33%	58.75%	64.22%
Average	59.11%	59.48%	59.97%

4.5 Results from Deep Learning Technique

Deep learning is an AI model that utilises numerous layers to understand the data progressively. This section will mention the structures and contemporary strategies for deep learning designs in AI models that find the correct representation for the inputted information.

Various deep learning Techniques were fine-tuned on all the ear datasets to detect the ear. Each dataset is split into 20% training and 80% test sets. The experiments used the Keras deep learning framework using the TensorFlow back-end. The models were evaluated using the popular evaluation metrics, accuracy, sensitivity, specificity, and area under the curve. The explanation of these rating matrices has been explained above in Chapter 3. All experiment performances are evaluated using the confusion matrix-based performance metrics.

The confusion matrices used to evaluate and classify:

- True Positives (TP) where the ears have been correctly classified as positive.
- True Negatives (TN) where the ears have been correctly classified as negative.
- False Positives (FP) where the ears have been incorrectly classified as positive.
- False Negatives (FN) where the ears have been incorrectly classified as negative.

Convolutional Neural Network

As discussed in 3.1, a Convolutional Neural Network (CNN) is an NN that joins two or more layers to produce one composite layer. The convolutional layer can learn features from the input data. When stacking many convolutional layers, the network can learn a hierarchy of increasingly complex features, [60]. A polling layer is usually added between successive convolutional layers to reinforce essential elements. The CNN reduces the number of parameters passed to the lower layers, depicted in Figure 4.3.

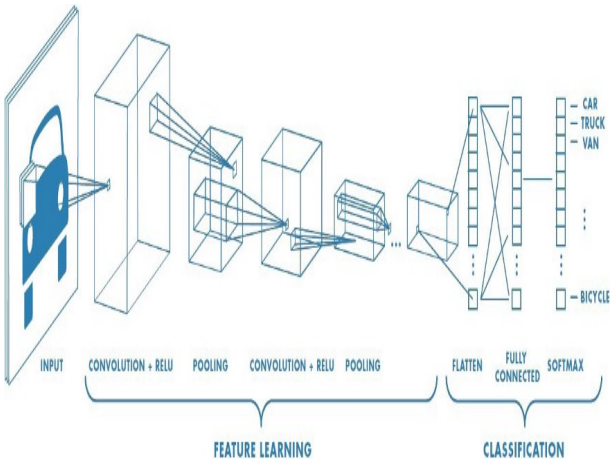


Figure 4.3: Diagram of Convolutional Neural Networks

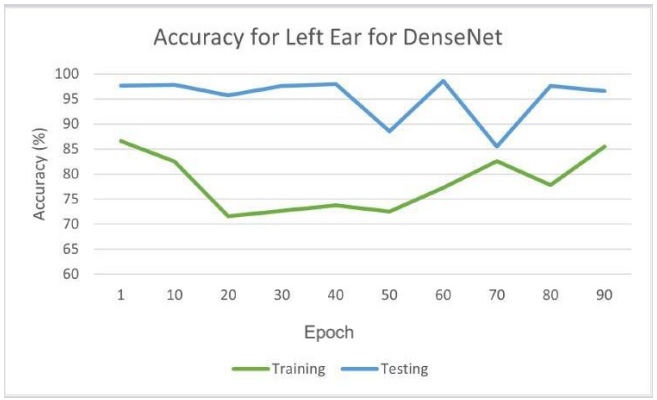
DenseNet Architecture [123] used in research is similar to ResNet but was created to fix the vanishing gradient problem. DenseNet utilises cross-layer connectivity by connecting each preceding layer to the next layer in a feed-forward manner. It was done to fix the ResNet by preserving identity transformations, which increased complexity. As it uses solid blocks, it allows for featuring maps of all previous layers to be used as the inputs into the subsequent layers.

The results are presented in Figures 4.4 and 4.5. It shows the accuracy and loss of these datasets. The DenseNet model used an average of 100 epochs, and the accuracy was determined using the test set. The models performed at extracting and learning discriminative features from the dataset. DenseNet Model for the left ear attains the best accuracy at epochs 70 - 98.65%, and the right ear attains the best accuracy at epochs 40 - 92.33%. The DenseNet results are noted in Table 4.11.

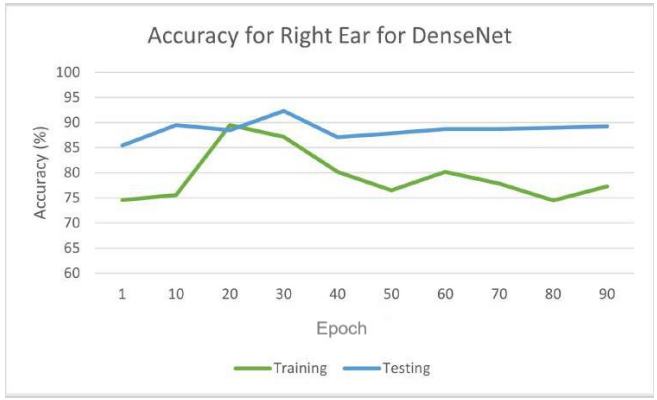
Minimal parameters are needed for the DenseNet, but the right ear performed worse than the left ear, which may have been because the images were down-sampled. It can be seen that performance improves as the model gets deeper. DenseNet begins to converge from 30 iterations, with little noise, until 30 iterations and then stabilises until 50 iterations, when overfitting starts.

Table 4.11: Performance of DenseNet models

	Left Ear		**Right Ear**	
Epoch	**Accuracy**	**Loss**	**Accuracy**	**Loss**
10	97.65	2.35	85.45	14.55
20	97.86	2.14	89.49	10.51
30	95.69	4.31	88.51	11.49
40	97.59	2.41	92.33	7.67
50	97.95	2.05	87.11	12.89
60	88.55	11.45	87.82	12.18
70	98.65	1.35	88.75	11.25
80	85.46	14.54	88.75	11.25
90	97.58	2.42	88.96	11.04
100	96.57	3.43	89.23	10.77

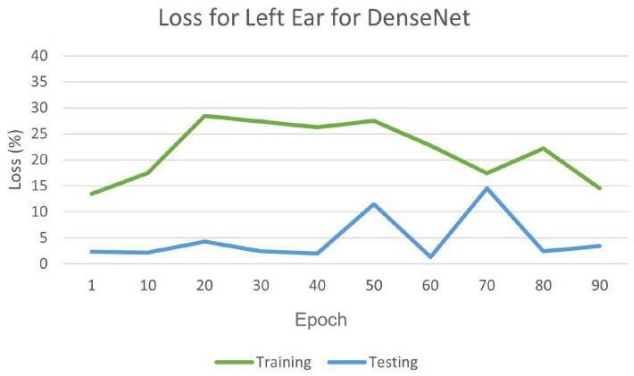


(a) Accuracy for the Left Ear for DenseNet

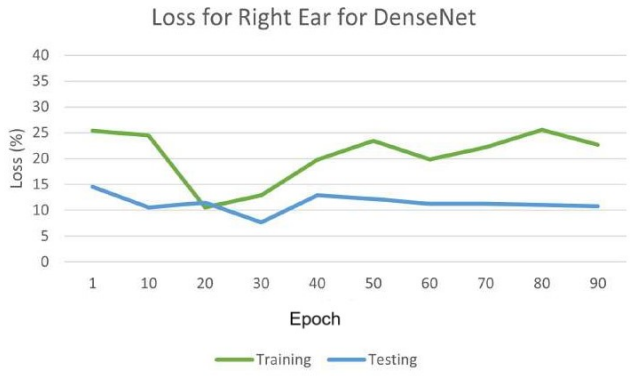


(b) Accuracy for the Right Ear for DenseNet

Figure 4.4: Accuracy for the ear dataset for DenseNet



(a) Loss for the Left Ear for DenseNet



(b) Loss for the Right Ear for DenseNet

Figure 4.5: Loss for the ear dataset for DenseNet

4.5.1 EfficientNet

As discussed in the paper in section 3.2, EfficientNet is a lightweight model based on the auto machine learning framework to develop a baseline EfficientNetB0 network and uniformly scaled up the depth, width and resolution using a simplified and effective compound coefficient to improve EfficientNet models B1-B8. The models performed efficiently and attained superiority over the existing CNN models on the other CNN datasets. EfficientNet is smaller and only requires a few parameters. They are faster and more generalisable in obtaining higher accuracy on other datasets' poplar for the transfer learning task.

The proposed thesis fine-tuned EfficientNet models B0-B8 on the dataset to detect the ears. In transferring the pre-trained EfficientNet to the ear dataset, the models were fine-tuned by adding a global average pooling to reduce the number of parameters and fix overfitting. The dense layers follow the global average pooling with a ReLU activation function and a dropout rate of 0.4 before the output last layer [85]. It is done with the SoftMax activation function to determine the probabilities of the input data to represent the ears, and this can be seen in Equation 4.1.

$$\sigma\left(q\right)_i = \frac{e^{q_i}}{\sum_N^{y=1} e^{q_y}} \tag{4.1}$$

Where σ is the SoftMax activation function, q represents the input vector to the output layer, i depicted from the exponential element e^{q_i}, N is the number of classes, and e^{q_y} represents the output vector of the exponential function.

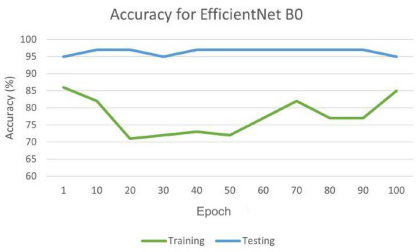
It is known that many iterations could lead to model overfitting, while too few can cause model underfitting. This thesis used an early stopping strategy. It configured approximately 90 training iterations before terminating, catering to early stopping to improve performance, and was applied to control overfitting and use gradient descent. The EfficientNet B0-B8 models were trained with 100 iterations (epochs). The batch size for each iteration was 32, and the momentum equalled 0.2 and was regulated. At the same time, categorical cross entropy is the loss function used to update weights at each iteration. Hyperparameters used were evaluated and found to perform optimally, and this can be defined in Equation 4.2.

$$\alpha = \alpha - n \cdot \triangle_{\alpha} J(\alpha; x^i; y^i) \tag{4.2}$$

Where $\triangle_{\alpha} J$ is the gradient of the loss about α, n is the defined learning rate α is the weight vector, while x and y are the respective training sample and label.

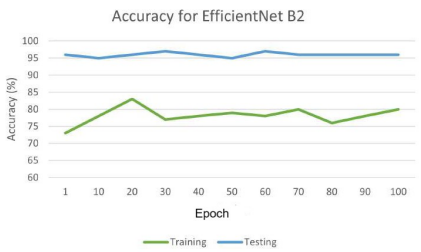
The results are presented in Figures 4.6 and 4.7 depicting the accuracy and loss of these datasets. The various EfficientNet models average at 100 epochs, and the accuracy is determined using the test set. The models performed at extracting and learning discriminative features from the dataset. EfficientNet-B8 attains the best accuracy at 98.45%, and the EfficientNet results are noted in Table 4.12.

An advantage of EfficientNet is that they are smaller with fewer parameters, faster, and obtain transfer learning successfully from the datasets. The worst performing EfficientNet is B2, as shown in Table 4.12. Even though it has minimal parameters, the reason that this performed poorly could have been because the images were down-sampled. It was done to conform to the model's image input size. Performance improves as the model gets deeper. EfficientNet-B0 started poorly, beginning to converge from 30 iterations with little noise until 30 iterations and then stabilising until 50 iterations when overfitting started. The best performing EfficientNet is B8, as shown in Table 4.12 because of the large number of parameters. It began to converge from 60 iterations and then stabilised until 90 iterations when overfitting started. It was found that when the dataset was large and had an equal number of classes, the results achieved were high. Determining the most suitable hyperparameters was one of the challenges faced, as was overfitting, which was limited due to the data samples.

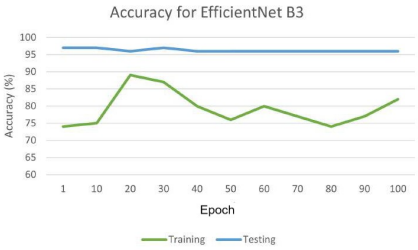


(a) Accuracy for EfficientNet B0

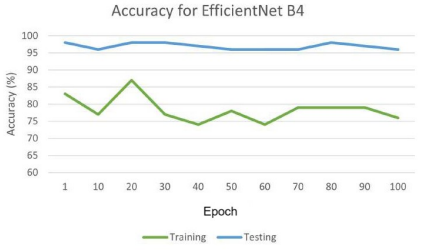
(b) Accuracy for EfficientNet B1



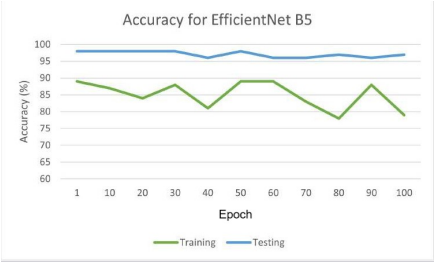
(c) Accuracy for EfficientNet B2



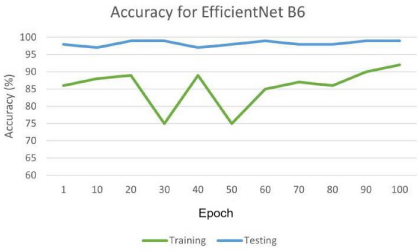
(d) Accuracy for EfficientNet B3



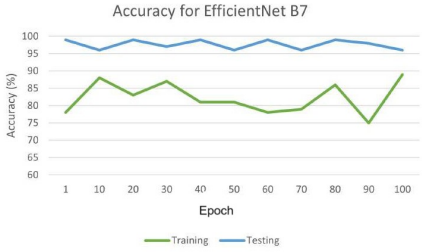
(e) Accuracy for EfficientNet B4



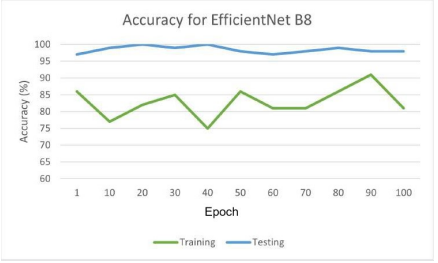
(f) Accuracy for EfficientNet B5



(g) Accuracy for EfficientNet B6



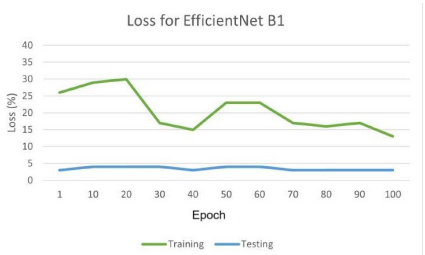
(h) Accuracy for EfficientNet B7



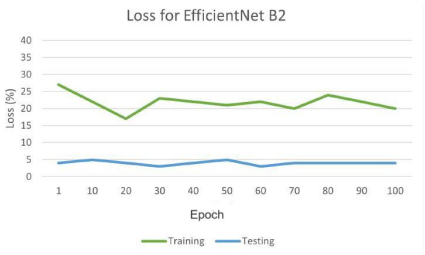
(i) Accuracy for EfficientNet B8

Figure 4.6: Accuracy for the ear dataset of each EfficientNet

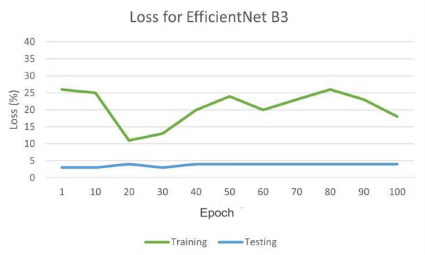
(a) Loss for EfficientNet B0



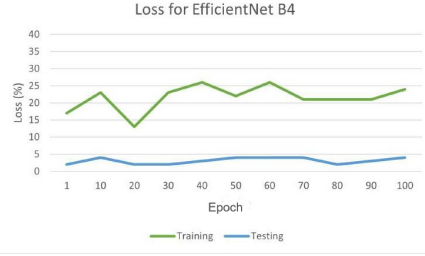
(b) Loss for EfficientNet B1



(c) Loss for EfficientNet B2



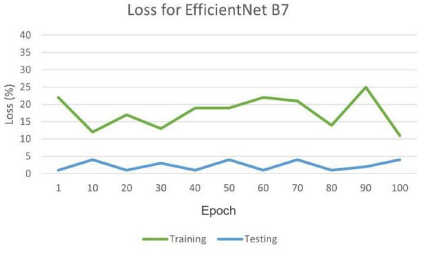
(d) Loss for EfficientNet B3



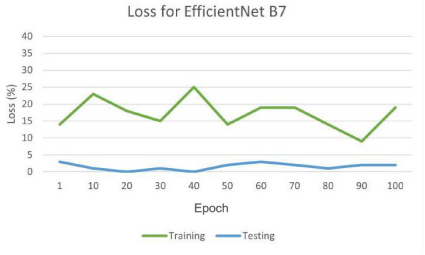
(e) Loss for EfficientNet B4

(f) Loss for EfficientNet B5

(g) Loss for EfficientNet B6



(h) Loss for EfficientNet B7



(i) Loss for EfficientNet B8

Figure 4.7: Loss for the ear dataset of each EfficientNet

Table 4.12: Performance of EfficentNet models

	EfficientNet																	
	B0		**B1**		**B2**		**B3**		**B4**		**B5**		**B6**		**B7**		**B8**	
Epoch	**Acc.**	**Loss**	**Acc.**	**Loss**	**Acc.**	**Loss**	**Acc.**	**Loss**	**Acc.**	**Loss**	**Acc.**	**Loss**	**Acc.**	**Loss**	**Acc.**	**Loss**	**Acc.**	**Loss**
1	95	5	97	3	96	4	97	3	98	2	98	2	98	2	99	1	97	3
10	97	3	96	4	95	5	97	3	96	4	98	2	97	3	96	4	99	1
20	97	3	96	4	96	4	96	4	98	2	98	2	99	1	99	1	100	0
30	95	5	96	4	97	3	97	3	98	2	98	2	99	1	97	3	99	1
40	97	3	97	3	96	4	96	4	97	3	96	4	97	3	99	1	100	0
50	97	3	96	4	95	5	96	4	96	4	98	2	98	2	96	4	98	2
60	97	3	96	4	97	3	96	4	96	4	96	4	99	1	99	1	97	3
70	97	3	97	3	96	4	96	4	96	4	96	4	98	2	96	4	98	2
80	97	3	97	3	96	4	96	4	98	2	97	3	98	2	99	1	99	1
90	97	3	97	3	96	4	96	4	97	3	96	4	99	1	98	2	98	2
100	95	5	97	3	96	4	96	4	96	4	97	3	99	1	96	4	98	2

4.5.2 Transformer Network Architecture

As discussed in the paper in section 3.3, transfer learning was adopted and helped with the pre-trained CNN model for large datasets to learn features of the target (right and left ears). It will transfer the features of the deep CNN models learned on other CNN models to this dataset. The number of deep CNN model parameters increases as the network gets deeper, which is used to achieve improved efficiency.

Hence, it requires many datasets for training, making it computationally complex. Applying these models directly on small and new databases results in feature extraction bias, overfitting, and poor generalisation. The pre-trained CNN modified and fine-tuned its structure to suit the dataset given. This concept of transfer learning is computationally expensive, has less training time, overcomes dataset limitations, improves performance, and is faster than training a model from the beginning. The proposed structure is represented in Figure 4.8.

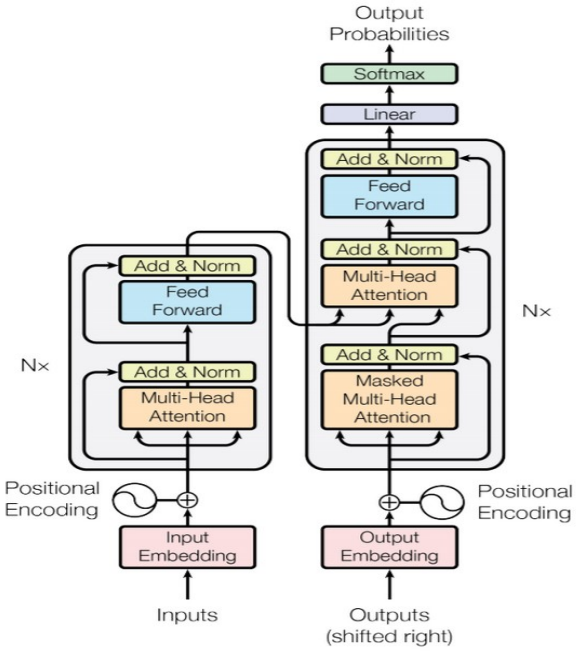


Figure 4.8: Block structure of the proposed model

The Transformer Network is an encoder-decoder architecture based on attention layers. The difference between a Convolutional Neural Network and a Transformer Network is that the data can be passed in parallel, which means that the GPU can be utilised effectively and efficiently. The speed of the training is also increased by

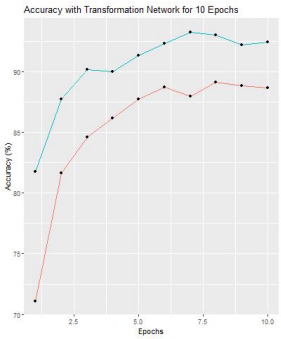
processing it in parallel. It is seen that the Transformation Network is based on a multi-headed attention layer, and by doing this, the vanishing gradient issue is overcome.

The results are presented in Figures 4.9 and 4.10, which is the accuracy and loss of these datasets. In the Transformer Network, the test set determines accuracy at different epochs. The models performed at extracting and learning discriminative features from the dataset. Transformer Network with 50 and 90 Epochs attains the best accuracy of 92.60 and 92.56%, and the Transformer Network results are noted in Table 4.13.

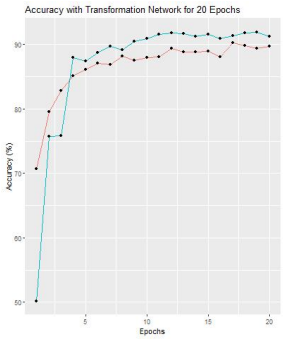
An advantage of Transformer Networks is that they are smaller with fewer parameters, faster, and obtain transfer learning successfully from the datasets. The worst performing was 20 epochs, as shown in Table 4.13. The reason that this performed poorly could have been because it did not have enough data to learn from. It was to conform to the model's image input size. Performance improves as the model gets deeper. On average, overfitting occurred at 30 iterations and stabilised at around 50. The best-performing Transformer Network is at epochs 50 and 90, as shown in Table 4.13 because of the large number of parameters. It began to converge from 30 iterations and then stabilised until 50 iterations when overfitting started. Determining the most suitable hyperparameters was one of the challenges faced, as was overfitting, which was limited due to the data samples.

Table 4.13: Performance of Transformer Network

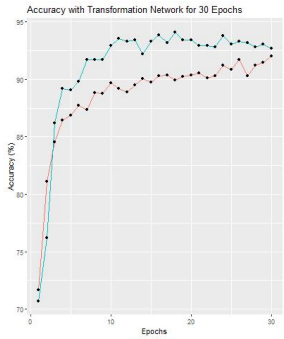
Epochs	**Accuracy (%)**	**Loss (%)**
10	90.42	9.58
20	87.06	12.94
30	91.10	8.90
40	90.97	9.03
50	92.60	7.40
60	91.74	8.26
70	91.81	8.19
80	92.18	7.82
90	92.56	7.44
100	91.91	8.09



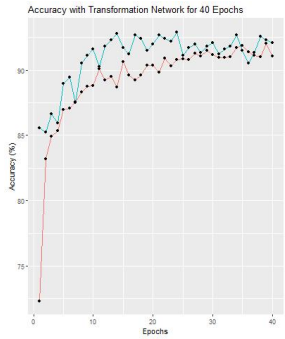
(a) Accuracy for Transformer Network for 10 Epochs



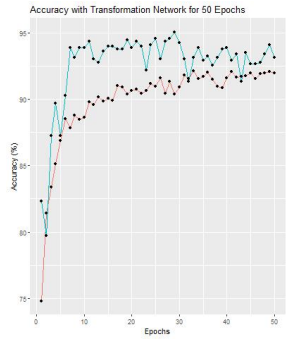
(b) Accuracy for Transformer Network for 20 Epochs



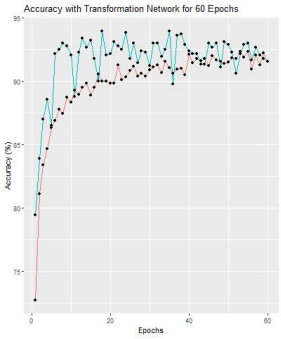
(c) Accuracy for Transformer Network for 30 Epochs



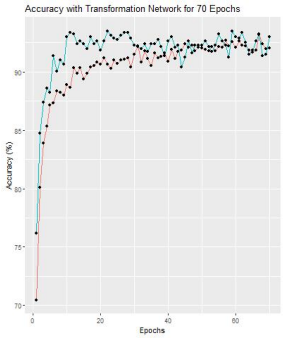
(d) Accuracy for Transformer Network for 40 Epochs



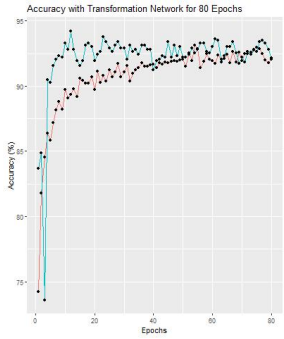
(e) Accuracy for Transformer Network for 50 Epochs



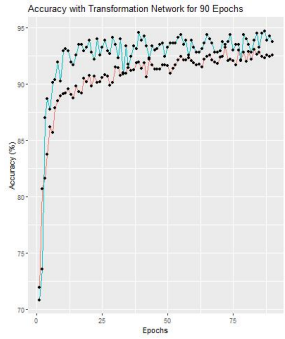
(f) Accuracy for Transformer Network for 60 Epochs



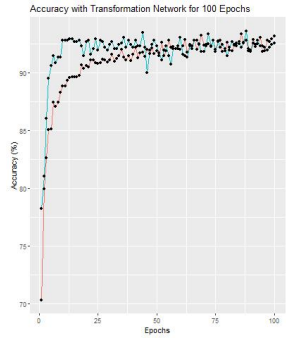
(g) Accuracy for Transformer Network for 70 Epochs



(h) Accuracy for Transformer Network for 80 Epochs

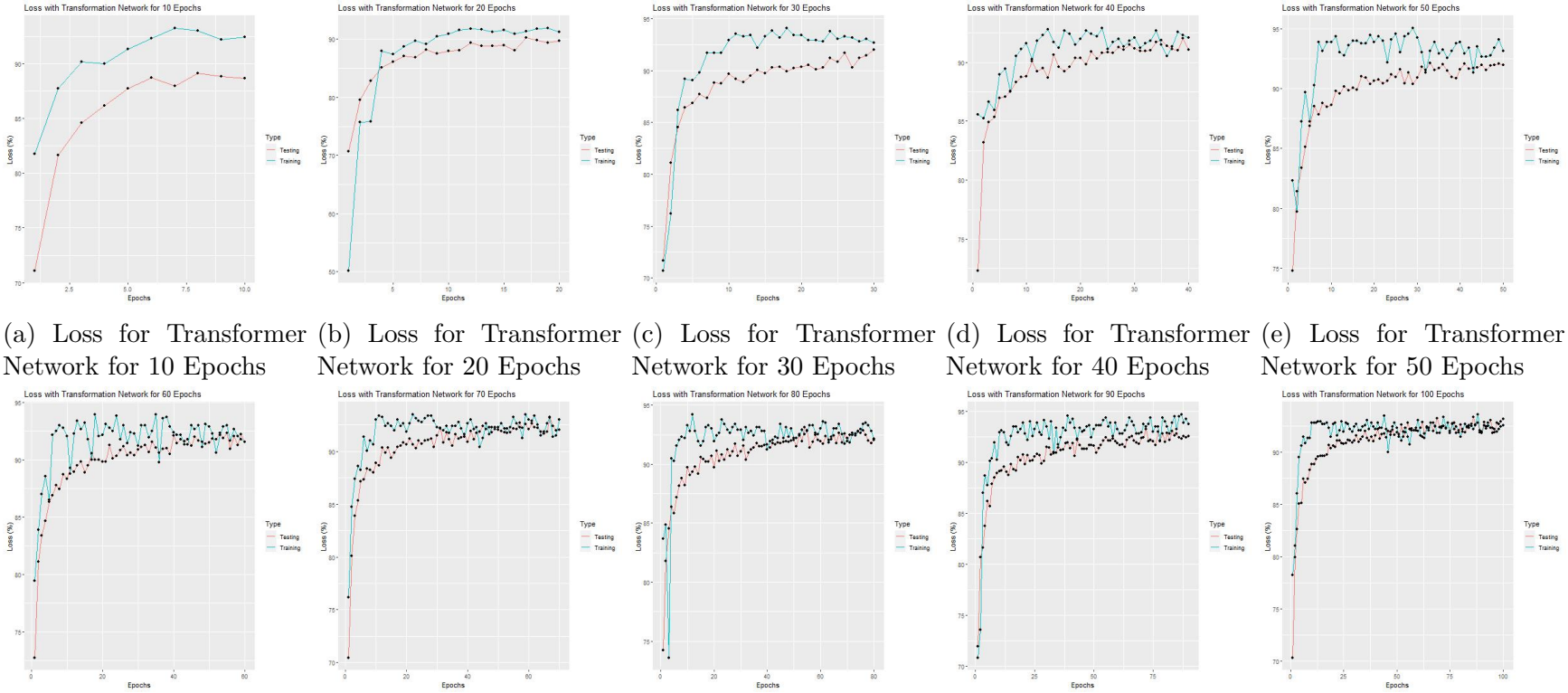


(i) Accuracy for Transformer Network for 90 Epochs



(j) Accuracy for Transformer Network for 100 Epochs

Figure 4.9: Accuracy for the ear dataset of each Transformer Network



(a) Loss for Transformer Network for 10 Epochs (b) Loss for Transformer Network for 20 Epochs (c) Loss for Transformer Network for 30 Epochs (d) Loss for Transformer Network for 40 Epochs (e) Loss for Transformer Network for 50 Epochs

(f) Loss for Transformer Network for 60 Epochs (g) Loss for Transformer Network for 70 Epochs (h) Loss for Transformer Network for 80 Epochs (i) Loss for Transformer Network for 90 Epochs (j) Loss for Transformer Network for 100 Epochs

Figure 4.10: Loss for the ear dataset of each Transformer Network

4.5.3 Lightweight Deep Learning with Model Compression

As discussed in the paper in section 3.4, Image pre-processing is a considerable part of the deep learning task. Most CNN models require a large dataset to learn to discriminate features suitably for making predictions and obtaining a good performance. As images in the datasets are of different sizes, the inputted images must be resized to conform to all the other CNN models. However, the features need to be preserved when resizing is performed.

The ReducedFireNet Model is a convolutional neural network (CNN) type that uses multiple fire modules. The ReducedFireNet Model consists of four Fire modules consisting of two layers: a 1 x 1 convolutional layer and a concatenation of 1 x 1 and 3 x 3 convolutional layers. A Max-pooling layer is applied to the first three Fire modules. The reason for the Max-pooling layer is to reduce the size of the input and increase computation time. It is also seen that it helps detect more robust features such as ears. The last layer of the matrix has a GlobalAveragePooling layer applied. A dropout layer is applied once the Max pooling layer has been applied on the second layer. The reason that this is done is to reduce overfitting and to make it more robust. Rectified Linear Unit (ReLU) is the activation function applied to each convolutional layer's output. Once the dense layer is created, a Softmax activation function, shown in Equation 4.3, is applied to classify the ear. The depiction of how the ReducedFireNet Model works is shown in Figure, 4.11.

$$\sigma\left(q\right)_i = \frac{e^{q_i}}{\sum_N^{y=1} e^{q_y}} \tag{4.3}$$

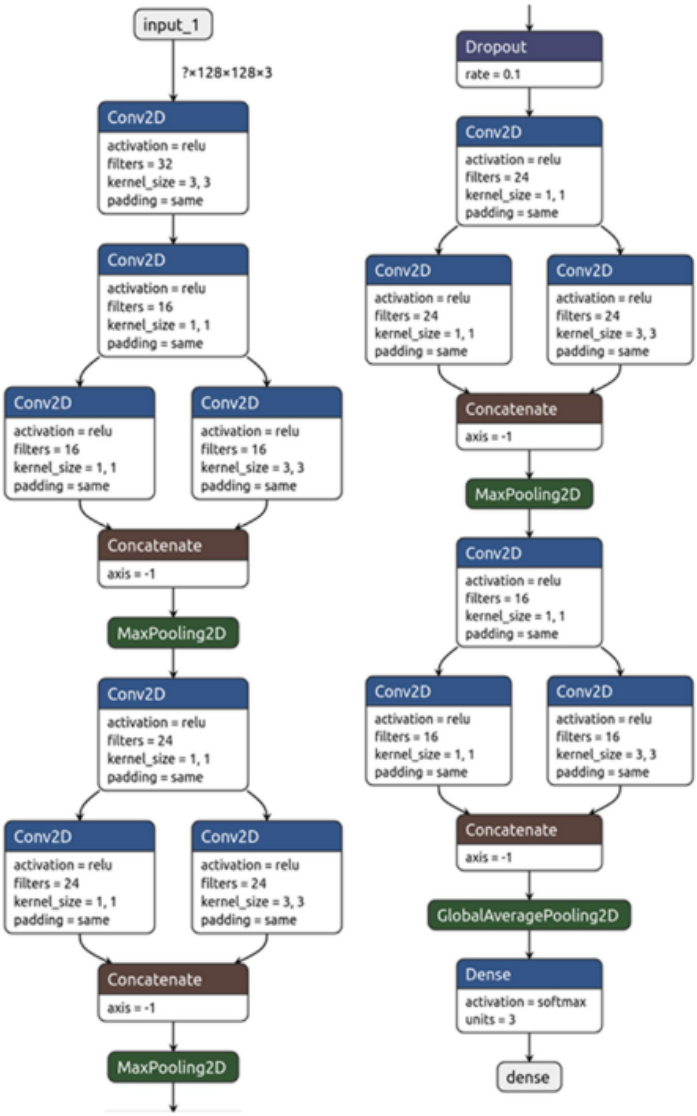


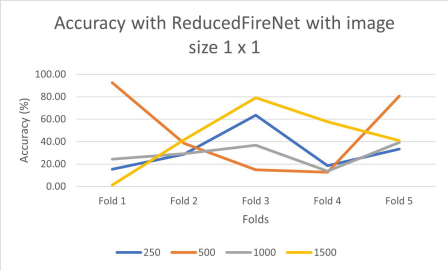
Figure 4.11: ReducedFireNet Model Configuration

The results are presented in Figures 4.12; this is the accuracy and loss of these datasets. The accuracy of the ReducedFireNet Model at a different number of images is used in the testing set. The models performed at extracting and learning discriminative features from the dataset. ReducedFireNet Model with the max number of images and image size 128 x 128 was the best accuracy at 87.91%.

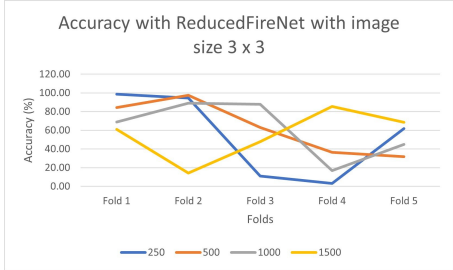
An advantage of the ReducedFireNet Model is that it reduces the input size and increases computation time. The worst performance was with the max number of images and an image size 1 x 1, as shown in Figure 4.12. The reason that this performed poorly could have been because it did not have enough data to learn from. It was done to conform to the model's image input size. It can be seen that performance improves as the model gets deeper. On average, it was seen that overfitting occurred at the max number of images and image size 16 x 16 iterations and stabilised at around the max number of images and image size 128 x 128, as shown in Table 4.15 because of the large number of parameters.

Table 4.14: Performance of ReducedFireNet Model

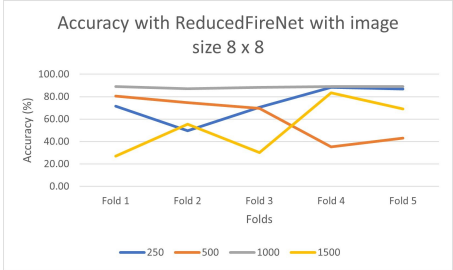
Size of image	**Accuracy (%)**	**Loss (%)**
1 x 1	44.15	55.85
3 x 3	55.44	44.56
8 x 8	53.00	47.00
16 x 16	44.98	55.02
32 x 32	52.56	47.44
64 x 64	73.87	26.13
128 x 128	87.91	12.09



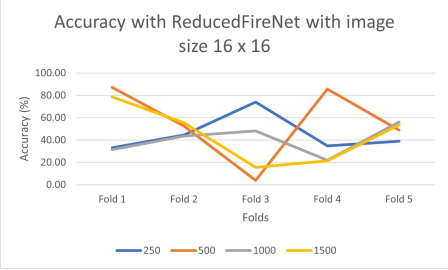
(a) Accuracy for ReducedFireNet Model with image size 1 x 1



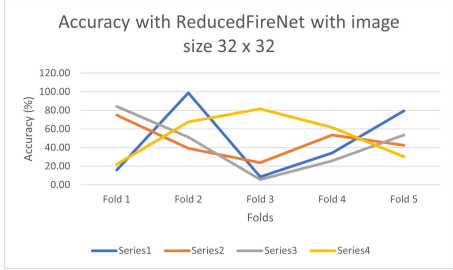
(b) Accuracy for ReducedFireNet Model with image size 3 x 3



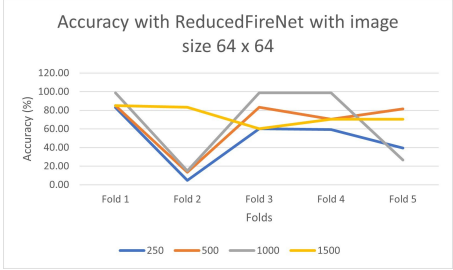
(c) Accuracy for ReducedFireNet Model with image size 8 x 8



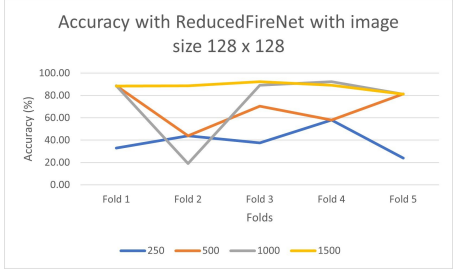
(d) Accuracy for ReducedFireNet Model with image size 16 x 16



(e) Accuracy for ReducedFireNet Model with image size 32 x 32



(f) Accuracy for ReducedFireNet Model with image size 64 x 64



(g) Accuracy for ReducedFireNet Model with image size 128 x 128

Figure 4.12: Accuracy for the ear dataset of each ReducedFireNet Model

4.5.4 General Discussion of the Deep Learning

This section explored efficient deep learning-based models for accurate ear biometrics recognition. The ears were extracted and identified from 2D profiles and facial images, focusing on both left and right ears. A range of machine learning techniques were explored, and innovative deep learning techniques: Transformer Network Architecture, 3.3, Lightweight Deep Learning with Model Compression3.4, and EfficientNet,3.2.

The experimental results showed that the Transformer Network achieved a high accuracy of 92.60% and 92.56% with epochs of 50 and 90, respectively. The proposed ReducedFireNet Model reduces the input size and increases computation time, but it detects more robust ear features. The EfficientNet variant B8 achieved a classification accuracy of 98.45%. The results showed that deep learning models can improve ear biometrics recognition when both ears are computed, as seen in Table 4.15 and 4.16.

Table 4.15: Performance of ReducedFireNet Model

Size of image	**Accuracy (%)**	**Loss (%)**
1 x 1	44.15	55.85
3 x 3	55.44	44.56
8 x 8	53.00	47.00
16 x 16	44.98	55.02
32 x 32	52.56	47.44
64 x 64	73.87	26.13
128 x 128	87.91	12.09

Table 4.16: Accuracy Achieved for All Deep Learning

	DenseNet		**EfficientNet - B8**		**Transformer Network**	
Epoch	**Accuracy(%)**	**Loss(%)**	**Accuracy(%)**	**Loss(%)**	**Accuracy(%)**	**Loss(%)**
1	-	-	97	3	-	-
10	92	8	99	1	90	10
20	94	6	100	0	87	13
30	92	8	99	1	91	9
40	95	5	100	0	91	9
50	93	7	98	2	93	7
60	88	12	97	3	92	8
70	94	6	98	2	92	8
80	87	13	99	1	92	8
90	93	7	98	2	93	7
100	93	7	98	2	92	8

4.6 Conclusion

In this chapter, the proposed framework is discussed and explained. The experimental results have been presented and discussed, showing that the proposed framework performs better and produces a more accurate ear identification rate for an image than other research studies.

The next chapter concludes the thesis and provides direction for future work.

Chapter 5

Conclusion and Future Works

5.1 Summary of work

This chapter concludes the thesis with a summary of the work presented in the previous chapters, discusses potential solutions to some of the thesis's limitations and presents future work that will further enhance the applicability of the ear identification system to real-world applications.

The thesis started by introducing the topic and explaining the importance and application areas of ear biometrics in Chapter 1. The thesis further presented a study of the state-of-the-art techniques used to obtain ear biometrics and a critical analysis of the performance of some of those models when evaluated on popular ear biometrics datasets in Chapter 2.

In Chapter 3, and the thesis introduced the methods and materials employed to classify ear images left and right. The thesis then explored various machine learning techniques such as Naïve Bayes, Decision Tree, K-Nearest Neighbor, and innovative deep learning techniques: Transformer Network Architecture, Lightweight Deep Learning with Model Compression and EfficientNet.

In the last chapter, the thesis presented a comprehensive analysis and discussion of our experimental results for the proposed machine learning algorithms and innovative deep learning techniques for ear identification from profile and facial images. The thesis further evaluated the performance accuracy of our system and compared it with state-of-the-art results. It describes our empirical and experimental results and the parameters for designing our ear models. Also, several tables are presented to describe the impact of different parameters on the prediction rate of ear classification.

Biometrics is the recognition of a human using biometric characteristics for identification, which may be physiological or behavioural. The physiological biometric features are the face, ear, iris, fingerprint, and handprint; behavioural biometrics are signatures, voice, gait patterns, and keystrokes. Numerous systems have been developed to distinguish biometric traits used in multiple applications, such as forensic investigations and security systems. With the COVID-19 pandemic, facial recognition systems failed due to users wearing masks; however, human ear recognition proved more suitable as it is visible. This thesis reviews the source of ear modelling, details the algorithms, methods and processing steps and finally tracks the input database's error and limitations for the ear identification results.

This thesis explores efficient deep learning-based models for accurate ear biometrics recognition. The ears were extracted and identified from 2D profiles and facial images, focusing on both left and right ears. A range of machine learning techniques were explored, such as Naïve Bayes, DecisionTree, K-Nearest Neighbor, and innovative deep learning techniques; Transformer Network Architecture, Lightweight Deep Learning with Model Compression and EfficientNet.

The feature vector combines Haralick texture Moments, Zernike Moments, Gabor Filter and Local Binary Pattern. All these feature vectors are then fused and nor-

malised to obtain a result. Naïve Bayes achieved 58.33% accuracy for the right ear, K-Nearest Neighbor achieved 60.20% for the left ear, and Decision Tree achieved 55.72% for the left ear.

DenseNet Deep Learning achieved 95.36% accuracy for the left ear and 88.64% for the right ear. The total sum of the left and right ear identification rates of 92.00% was achieved. Further investigation was done to ascertain if the ear could be identified by developing a CNN using EfficientNet. The nine variants of EfficientNet were fine-tuned and implemented on multiple publicly available ear databases. These experiments showed that EfficientNet variant B8 achieved the best accuracy of 98.45%, proving that the EfficientNet on ear recognition performed superior over the existing CNN models.

Additional tests were carried out to see if accuracy could be improved with a pretrained dataset. The methods used were a lightweight method called the ReducedFireNet and a Transformation Network model. Both models used the same images in the testing set. The ReducedFireNet model performs by extracting and learning discriminative features from the dataset. The best accuracy was 87.91% with an image size of 128 x 128. The ReducedFireNet Model reduces the input size and increases computation time, but it detects more robust ear features. The next model considered was the Transformer Network. The test set determines accuracy at different epochs. The models performed at extracting and learning discriminative features from the dataset. Transformer Network with 50 and 90 epochs attains the best accuracy of 92.60 and 92.56%, respectively.

The results showed that deep learning models can improve ear biometrics recognition when both ears are computed.

5.2 Contribution to Knowledge

This thesis's main contribution is improving the accuracy of characterising ear identification with facial or profile images. According to the set objectives, the following were achieved:

- The thesis presented a comprehensive review of various traditional ear identification tests. It looked at a range of machine learning techniques employed in biometric and ear identification; this was done by doing testing against specific machine learning techniques using prominent ear datasets. The thesis highlights state-of-the-art deep learning architectures, stating their strengths and weaknesses, evaluation metrics, and performance of the current ear identification classifiers.
- The thesis designed a robust image-processing algorithm that handles the pre-processing of images, which deals with contrast enhancement, feature extraction, and noise removal, before feeding it to the design machine learning and CNN model.
- It developed novel deep learning architectures and fine-tuned pre-trained CNN to detect ear areas and effectively classify them as either the left or right ear.
- This thesis demonstrates that a model's performance accuracy and sensitivity can be improved through deep learning.
- Demonstrate that training a model on a dataset and testing it on a different dataset that does not originate from the training subset gives a better generalisation of the ear identification accuracy.

Appendix A

Proof of Submission of Unpublished Article Number Five

ICCCI 2024 submission 97

I2

ICCCI 2024<iccci2024@easychair.org>
To: Aimee Booysens (210501411)

Dear authors,

We received your submission to ICCCI 2024 (16th International Conference on Computational Collective Intelligence):

Authors : Aimee Booysens and Serestina Viriri
Title : Lightweight Deep Learning with Model Compression for Ear Recognition
Number : 97
Track : Main Track

The submission was uploaded by Serestina Viriri <viriris@ukzn.ac.za>.
You can access it via the ICCCI 2024 EasyChair Web page

https://easychair.org/conferences/?conf=iccci2024

Thank you for submitting to ICCCI 2024.

Best regards,
EasyChair for ICCCI 2024.

Bibliography

[1] Ayman Abaza. *High performance image processing techniques in automated identification systems*. West Virginia University, 2008. 88

[2] Ayman Abaza, Christina Hebert, and Mary Ann F Harrison. Fast learning ear detection for real-time surveillance. In *2010 fourth IEEE international conference on biometrics: theory, applications and systems (BTAS)*, pages 1–6. IEEE, 2010.

[3] Ayman Abaza, Arun Ross, Christina Hebert, Mary Ann F. Harrison, and Mark S. Nixon. A survey on ear biometrics. *ACM Comput. Surv.*, 45(2), March 2013. 2

[4] Ayman Abaza, Arun Ross, Christina Hebert, Mary Ann F Harrison, and Mark S Nixon. A survey on ear biometrics. *ACM computing surveys (CSUR)*, 45(2):1–35, 2013.

[5] Mohamed Abdel-Mottaleb and Jindan Zhou. Human ear recognition from face profile images. In *International Conference on Biometrics*, pages 786–792. Springer, 2006.

[6] Thomas Abeel, Yves Van de Peer, and Yvan Saeys. Java-ml: A machine learning library. *Journal of Machine Learning Research*, 10:931–934, 2009.

[7] Ivo Alberink and Arnout Ruifrok. Performance of the fearid earprint identification system. *Forensic science international*, 166(2-3):145–154, 2007. 88

[8] Ahmed M Alkababji and Omar H Mohammed. Real time ear recognition using deep learning. *Telkomnika*, 19(2):523–530, 2021.

[9] Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S

Awwal, and Vijayan K Asari. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, 2018.

[10] Hammam Alshazly, Christoph Linse, Erhardt Barth, and Thomas Martinetz. Ensembles of deep learning models and transfer learning for ear recognition. *Sensors*, 19(19):4139, 2019.

[11] Hammam Alshazly, Christoph Linse, Erhardt Barth, and Thomas Martinetz. Handcrafted versus cnn features for ear recognition. *Symmetry*, 11(12):1493, 2019.

[12] Hammam Alshazly, Christoph Linse, Erhardt Barth, and Thomas Martinetz. Deep convolutional neural networks for unconstrained ear recognition. *IEEE Access*, 8:170295–170310, 2020.

[13] Gandhimathi Amirthalingam and G Radhamani. A multimodal approach for face and ear biometric system. *International Journal of Computer Science Issues (IJCSI)*, 10(5):234, 2013.

[14] Saeeduddin Ansari and Phalguni Gupta. Localization of ear using outer helix curve of the ear. In *2007 International Conference on Computing: Theory and Applications (ICCTA'07)*, pages 688–692. IEEE, 2007.

[15] Banafshe Arbab-Zavar and Mark S Nixon. On shape-mediated enrolment in ear biometrics. In *International Symposium on Visual Computing*, pages 549–558. Springer, 2007.

[16] Sarah Adel Bargal, Alexander Welles, Cliff R Chan, Samuel Howes, Stan Sclaroff, Elizabeth Ragan, Courtney Johnson, and Christopher Gill. Image-based ear biometric smartphone app for patient identification in field settings. In *VISAPP (3)*, pages 171–179, 2015.

[17] Venkatesha Basrur, Feng Yang, Tsuneto Kushimoto, Youichiro Higashimoto, Ken-ichi Yasumoto, Julio Valencia, Jacqueline Muller, Wilfred D Vieira, Hidenori Watabe, Jeffrey Shabanowitz, et al. Proteomic analysis of early melanosomes: identification of novel melanosomal proteins. *Journal of proteome research*, 2(1):69–79, 2003.

[18] Sebastian Berisha. Image classification using gabor filters and machine learning. 2009.

[19] Bir Bhanu and Hui Chen. *Ear Biometrics, 3D*, pages 241–248. Springer US, Boston, MA, 2009. 3

[20] Lei Bi, Jinman Kim, Euijoon Ahn, Ashnil Kumar, Michael Fulham, and Dagan Feng. Dermoscopic image segmentation via multistage fully convolutional networks. *IEEE Transactions on Biomedical Engineering*, 64(9):2065–2074, 2017.

[21] Francesco Bonanno, Giacomo Capizzi, Grazia Lo Sciuto, Christian Napoli, Giuseppe Pappalardo, and Emiliano Tramontana. A cascade neural network architecture investigating surface plasmon polaritons propagation for thin metals in openmp. In *International Conference on Artificial Intelligence and Soft Computing*, pages 22–33. Springer, 2014.

[22] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[23] Mark Burge and Wilhelm Burger. Ear biometrics. In *Biometrics*, pages 273–285. Springer, 1996. 5, 6

[24] M. A. Carreira-Perpinan. Compression neural networks for feature extraction: Application to human recognition from ear images, 1995. 87

[25] Modesto Castrillon-Santana, Javier Lorenzo-Navarro, and Daniel Hernandez-Sosa. An study on ear detection and its applications to face detection. In *Conference of the Spanish Association for Artificial Intelligence*, pages 313–322. Springer, 2011.

[26] Kyong Chang, Kevin W Bowyer, Sudeep Sarkar, and Barnabas Victor. Comparison and combination of ear and face images in appearance-based biometrics. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1160–1165, 2003.

[27] Hui Chen and Bir Bhanu. Human ear recognition in 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):718–737, 2007.

[28] Long Chen, Zhichun Mu, Baoqing Zhang, and Yi Zhang. Ear recognition from one sample per person. *PloS one*, 10(5):e0129505, 2015.

[29] Francois Chollet et al. *Deep learning with Python*, volume 361. Manning New York, 2018.

[30] Michal Choraś. Ear biometrics based on geometrical method of feature extraction. In *International Conference on Articulated Motion and Deformable Objects*, pages 51–61. Springer, 2004.

[31] Michal Choras. Image feature extraction methods for ear biometrics–a survey. In *6th International Conference on Computer Information Systems and Industrial Management Applications (CISIM'07)*, pages 261–265. IEEE, 2007.

[32] Debbrota Paul Chowdhury, Sambit Bakshi, Guodong Guo, and Pankaj Kumar Sa. On applicability of tunable filter bank based feature for ear biometrics: a study from constrained to unconstrained. *Journal of medical systems*, 42(1):1–20, 2018.

[33] Dan Cireşan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*, 2012.

[34] Alastair H Cummings, Mark S Nixon, and John N Carter. A novel ray analogy for enrolment of ear biometrics. In *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–6. IEEE, 2010.

[35] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015.

[36] Samuel Dodge, Jinane Mounsef, and Lina Karam. Unconstrained ear recognition using deep neural networks. *IET Biometrics*, 7(3):207–214, 2018.

[37] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[38] Žiga Emeršič, Luka L Gabriel, Vitomir Štruc, and Peter Peer. Convolutional encoder–decoder networks for pixel-wise ear detection and segmentation. *IET Biometrics*, 7(3):175–184, 2018.

[39] Ziga Emersic, BS Harish, Weronika Gutfeter, Jalil Nourmohammadi Khiarak, Andrzej Pacut, Earnest Hansley, Mauricio Pamplona Segundo, Sudeep Sarkar, Hyeonjung Park, Gi Pyo Nam, et al. The unconstrained ear recognition challenge 2019-arxiv version with appendix. *arXiv preprint arXiv:1903.04143*, 2019.

[40] Žiga Emeršič, Blaž Meden, Peter Peer, and Vitomir Štruc. Evaluation and analysis of ear recognition models: performance, complexity and resource requirements. *Neural computing and applications*, pages 1–16, 2018.

[41] Ziga Emersic and Peter Peer. Ear biometric database in the wild. In *2015 4th international work conference on bioinspired intelligence (IWOBI)*, pages 27–32. IEEE, 2015.

[42] Žiga Emeršič, Dejan Štepec, Vitomir Štruc, and Peter Peer. Training convolutional neural networks with limited training data for ear recognition in the wild. *arXiv preprint arXiv:1711.09952*, 2017.

[43] Žiga Emeršič, Dejan Štepec, Vitomir Štruc, Peter Peer, Anjith George, Adii Ahmad, Elshibani Omar, Terranee E Boult, Reza Safdaii, Yuxiang Zhou, et al. The unconstrained ear recognition challenge. In *2017 IEEE international joint conference on biometrics (IJCB)*, pages 715–724. IEEE, 2017. 87

[44] Ziga Emersic, Vitomir Struc, and Peter Peer. Ear recognition: More than a survey. *Neurocomputing*, 255:26–39, 2017. 5

[45] Ziga Emersic, Vitomir Struc, and Peter Peer. Ear recognition: More than a survey. *Neurocomputing*, 255:26–39, 2017. 87

[46] Žiga Emeršič, Vitomir Štruc, and Peter Peer. Ear recognition: More than a survey. *Neurocomputing*, 255:26–39, 2017.

[47] Ziga Emersic, Vitomir Struc, and Peter Peer. Ear Recognition: More Than a Survey. *Neurocomputing*, 2017.

[48] Žiga Emeršič and Peter Peer. Ear biometric database in the wild. pages 27–32, 07 2015. 87

[49] Ziga Emesic and Peter Peer. Toolbox for ear biometric recognition evaluation. In *IEEE EUROCON 2015-International Conference on Computer as a Tool (EUROCON)*, pages 1–6. IEEE, 2015.

[50] Luis Alvarez Esther Gonzalez and Luis Mazorra. *AMI Ear Database*, 2018 (accessed February 3, 2014).

[51] Dariusz Frejlichowski and Natalia Tyszkiewicz. The west pomeranian university of technology ear database – a tool for testing biometric algorithms. 87

[52] Stefan Fritsch, Frauke Guenther, and Maintainer Frauke Guenther. Package 'neuralnet'. *Training of Neural Networks*, 2019.

[53] Yun Fu, Liangliang Cao, Guodong Guo, and Thomas S Huang. Multiple feature fusion by subspace learning. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 127–134, 2008.

[54] Pedro Luis Galdámez, William Raveane, and Angélica González Arrieta. A brief review of the ear recognition process using deep neural networks. *Journal of Applied Logic*, 24:62–70, 2017.

[55] Esther Gonzalez, Luis Alvarez, and Luis Mazorra. Ami ear database, 2012. 87

[56] M. Haghighat, S. Zonouz, and M. Abdel-Mottaleb. Cloudid: Trustworthy cloud-based and cross-enterprise biometric identification. *Expert Systems with Applications*, 42:7905–7916, 2015. 89, 90, 91, 94, 95

[57] Rami R Hallac, Jeon Lee, Mark Pressler, James R Seaward, and Alex A Kane. Identifying ear abnormality from 2d photographs using convolutional neural networks. *Scientific reports*, 9(1):1–6, 2019.

[58] Earnest E Hansley, Maurício Pamplona Segundo, and Sudeep Sarkar. Employing fusion of learned and handcrafted features for unconstrained ear recognition. *IET Biometrics*, 7(3):215–223, 2018.

[59] Earnest Eugene Hansley. Identification of individuals from ears in real world conditions. 2018.

[60] Jeff Heaton. Ian goodfellow, yoshua bengio, and aaron courville: Deep learning, 2018. 97

[61] Vinh Truong Hoang. Earvn1.0: A new large-scale ear images dataset in the wild. *Data in Brief*, 27:104630, 2019. 87

[62] Raabid Hussain, Alain Lalande, Kibrom Berihu Girum, Caroline Guigou, and Alexis Bozorg Grayeli. Automatic segmentation of inner ear on ct-scan using

auto-context convolutional neural network. *Scientific Reports*, 11(1):1–10, 2021.

[63] Fevziye Irem Eyiokur, Dogucan Yaman, and Hazım Kemal Ekenel. Domain adaptation for ear recognition using deep convolutional neural networks. *arXiv e-prints*, pages arXiv–1803, 2018.

[64] Nursuriati Jamil, AA Almisreb, Syed Mohd Zahid Syed Zainal Ariffin, N Md Din, and Raseeda Hamzah. Can convolution neural network (cnn) triumph in ear recognition of uniform illumination invariant? 2018.

[65] Jens Kleesiek, Gregor Urban, Alexander Hubert, Daniel Schwarz, Klaus Maier-Hein, Martin Bendszus, and Armin Biller. Deep mri brain extraction: A 3d convolutional neural network for skull stripping. *NeuroImage*, 129:460–469, 2016.

[66] Aviwe Kohlakala and Johannes Coetzer. Ear-based biometric authentication through the detection of prominent contours. *SAIEE Africa Research Journal*, 112(2):89–98, 2021.

[67] A Kumar. Iit delhi ear database version 1.0., 2007. 87

[68] Ajay Kumar and Chenye Wu. Automated human identification using ear imaging. *Pattern Recognition*, 45(3):956–968, 2012. 5, 6

[69] Francis Quintal Lauzon. An introduction to deep learning. In *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, pages 1438–1439. IEEE, 2012.

[70] Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.

[71] Zhang Li, Zheyu Hu, Jiaolong Xu, Tao Tan, Hui Chen, Zhi Duan, Ping Liu, Jun Tang, Guoping Cai, Quchang Ouyang, et al. Computer-aided diagnosis of lung carcinoma using deep learning-a pilot study. *arXiv preprint arXiv:1803.05471*, 2018.

[72] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[73] Alessio Lomuscio and Lalit Maganti. An approach to reachability analysis for feed-forward relu neural networks. *arXiv preprint arXiv:1706.07351*, 2017.

[74] Daniel Lowd and Pedro Domingos. Naive bayes models for probability estimation. In *Proceedings of the 22nd international conference on Machine learning*, pages 529–536. ACM, 2005.

[75] Michal Majka. *naivebayes: High Performance Implementation of the Naive Bayes Algorithm in R*, 2019. R package version 0.9.7.

[76] Jameson Merkow, Brendan Jou, and Marios Savvides. An exploration of gender identification using only the periocular region. In *BTAS*, pages 1–5, 2010.

[77] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, Gilbert Maitre, et al. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966. Citeseer, 1999. 88

[78] D Meyer, E Dimitriadou, K Hornik, A Weingessel, F Leisch, CC Chang, and CC Lin. Misc functions of the department of statistics, probability theory group (formerly: E1071). *Package e1071. TU Wien*, 2015.

[79] Shervin Minaee, Amirali Abdolrashidi, Hang Su, Mohammed Bennamoun, and David Zhang. Biometric recognition using deep learning: A survey. *arXiv preprint arXiv:1912.00271*, 2019.

[80] MD Moniruzzaman and Syed Islam. Automatic ear detection using deep learning. 2017.

[81] Colm Mulcahy. Image compression using the haar wavelet transform. *Spelman Science and Mathematics Journal*, 1(1):22–31, 1997.

[82] Mohammed Hasan Mutar, Essam Hammodi Ahmed, and HO Majid Razaq Mohamed ALsemawi. Ear recognition system using random forest and histograms of oriented gradients techniques. *Solid State Technology*, 63(4):8740–8748, 2020.

[83] Loris Nanni and Alessandra Lumini. A multi-matcher for ear authentication. *Pattern Recognition Letters*, 28(16):2219–2226, 2007.

[84] Imran Naseem, Roberto Togneri, and Mohammed Bennamoun. Sparse representation for ear biometrics. In *International Symposium on Visual Computing*, pages 336–345. Springer, 2008.

[85] Mustapha Oloko-Oba and Serestina Viriri. Ensemble of efficientnets for the diagnosis of tuberculosis. *Computational Intelligence and Neuroscience*, 2021, 2021. 102

[86] Ibrahim Omara, Xiaohe Wu, Hongzhi Zhang, Yong Du, and Wangmeng Zuo. Learning pairwise svm on deep features for ear recognition. In *Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on*, pages 341–346. IEEE, 2017.

[87] Anika Pflug and Christoph Busch. Ear biometrics: a survey of detection, feature extraction and recognition methods. *IET biometrics*, 1(2):114–129, 2012. 5, 6

[88] P.Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998. 88

[89] Surya Prakash and Phalguni Gupta. An efficient ear localization technique. *Image and Vision Computing*, 30(1):38–50, 2012.

[90] Surya Prakash and Phalguni Gupta. An efficient ear recognition technique invariant to illumination and pose. *Telecommunication Systems*, 52(3):1435–1448, 2013.

[91] Surya Prakash, Umarani Jayaraman, and Phalguni Gupta. Connected component based technique for automatic ear detection. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 2741–2744. IEEE, 2009. 88

[92] Ramar Ahila Priyadharshini, Selvaraj Arivazhagan, and Madakannu Arun. A deep learning approach for person identification using ear biometrics. *Applied Intelligence*, pages 1–12, 2020.

[93] K Radhika, K Devika, T Aswathi, P Sreevidya, V Sowmya, and KP Soman. Performance analysis of nasnet on unconstrained ear recognition. In *Nature Inspired Computing for Data Science*, pages 57–82. Springer, 2020.

[94] Mahbubur Rahman, Md Rashedul Islam, Nazmul Islam Bhuiyan, Bulbul Ahmed, and Md Aminul Islam. Person identification using ear biometrics. *International Journal of The Computer, the Internet and Management*, 15(2):1–8, 2007.

[95] Rui Raposo, Edmundo Hoyle, Adolfo Peixinho, and Hugo ProenĂ§a. Ubear: A dataset of ear images captured on-the-move in uncontrolled conditions. 04 2011. 88

[96] William Raveane, Pedro Luis Galdamez, and Maria Angelica Gonzalez Arrieta. Ear detection and localization with convolutional neural networks in natural images and videos. *Processes*, 7(7):457, 2019.

[97] KR Resmi and G Raju. Automatic 2d ear detection: A survey.

[98] Arun Ross and Ayman Abaza. Human ear recognition. *Computer*, 44(11):79–81, 2011.

[99] Curtis T Rueden, Johannes Schindelin, Mark C Hiner, Barry E DeZonia, Alison E Walter, Ellen T Arena, and Kevin W Eliceiri. Imagej2: Imagej for the next generation of scientific image data. *BMC bioinformatics*, 18(1):1–26, 2017.

[100] S Hma Salah, H Du, and N Al-Jawad. Fusing local binary patterns with wavelet features for ethnicity identification. In *Proc. IEEE Int. Conf. Signal Image Process*, volume 21, pages 416–422, 2013.

[101] Sudeep Sarkar, Mauricio Pamplona Segundo, and Earnest Eugene Hansley. Unconstrained ear recognition using a combination of deep learning and handcrafted features, September 24 2019. US Patent 10,423,823.

[102] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[103] Dasari Shailaja and Phalguni Gupta. A simple geometric approach for ear recognition. In *9th International Conference on Information Technology (ICIT'06)*, pages 164–167. IEEE, 2006.

[104] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database of human faces. Technical Report CMU-RI-TR-01-02, Carnegie Mellon University, Pittsburgh, PA, January 2001. 88

[105] Terence Sim, Simon Baker, and Maan Bsat. *The CMU pose, illumination, and expression (PIE) database of human faces.* Citeseer, 2001.

[106] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[107] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[108] Michael Reed Teague. Image analysis via the general theory of moments*. *JOSA*, 70(8):920–930, 1980.

[109] Terry M Therneau, Elizabeth J Atkinson, et al. An introduction to recursive partitioning using the rpart routines. Technical report, Technical report Mayo Foundation, 1997.

[110] Liang Tian and Zhichun Mu. Ear recognition based on deep convolutional network. In *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 437–441. IEEE, 2016.

[111] Arkadiusz Tomczyk and Piotr S Szczepaniak. Ear detection using convolutional neural network on graphs with filter rotation. *Sensors*, 19(24):5510, 2019.

[112] Vallabh Vidyanagar. Oval shape detection and support vector machine based approach for human ear detection from 2d profile face image. *International Journal of Hybrid Information Technology*, 7(5):113–120, 2014.

[113] N.-S. Vu, H. M. Dee, and A. Caplier. Face recognition using the POEM descriptor. 45(7):2478–2488, 2012.

[114] Ngoc-Son Vu, Hannah M Dee, and Alice Caplier. Face recognition using the poem descriptor. *Pattern Recognition*, 45(7):2478–2488, 2012.

[115] Zhi-qin Wang and Xiao-dong Yan. Multi-scale feature extraction algorithm of ear image. In *2011 International Conference on Electric Information and Control Engineering*, pages 528–531. IEEE, 2011.

[116] Yu Weng, Tianbao Zhou, Yujie Li, and Xiaoyu Qiu. Nas-unet: Neural architecture search for medical image segmentation. *IEEE Access*, 7:44247–44257, 2019.

[117] Ping Yan and Kevin W Bowyer. *Ear biometrics in human identification*. University of Notre Dame PhD in computer science and engineering, 2006.

[118] Ping Yan and Kevin W Bowyer. Biometric recognition using 3d ear shape. *IEEE Transactions on pattern analysis and machine intelligence*, 29(8):1297–1308, 2007.

[119] Ping Yan and KevinW Bowyer. Empirical evaluation of advanced ear biometrics. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, pages 41–41. IEEE, 2005. 88

[120] Takaya Yuizono, Yu Wang, Kiminori Satoh, and Shigeru Nakayama. Study on individual recognition for ear images by using genetic local search. In *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600)*, volume 1, pages 237–242. IEEE, 2002.

[121] Yi Zhang, Zhi-Chun Mu, Li Yuan, Chen Yu, and Liu Qing. *USTB-Helloear: A Large Database of Ear Images Photographed Under Uncontrolled Conditions*, pages 405–416. 12 2017. 87

[122] Yi Zhang and Zhichun Mu. Ear detection under uncontrolled conditions with multiple scale faster region-based convolutional neural networks. *Symmetry*, 9(4):53, 2017. 5, 6

[123] Yi Zhang, Zhichun Mu, Li Yuan, and Chen Yu. Ear verification under uncontrolled conditions with convolutional neural networks. *IET Biometrics*, 7(3):185–198, 2018. 98

[124] Yi Zhang, Zhichun Mu, Li Yuan, Chen Yu, and Qing Liu. Ustb-helloear: A large database of ear images photographed under uncontrolled conditions. In *International Conference on Image and Graphics*, pages 405–416. Springer, 2017.

[125] Sanping Zhou, Fei Wang, Zeyi Huang, and Jinjun Wang. Discriminative feature learning with consistent attention regularization for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8040–8049, 2019.

[126] Yuxiang Zhou and Stefanos Zaferiou. Deformable models of ears in-the-wild for alignment and recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 626–633. IEEE, 2017.