

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA  
Ministry Of Higher Education and Scientific Research  
UNIVERSITÉ 8 MAI 1945 - GUELMA



Faculty of Sciences and Technology  
Department of Electronics and Telecommunications

LABORATOIRE PROBLEMES INVERSES, MODELISATION, INFORMATION ET SYSTEMES  
(PIMIS)

**Doctoral Dissertation**

Submitted in partial fulfilment of the requirements for  
Doctorate degree 3<sup>rd</sup> Cycle LMD in Department of Electronics and Telecommunications

**Domain:** Sciences and Technology **Sector:** Electronics

**Specialty:** Security and biometrics

Presented by: **Rafik BOUAOUINA**

Entitled

**Biometric Recognition using Deep Learning**

Defended on: October 10<sup>th</sup> 2024

Examination Committee :

President	Houcine BOUROUBA	Professor	University of Guelma
Director	Amir BENZAOU	Professor	University of Skikda
Co-Director	Hakim DOGHMANE	MCA	University of Guelma
Examiner	Mohamed NEMISSI	Professor	University of Guelma
Examiner	Kamel MESSAOUDI	Professor	University of Souk-Ahrass

2023/2024

## *Dedications*

*I dedicate this work to my parents*

*May they find here the testimony of my deep gratitude and  
acknowledgment*

*To my wife and my children who give love and liveliness.*

*To all those who have helped me - directly or indirectly - and those  
who shared with me the emotional moments during the  
accomplishment of this work and who warmly supported and  
encouraged throughout my journey.*

*To all my friends who have always encouraged me, and to whom I  
wish more success.*

*Thanks!*

*Rafik BOUAOUINA*



# Abstract

In forensic science, security, identity, and verification, biometrics are essential. Over the past three decades, research has concentrated on developing dependable systems with human behavioural and biological characteristics, each with its own advantages and difficulties.

The human ear is emerging as a promising biometric modality due to its unique characteristics and advantages over other methods like face, fingerprint, or iris scanning. Ear biometrics offers a seamless process without requiring physical movements from users. Recent years have seen increased attention and progress in ear biometrics, covering detection, preparation, feature extraction, verification, and identification. Taking advantage of this great development in deep learning due to improvements in computer processing power, data availability, and innovative algorithms. More specifically, convolutional neural networks (CNNs), which have achieved remarkable success in areas such as computer vision tasks.

We conducted a comprehensive experimental study introducing novel methodologies to improve ear identification. Two main contributions are devoted in this work; the first one introduces a practical approach called Mean-Class Activation Maps with CNNs (Mean-CAM-CNN) to tackle issues related to image classification by focusing on discriminative regions of ear images. The Mean-CAM framework addresses intra-class variability by using a guided mask to crop relevant areas based on mean heat maps. This cropped region is then utilised for training a CNN, therefore enhancing discriminative classification performance. In the second contribution, we aimed to combine deep convolutional generative adversarial network DCGAN model with Mean-CAM technique, where DCGAN model used as pre-processing step to colourise and enhance the ear images of the used datasets. The proposed approach was evaluated on two ear recognition datasets, AMI and AWE, showing significant improvements with Rank-1 recognition rates of 100 % for AMI and 76.25 % for AWE.

**Key words:** *biometrics, ear recognition, generative adversarial networks, convolutional neural networks, class activation map, attention networks.*

## ملخص

في مجال الطب الشرعي والأمن والهوية والتحقق، تعتبر القياسات الحيوية ضرورية. على مدى العقود الثلاثة الماضية، ركز البحث على تطوير أنظمة يمكن الاعتماد عليها ذات خصائص سلوكية وبيولوجية بشرية، ولكل منها مزاياها وصعوباتها الخاصة.

تبرز الأذن البشرية كطريقة قياس حيوية واعدة بسبب خصائصها الفريدة ومزاياها مقارنة بالطرق الأخرى مثل مسح الوجه أو بصمات الأصابع أو قرحة العين. تقدم القياسات الحيوية للأذن عملية سلسلة دون الحاجة إلى حركات جسدية من المستخدمين. شهدت السنوات الأخيرة زيادة الاهتمام والتقدم في القياسات الحيوية للأذن، والتي تغطي الكشف والإعداد واستخراج السمات والتحقق والتعريف. الاستفادة من هذا التطور الكبير في التعلم العميق بسبب التحسينات في قوة معالجة الكمبيوتر وتوافر البيانات والخوارزميات المبتكرة. وبشكل أكثر تحديدًا، الشبكات العصبية التلافيفية (CNNs)، والتي حققت نجاحًا ملحوظًا في مجالات مثل مهام الرؤية الحاسوبية.

أجرينا دراسة تجريبية شاملة قدمت منهجيات جديدة لتحسين التعرف على الأذن. تم تخصيص مساهمتين رئيسيتين في هذا العمل؛ يقدم الأول نهجًا عمليًا يسمى خرائط تنشيط الفئة المتوسطة مع الشبكات العصبية التلافيفية -Mean-CAM) (CNN لمعالجة القضايا المتعلقة بتصنيف الصور من خلال التركيز على المناطق التمييزية لصور الأذن. يعالج إطار عمل Mean-CAM التباين داخل الفئة باستخدام قناع موجه لقص المناطق ذات الصلة بناء على خرائط الحرارة المتوسطة. ثم يتم استخدام هذه المنطقة المقصودة لتدريب CNN، وبالتالي تعزيز أداء التصنيف التمييزي. في المساهمة الثانية، كنا نهدف إلى الجمع بين نموذج DCGAN للشبكة التنافسية التوليدية التلافيفية العميقة وتقنية Mean-CAM، حيث تم استخدام نموذج DCGAN كخطوة معالجة مسبقة لتلوين وتحسين صور الأذن لمجموعات البيانات المستخدمة. تم تقييم النهج المقترح على مجموعتي بيانات التعرف على الأذن، AMI و AWE، مما أظهر تحسينات كبيرة مع معدلات التعرف على المرتبة الأولى بنسبة 100% لـ AMI و 76.25% لـ AWE.

**كلمات مفتاحية:** القياسات الحيوية، التعرف على الأذن، الشبكات التوليدية التنافسية، الشبكات العصبية التلافيفية، خريطة تنشيط الفئة.

# Résumé

En sciences forensiques, en sécurité, en identité et en vérification, la biométrie est essentielle. Au cours des trois dernières décennies, la recherche s’est concentrée sur le développement de systèmes fiables avec des caractéristiques comportementales et biologiques humaines, chacune avec ses propres avantages et difficultés.

L’oreille humaine émerge comme une modalité biométrique prometteuse en raison de ses caractéristiques uniques et de ses avantages par rapport à d’autres méthodes comme la numérisation du visage, des empreintes digitales ou de l’iris. La biométrie auriculaire offre un processus transparent sans nécessiter de mouvements physiques de la part des utilisateurs. Ces dernières années, une attention et des progrès accrus ont été réalisés dans le domaine de la biométrie auriculaire, couvrant la détection, la préparation, l’extraction de caractéristiques, la vérification et l’identification. Profitant de ce grand développement dans l’apprentissage profond en raison des améliorations de la puissance de traitement informatique, de la disponibilité des données et des algorithmes innovants. Plus précisément, les réseaux neuronaux convolutionnels (CNN), qui ont obtenu un succès remarquable dans des domaines tels que les tâches de vision par ordinateur.

Nous avons mené une étude expérimentale complète introduisant de nouvelles méthodologies pour améliorer l’identification des oreilles. Deux contributions principales sont consacrées à ce travail ; La première présente une approche pratique appelée Mean-Class Activation Maps with CNNs (Mean-CAM-CNN) pour résoudre les problèmes liés à la classification d’images en se concentrant sur les régions discriminantes des images d’oreille. Le cadre Mean-CAM aborde la variabilité intra-classe en utilisant un masque guidé pour recadrer les zones pertinentes en fonction des cartes de chaleur moyennes. Cette région recadrée est ensuite utilisée pour entraîner un CNN, améliorant ainsi les performances de classification discriminante. Dans la deuxième contribution, nous avons cherché à combiner le modèle DCGAN de réseau antagoniste génératif convolutionnel profond avec la technique Mean-CAM, où le modèle DCGAN est utilisé comme étape de prétraitement pour colorer et améliorer les images d’oreille des ensembles de données utilisés. L’approche proposée a été évaluée sur deux ensembles de données de reconnaissance d’oreille, AMI et AWE, montrant des améliorations significatives avec des taux de reconnaissance de rang 1 de 100,00 % pour AMI et 76,25 % pour AWE.

**Mots clés :** *Biométrie, Reconnaissance de l’oreille, confrontation générative réseaux, réseaux de neurones convolutifs, carte d’activation de classe, réseaux d’attention.*

# Contents

List of Figures

List of Tables

List of Abbreviations

<b>General introduction</b>	<b>1</b>
<b>1 Fundamentals of Biometrics</b>	<b>4</b>
1.1 Introduction . . . . .	5
1.2 Definition of biometrics . . . . .	5
1.3 Identification vs. Verification . . . . .	6
1.4 Biometric systems . . . . .	6
1.5 Applications of biometric systems . . . . .	9
1.6 Performance evaluation . . . . .	11
1.6.1 Verification system error rates . . . . .	12
1.6.2 Identification system error rates . . . . .	14
1.7 Biometric modalities . . . . .	15
1.7.1 Fingerprint . . . . .	16
1.7.2 Face . . . . .	17
1.7.3 Iris . . . . .	18
1.7.4 Ear . . . . .	19
1.7.5 Hand geometry . . . . .	19
1.7.6 Gait . . . . .	20
1.7.7 Signature . . . . .	21
1.7.8 Retina . . . . .	21
1.8 Comparison between biometric modalities . . . . .	22
1.9 Why this thesis focuses on ear biometrics? . . . . .	24

1.10	Conclusion	24
<b>2</b>	<b>Ear Biometrics</b>	<b>26</b>
2.1	Introduction	27
2.2	Motivation: Why ear biometric?	27
2.3	Characteristics of the ear	28
2.4	Ear recognition system	29
2.4.1	Ear detection	29
2.4.2	Pre-processing or normalisation	30
2.4.3	Feature extraction	30
2.4.4	Classification or identification	30
2.5	Ear recognition methods	31
2.5.1	Handcrafted methods	31
2.5.2	Deep learning methods	32
2.6	Open challenges	36
2.7	Conclusion	37
<b>3</b>	<b>Review of Deep Learning</b>	<b>39</b>
3.1	Introduction	40
3.2	Artificial intelligence, machine learning and deep learning	40
3.3	Artificial neural network	42
3.3.1	Architecture of an artificial neural network	43
3.3.2	Activation functions	44
3.3.3	Model training	48
3.4	Convolutional neural networks (CNNs)	49
3.4.1	Convolutional layers	50
3.5	CNN architectures	53
3.5.1	AlexNet	53
3.5.2	Visual geometry group (VGG)	54
3.5.3	GoogleNet	54
3.5.4	ResNet	55
3.5.5	Inception	56
3.5.6	DenseNet	57
3.5.7	ResNeXt	58
3.5.8	Wide ResNet	59

3.5.9	Recent advancements in convolutional neural networks . . . . .	59
3.6	Deep learning applications . . . . .	60
3.7	Challenges in deep learning . . . . .	61
3.8	Conclusion . . . . .	62
<b>4</b>	<b>Ear Recognition using Mean-Class activation Maps and Convolutional Neural Networks</b>	<b>64</b>
4.1	Introduction . . . . .	65
4.2	Materials and methods . . . . .	65
4.2.1	ResNet-50 . . . . .	65
4.2.2	Class activation map (CAM) . . . . .	66
4.2.3	Data augmentation . . . . .	69
4.2.4	Proposed approach . . . . .	70
4.3	Experimental analysis . . . . .	76
4.3.1	Data sets . . . . .	76
4.3.2	Experimental protocols and setups . . . . .	77
4.3.3	Ablation analysis . . . . .	77
4.3.4	Comparison . . . . .	84
4.4	Conclusion . . . . .	85
<b>5</b>	<b>Boosting the Performance of Deep Ear Recognition Systems Using Generative Adversarial Networks and Mean Class Activation Maps</b>	<b>87</b>
5.1	Introduction . . . . .	88
5.2	Proposed approach . . . . .	88
5.2.1	Preprocessing . . . . .	88
5.2.2	Feature extraction/classification . . . . .	91
5.3	Experimental analysis . . . . .	92
5.3.1	Evaluation protocols and setup . . . . .	93
5.3.2	Experiment 1 . . . . .	93
5.3.3	Experiment 2 . . . . .	94
5.3.4	Experiment 3 . . . . .	98
5.4	Comparison . . . . .	98
5.5	Conclusion . . . . .	100
	<b>Bibliography</b>	<b>105</b>

# List of Figures

1.1	Representation of the structure of a biometric system. . . . .	7
1.2	Distribution of error rates in a biometric verification system. . . . .	13
1.3	Illustration of ROC (left) and DET (right) curves. . . . .	14
1.4	CMC curve representation. . . . .	15
1.5	Fingerprint patterns. . . . .	17
1.6	Face recognition. . . . .	18
1.7	Iris sample. . . . .	18
1.8	Anatomy of the ear. . . . .	19
1.9	An example of hand geometry reader. . . . .	20
1.10	Gait cycle. . . . .	20
1.11	Online and off-line signatures. . . . .	21
1.12	A retinal image. . . . .	22
2.1	Anatomical structure of the outer ear shape. . . . .	29
2.2	Flowchart of an ear biometric recognition system. . . . .	29
2.3	An example of ear detection. . . . .	30
2.4	Example of ear normalisation [32]. . . . .	30
3.1	Machine learning vs. deep learning feature extraction. . . . .	42
3.2	The structure of a perceptron. . . . .	42
3.3	Architecture of an ANN. . . . .	43
3.4	A binary activation function. . . . .	44
3.5	A linear activation function. . . . .	45
3.6	Sigmoid activation function curve. . . . .	46
3.7	Tanh activation function curve. . . . .	46
3.8	Illustration of ReLU and Leaky ReLU activation functions. . . . .	47
3.9	A simple CNN architecture. . . . .	50

3.10	A graphical example of the convolution process. . . . .	51
3.11	Pooling processes. . . . .	52
3.12	AlexNet architecture. . . . .	54
3.13	VGG architecture. . . . .	54
3.14	Inception module of GoogleNet. . . . .	55
3.15	A residual block. . . . .	56
3.16	A basic schematic for the Inception Residual unit. . . . .	57
3.17	A schematic illustration of a 3-layer dense block used in the DenseNet. . . . .	58
3.18	ResNeXt building block. . . . .	59
3.19	Examples of DL applications. . . . .	61
4.1	CAMs of several images of different classes from AWE and AMI data sets. The maps indicate the relevant image areas employed for image classification. . . . .	67
4.2	Class Activation Mapping: The anticipated class score is projected back to the preceding convolutional layer (CAM) to construct the class activation maps. CAM emphasises the relevant areas that are distinctive to each class. . . . .	69
4.3	A graphical flowchart of the proposed Mean-CAM-CNN framework. . . . .	71
4.4	A graphical flowchart of the Mean-CAM process. . . . .	73
4.5	The process of generating discriminative regions for the AMI dataset. . . . .	74
4.6	The process of generating discriminative regions for the AWE dataset. . . . .	75
4.7	CMC curves produced on the testing set from the AMI database using the feature extraction strategy: a) Without data augmentation. b) With data augmentation. . . . .	79
4.8	CMC curves produced on the testing set from the AWE database using the feature extraction strategy: a) Without data augmentation. b) With data augmentation. . . . .	80
4.9	CMC curves produced on the testing set from the AMI database using the fine-tuning strategy: a) Without data augmentation. b) With data augmentation. . . . .	82
4.10	CMC curves produced on the testing set from the AWE database using the fine-tuning strategy: a) Without data augmentation. b) With data augmentation. . . . .	83
5.1	U-Net design for the generative paradigm. . . . .	89
5.2	Design of the discriminative model. . . . .	89
5.3	The proposed DCGAN model is used to colour, enhance, and resize ear images.: (a) original images; (b) enhanced images. . . . .	90



5.4	An illustration of the suggested mean-CAM-CNN framework's flowchart. . .	92
5.5	CMCs for the testing sets of (a) AMI and (b) AWE datasets: comparative analysis of ResNet-50 and Mean-CAM-CNN with varied threshold values ( $\tau$ ).	96

# List of Tables

1.1	Most of used biometric modalities. . . . .	16
1.2	Comparison between different biometric modalities. . . . .	23
2.1	Handcrafted approaches comparative summary. . . . .	34
2.2	Deep learning approaches comparative summary. . . . .	35
4.1	ResNet-50 structural parameters. . . . .	66
4.2	Rank-1, Rank-5, and AUC results (%) using the feature extraction strategy for the AMI data set. . . . .	78
4.3	Rank-1, Rank-5, and AUC results (%) using the feature extraction strategy for the AWE data set. . . . .	79
4.4	Rank-1, Rank-5, and AUC results (%) using the fine-tuning strategy for the AMI data set. . . . .	81
4.5	Rank-1, Rank-5, and AUC results (%) using the fine-tuning strategy for the AWE data set. . . . .	82
4.6	Comparing Mean-CAM-CNN Rank-1 recognition rate with several competing approaches using AMI and AWE ear recognition data sets. . . . .	85
5.1	Comparative evaluation of fine-tuned CNN architectures for ear recognition. . . . .	94
5.2	Ear recognition results using ResNet-50 and Mean-CAM-CNN across different threshold values of the $\tau$ parameter. . . . .	95
5.3	Visualisation and performance analysis of model predictions: a comparative study using the Mean-CAM technique on ear images (B/L pred denotes baseline prediction, and P represents probability). . . . .	97
5.4	Comparative analysis of ear recognition outcomes: Evaluating the influence of preprocessing on Mean-CAM-CNN performance. . . . .	98

5.5 Comparative analysis of Rank-1 recognition rates: evaluating the DCGAN + Mean-CAMCNN approach against competing methods on AMI and AWE datasets. . . . .	100
--	-----

# List of Abbreviations

**DTCW:** Dual tree complex wavelet

**1D-LBP:** Glocal binary pattern

**CNN:** Convolutional neural network

**VGG:** Visual geometry group

**DCGAN:** Deep convolutional generative adversarial network

**SVM:** Support vector machine

**CFDCNet:** Channel features and dynamic convolution

**MDFNet:** Magnitude and direction alongside responses of data-driven filters

**NLP:** Natural language processing

**AI:** Artificial intelligence

**ML:** Machine learning

**ANN:** Artificial neuronal network

**AI:** Artificial intelligence

**ML:** Machine learning

**ANN:** An artificial neuronal network

**MLP:** multi-layer perceptron

**ReLU:** Rectified linear unit function

**FC:** Fully connected

**DL:** Deep Learning

**ZB:** zettabytes

**CAM:** Class Activation Map

**GAP:** Global average pooling

**GMP:** global max-pooling

**AUC:** Area under the curve

**MAI:** The mathematical analysis of images

**AWE:** Annotated web ears

**DMC:** dynamic matrix control

**CNN:** convolution neural network

**Mean-CAM:** Mean-Class activation Map

**DCGAN:** Deep convolutional generative adversarial network

# General introduction

## **Context**

Biometrics is the process of automatically identifying a person by analysing their physiological or behavioural features. This identification method is superior to conventional methods that rely on passwords and personal identification numbers (PINs) for several reasons. Firstly, it requires the person to be physically present at the point of identification. Additionally, it utilises biometric techniques, eliminating the need to remember a password or carry a token. Given the growing reliance on computers for information technology, it is imperative to impose limitations on accessing sensitive and personal data. By substituting PINs, biometric approaches have the ability to prevent illegal access to or deceitful use of daily human technologies, including cellular phones, smart cards, workstations, and computer networks. PINs and passwords have the potential to be forgotten, whereas token-based identity systems like passports and driver's licenses are susceptible to forgery, theft, or loss. Biometric identification systems are currently seeing renewed interest.

Various human traits have been explored and evaluated. These modalities can be further divided into sub-categories based on their position in the human body. These sub-categories include attributes of the hand region (such as fingerprint and hand geometry), attributes of the facial region (such as face and ear), attributes of the ocular region (such as iris and retina), behavioural attributes (such as walking style and electronic signature), and medical-chemical attributes (such as bones, odor, and DNA).

Ear recognition has gained significant interest in recent years as a biometric recognition system. This thesis focuses on enhancing accuracy, creating resilient recognition algorithms, and investigating novel uses in the fields of security, access control, and forensic identification. This is accomplished by using the advancements achieved in the domain of deep learning and computer vision.

Ear biometric systems encounter a number of challenges and difficulties, some of which are comparable to the issues seen in other modalities. These difficulties persist despite the extensive study undertaken on ear biometrics. For instance, some factors that might affect the quality of an image are lighting, pose, occlusions, and image resolution. Given these constraints, ear datasets may be classified into two distinct categories: constrained and unconstrained. In our research, we specifically targeted unconstrained ear datasets due to their significant difficulty to the scientific community.

## **Research objectives**

The objectives of this research are:

- Algorithms based on Deep Learning, in particular convolution networks, have quickly

become a methodology of choice for image analysis. The objective of this research work is to develop and propose effective approaches in feature extraction and classification based on the Deep Learning technique in order to model, classify, and recognise individuals by biometric characteristics under unconstrained conditions. In other words, the aim is to improve the performance of biometric identification under uncontrolled conditions.

- Proposing a practical approach called mean-class activation maps with CNNs, which extracts and considers only the discriminative regions of the entire image.
- Presenting a practical approach leveraging the combination of DCGAN and Mean-CAM-CNN for addressing the intricate challenge of ear biometric recognition.

### **Thesis overview**

The thesis is structured as follows:

Chapter 1 presents a comprehensive introduction to biometrics, including an explanation of the biometric recognition systems and an analysis of the factors that impact the performance of a biometric system.

Chapter 2 explores the ear recognition system. It examines relevant studies in the subject and classifies them into two primary categories: handcrafted methods and deep learning methods.

Chapter 3 focuses on the topic of deep learning and its practical uses. It provides a comprehensive overview of machine learning and its various branches, including artificial neural networks and convolutional neural networks, along with their distinct structures.

Chapter 4 introduces a new approach to biometric ear recognition using mean class activation maps and convolutional neural networks to address the problems of irrelevant noise, blur, or occluded regions that can affect the CNN features derived from a global image.

In chapter 5, the issue of greyscale ear images is addressed, along with the method of colourising them to improve the model's performance. Subsequently, a practical feature extraction and classification technique, referred to as Mean-CAM-CNN, is introduced. This technique leverages mean-class activation maps in conjunction with CNNs.



# Chapter 1

## Fundamentals of Biometrics

## 1.1 Introduction

In this technological era, biometrics has become crucial across various domains including security, personal identification, verification systems, and forensic science. Over the past three decades, extensive research has been conducted to develop robust biometric systems utilising fingerprints, voice, iris, face, gait, keystroke, and ear recognition, each with its own set of benefits and limitations.

The structure of this chapter is as follows: Section 2 provides a clear definition and context for the concept of biometrics. Section 3 specifically addresses the issues related to biometric identification and verification, along with the performance assessment criteria that are pertinent to these challenges. The fourth section of the document discusses the general structure of the biometric recognition system. Section 5 focuses on the prominent biometric applications and the latest advancements in the field. Section 6 provides an overview of the assessment techniques that are subsequently discussed in the seventh section, focusing specifically on biometric modalities. Section 8 covers the process of choosing a biometric modality for a specific application. Section 9 elucidates the reasons on why this thesis has focused on ear biometrics. The last portion serves as the concluding part of this chapter.

## 1.2 Definition of biometrics

The term "biometrics" originates from the Greek words "bios" (life) and "metron" (measure), confirming or identifying a person's identity through various body parts such as a fingerprint or face pattern, and behavioural characteristics, such as handwriting or walking patterns [1]. The International Organization for Standardization (ISO) defines biometrics as "the automated recognition of individuals based on their biological and behavioural characteristics" (ISO/IEC 2382-37, 2012).

Whenever the term biometrics is used, it refers to the ability to measure the human or biological existence. This technology is fundamentally about measuring human existence to verify identity, ensuring that these characteristics are permanently linked to the individual.

Biometrics has transformed the field of identification, offering a reliable method for detecting subtle differences quickly. It provides enhanced security, safety, authentication, and precision. Initially reliant on passcodes and PINs, security measures have evolved to utilise the unique physiological traits of individuals [2], reflecting the growing demand for robust protection in all aspects of life.

## 1.3 Identification vs. Verification

Biometrics is a pattern recognition system that uses human biological and behavioural characteristics for authentication [3]. The system identifies an individual by matching similar characteristics. There are two essential modes of human authentication in biometric systems: identification and verification.

1. Identification in biometric systems is the process of determining a person's identity by comparing their biometric features against multiple records in a database, a method known as "one-to-many" matching. This process requires comparing the biometric data presented with numerous stored templates to find a match. Due to the need to search through extensive databases, identification is generally more computationally intensive and slower than verification. Identification is used widely in scenarios where establishing an individual's identity is necessary without any prior identity claim, such as in criminal investigations, surveillance systems, and access control settings where identity cards are impractical.
2. Verification in biometric systems, also known as "one-to-one" matching, involves confirming an individual's claimed identity by comparing their presented biometric data against a previously enrolled template linked to that identity. This process is essentially a check of the assertion: "Is this person who they claim to be?" Verification is commonly used for personal authentication tasks, such as logging into a smartphone, accessing a bank account, or entering a secure facility. Typically, the user claims an identity (e.g., by entering a username) and then provides a biometric sample to support that claim. Although generally faster and less resource-intensive than identification, verification systems must maintain high accuracy in matching to prevent false rejections (denying access to the legitimate user) and false acceptances (granting access to an impostor) [4].

## 1.4 Biometric systems

A biometric system recognises specific characteristics of an individual through the use of mathematical algorithms and biometric data. These systems have various applications, some of which require users to enrol in advance, while others do not necessitate this preliminary step [5]. A good biometric system includes four components that complete the system architecture [6]. Figure 1.1 shows the structure of a biometric system [7].

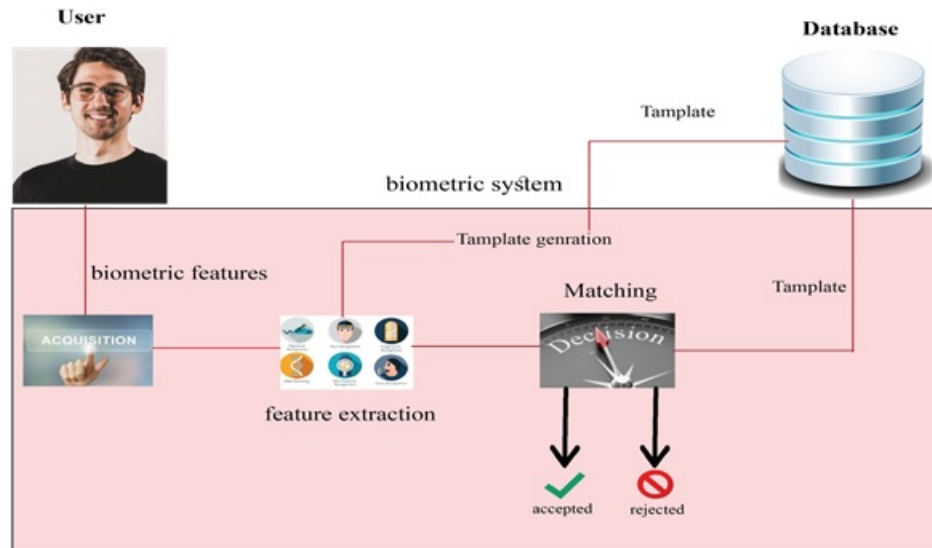


Figure 1.1: Representation of the structure of a biometric system.

Biometric architecture can perform all reliable and impenetrable verification methods to confirm the identity of a person.

1. **Data collection:** An essential stage in the functioning of biometric systems, when specialised technologies are used to gather physical or behavioural samples. This stage has substantial implications for the efficiency and reliability of the system. The importance of the obtained data lies on its quality. If the quality of the sample falls below a specified threshold due to variables such as poor positioning, illumination, or other environmental conditions, the system might request for an alternative sample. To maintain recognition and identification accuracy, it is important to ensure that only data of high quality is inputted into the system.
2. **Feature extraction process:** Digital signal processing (DSP) techniques play a significant role in the feature extraction process within biometric systems, where specific characteristics are identified and extracted from the raw biometric samples collected during the data collection phase. This process converts the raw data into a more manageable set of data points, known as features, which uniquely identify an individual.
3. **Matching:** In a biometric system, matching refers to the process of comparing a newly captured biometric sample with the stored templates in the system's database to determine the identity of an individual. Various distance metrics, similarity measures, or matching algorithms are used to quantify the similarity or dissimilarity between the

feature vectors. The choice of the matching algorithm and distance metric depends on the nature of the biometric modality and the feature representation used.

4. **Decision-making process:** It involves determining whether the newly collected sample's extracted features correspond to a match or a mismatch when compared to the stored templates.

To ensure accurate and reliable identification and verification in a biometric system, particular characteristics are essential for the effectiveness and robustness of the biometric system [8]. The following are the fundamental characteristics of an excellent biometric modality:

- **Universality:** It refers to the presence of the biometric trait in every individual. An excellent biometric modality should be universally present across the target population, ensuring that it can be used for widespread identification and verification purposes.
- **Distinctiveness:** A high-quality biometric modality should possess distinct and unique characteristics that enable it to differentiate individuals effectively. The feature should capture individual variations within the biometric trait, ensuring that each person's biometric data is distinguishable from others.
- **Permanence:** The quality of permanence implies the stability and consistency of the biometric trait over time. A reliable biometric feature should exhibit minimal changes over an individual's lifetime, thus ensuring the steadiness of the biometric data and its applicability for long-term identification.
- **Collectability:** The biometric feature should be easily collectible. It should be feasible to capture the biometric data in various operational environments without causing discomfort or inconvenience to the individuals.
- **Performance:** An excellent biometric feature should demonstrate high performance in terms of accuracy, robustness, and resistance to spoofing attacks.
- **Acceptance:** The perceived usefulness and ease of use of biometric systems play a significant role in individuals' acceptance.
- **Circumvention:** In the context of biometric systems, circumvention refers to the ability of individuals to deceive or bypass the security measures of the system, thereby compromising its effectiveness.

An effective biometric system must possess satisfactory accuracy and adequate recognition speed in relation to the necessary resources, be non-hazardous to users, widely accepted by the public, and sufficiently resilient against fraudulent techniques. Diverse biometric modalities are used in various applications [9]. Every biometric modality has distinct advantages and disadvantages, and the selection often relies on the specific application at hand. None of the biometric modalities fully satisfy the demands of all applications.

## 1.5 Applications of biometric systems

The requirement has covered biometric technology for a large-scale identity management system whose functionality depends on accurately determining a person's identity in the context of numerous applications [2, 3]. A biometric system protects the system to which it is applied. Almost every component of life is becoming a computer-based system, and computers need security more than anything else. Computers and every area of life want reliability and protection for their existence; therefore, they rely entirely on the biometric system. Biometric systems have multiple and multidimensional applications in today's era. Each biometric modality has its advantages, limitations, and applications [9]. The important applications of biometric systems can be summarised as follows:

- **Legal applications:** The field of justice and law enforcement has a long-standing history with biometric technology, leading to significant advancements in identity management. Currently, the police department utilises a genuinely multimodal approach to biometrics. Fingerprint, facial, and vocal recognition technologies serve a distinct purpose in enhancing public safety and facilitating the identification of individuals we are seeking.
- **Government applications:**
  - Border Control and Airports: One of the most important application areas of biometric technology is at every border. Biometric technology aids in automating the border-crossing process. Reliable initiatives of automated passenger screening are further making international passenger travel easier and border controls better, compared to ever before.
  - Healthcare: In the healthcare sector, biometrics is used to introduce an improved model. Medical records are one of the most important records in one's life; doctors need to get them on time, and they need to be accurate. Poor security and bad

accounting can mean the difference between an on-time accurate diagnosis and health fraud.

- **Commercial applications:**

- Security: As global connectedness grows, existing security approaches are no longer adequate to secure what matters. Fortunately, biometric technology is more accessible than ever, ready to secure and simplify everything from vehicle doors to phone PINs.
- Finance: Biometric technology is widely used in financial identification, verification, and authentication to enhance the safety, convenience, and accountability of banking, purchasing, and account management. Biometric solutions in the financial sector verify a customer's identity by capturing and comparing their distinct biometric traits with a recorded model, therefore safeguarding access to sensitive financial information. Modern banking solutions and payment technologies employ various biometric modalities, such as fingerprints, iris scans, voice recognition, facial recognition, palm vein scans, behavioural analysis, and other forms of biometric identification. These modalities are used individually or in combination as a multifactorial system to secure accounts and protect against fraudulent activities.
- Mobile: Mobile biometric solutions exist at the convergence of connection and identification. These systems include one or more biometric measures for the purpose of authentication or identification, and make use of smartphones, tablets, other portable devices, wearable technologies, and the Internet of Things for flexible deployment options. Mobile biometrics is becoming significant because to the adaptability offered by contemporary mobile technology and the widespread use of mobile paradigms in several sectors, including consumer, public, and private domains.

- **Eye movements and tracking applications:**

- The automotive industry has identified a well-established correlation between eye movement and attentiveness. Therefore, monitoring the eye movements of the auto-mobile driver may be quite beneficial in quantifying the level of somnolence, fatigue, or lethargy. One may determine the driver's tiredness by measuring either the length and intensity of blinks or the amount of ocular activity [10].

- Screen navigation is a crucial application for those with impairments. The program utilises cameras to monitor and analyse a person’s eye movements, enabling them to navigate a web page, generate text, or execute tasks by selecting buttons on a computer or mobile device. Consequently, this kind of application is receiving more attention lately because of the rapid advancement and the increasing need for new methods of screen navigation, particularly on mobile device platforms.

## 1.6 Performance evaluation

Multiple important factors influence the efficiency of the biometric system. These aspects further contribute to the assessment of the system’s performance. The key considerations are security, scalability, cost-effectiveness, usability, accuracy, and speed.

- **Security:** The integrity of the system directly impacts the effectiveness of the verification system. If the biometric system lacks robust security measures, it will be unable to safeguard any other applications.
- **Speed:** The system should exhibit rapid performance since the consumer desires little waiting time.
- **Scalability:** It refers to the capacity of the system to adapt and function well when new features are introduced or when the system is expanded in size.
- **Usability:** The system should possess a high level of user-friendliness to accommodate individuals from all backgrounds. The interface should possess a high degree of user-friendliness and clarity.
- **Accuracy:** The system should provide precise outcomes for a given input [11].

Several other parameters assess the performance and operation of the system. Performance assessment is essential since it allows the owner to thoroughly analyse and assess the system, enabling them to effectively manage it and identify any faults that may impact its performance.

Performance metrics of a biometric system are used to assess the efficiency and accuracy of the system in identifying or verifying persons via several rates. According to ISO/IEC 19795-1, this is a standard set by The International Organisation for Standardisation (ISO). The fundamental error rates are as follows:



- **Failure-to-Acquire Rate (FTA):** FTA measures the proportion of verification or identification attempts where the biometric system fails to acquire the necessary biometric information.
- **Failure-to-Enrol Rate (FTE):** FTE measures the proportion of individuals for whom the system could not generate a biometric model during enrolment. This can occur for various reasons, such as the individual having no fingerprints due to genetic reasons, medical conditions, or professional damage.
- **False Non-Match Rate (FNMR):** FNMR measures the proportion of times the system incorrectly rejects a valid match. This occurs when the acquired biometric data fails to match the stored biometric model of the same individual, resulting in a false rejection.
- **False Match Rate (FMR):** FMR measures the proportion of times the system incorrectly accepts a match between the acquired biometric data and the biometric model of another individual. This results in a false acceptance.

### 1.6.1 Verification system error rates

The system's performance may be verified by considering various factors, such as the False Accept Rate (FAR), False Reject Rate (FRR), and Crossover Rate (CER) or Equal Error Rate (EER) [12]. Usually, this necessary compromise between them is achieved by modifying a threshold. The performance in question may be represented by using a Receiver Operator Characteristic (ROC) curve.

- **False Rejection and False Acceptance Error Rates:** The False Reject Rate or False Non-Match Rate (FRR or FNMR) measures the probability that the system will not recognise a match between the input pattern and its corresponding template in the database. This rate reflects the percentage of legitimate inputs that are mistakenly rejected. The FRR is defined mathematically as follows:

$$FRR = \frac{\text{Number of False Rejections}}{\text{Total Genuine Attempts}} \times 100 \quad (1.1)$$

- **False Accept Rate or False Match Rate (FAR or FMR):** It represents the system's probability of making an incorrect match of an input pattern with a non-matching template from the database. It gives the percentage of invalid inputs that

are accepted. The FAR is defined as follows:

$$FAR = \frac{\text{Number of False Acceptances}}{\text{Total Impostor Attempts}} \times 100 \quad (1.2)$$

In order to achieve the highest level of security, it is necessary to maintain a balance between FAR and FRR. Figure 1.2 displays the error rate of the biometric system in verification mode. It is evident that a stricter criterion reduces the FAR but increases the FRR, resulting in fewer impostors being mistakenly identified as matches. In contrast, reducing the threshold decreases the FRR but increases the FAR, resulting in fewer legitimate instances being rejected. This compromise is vital for setting up the system for certain applications.

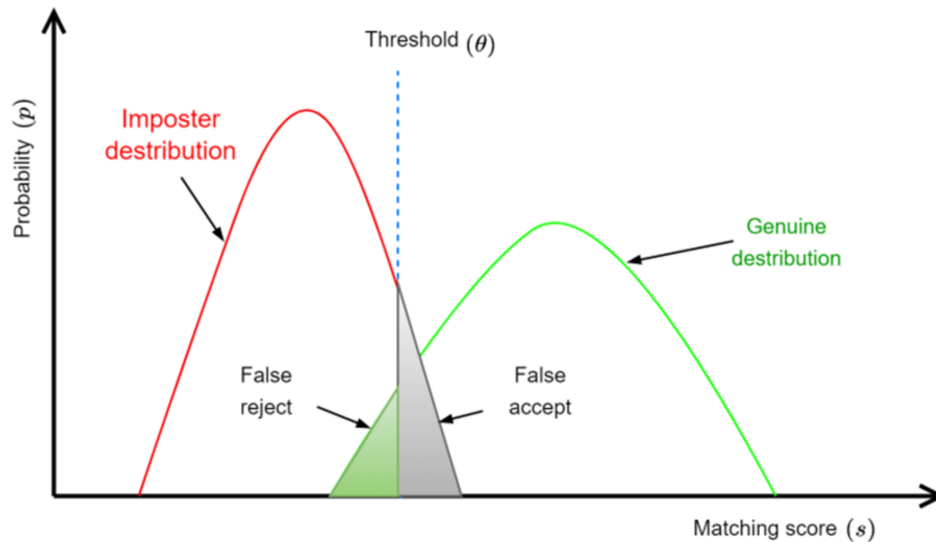


Figure 1.2: Distribution of error rates in a biometric verification system.

As the score threshold rises, fewer imposters incorrectly identified as matches. Simultaneously, more genuine attempts are incorrectly classified as non-matches.

- **Receiver Operating Characteristic Curve (ROC):** This curve is a statistical measure used to evaluate the performance of a binary classification model. ROC plot provides a graphical representation of the balance between the FAR and the FRR, as shown in Figure 1.3. Typically, the matching algorithm makes a judgment by using a threshold to establish the level of similarity between the input and a template that is required for it to be classified as a match. Decreasing the threshold will result in a decrease in false non-matches but an increase in false accepts. Similarly, raising the threshold will decrease the FAR but raise FRR. An often seen variant is the Detection

Error Trade-off (DET), which is derived by using normal deviate scales on both axes. This graph provides a clearer representation of the variations in performance at higher levels (with fewer mistakes) [13].

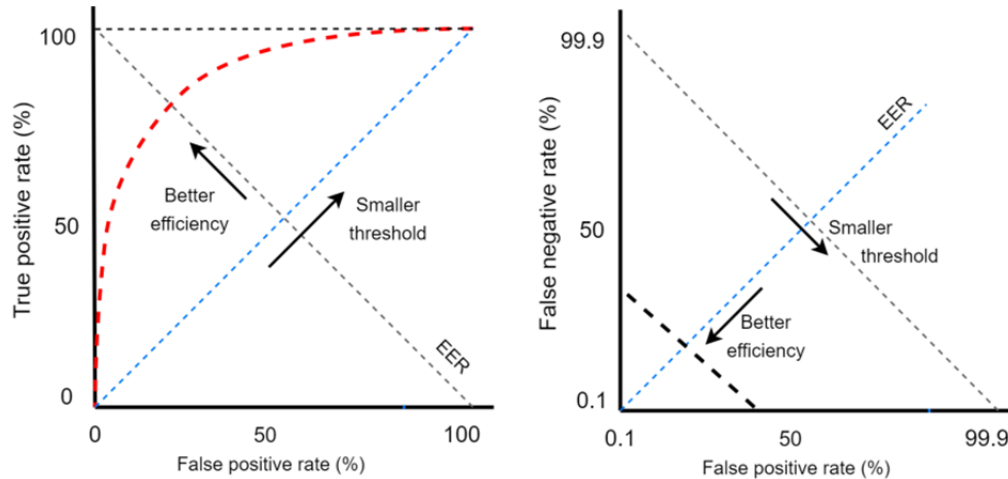


Figure 1.3: Illustration of ROC (left) and DET (right) curves.

The Equal Error Rate (EER), also known as the Crossover Error Rate (EER or CER), refers to the rate at which the number of mistakes in accepting and rejecting decisions are the same. EER value may be readily derived from the ROC curve. EER provides a quick way to compare the accuracy of systems with various ROC curves. Typically, the system with the lowest EER is the most accurate. The point when the FAR and FRR have equal values may be determined by analysing the ROC plot. A system is regarded more accurate when it has a lower EER.

### 1.6.2 Identification system error rates

The Recognition Rate (RR) is a metric that may be used to evaluate the effectiveness of an identification system. The information provided by RR is clear and direct. The proportion of participants who were previously enrolled and had their identities properly recognised may be mathematically stated as shown in Eq. 1.3.

$$RR = \frac{\text{Number of correct recognised images}}{\text{Total number of images}} \times 100 \quad (1.3)$$

CMC curve, also known as the Cumulative Match Characteristic curve, is a performance graph that is widely used in biometric systems, particularly in identification settings. It

quantifies the probability of finding a correct match within a certain number of highest-ranked matches given by the system. Essentially, it demonstrates the efficacy of a biometric system in accurately comparing a specific biometric sample with a database of samples. As illustrated in Figure 1.4, the rank is represented along the horizontal axis of the CMC curve, while the identification accuracy or probability of correct identification is depicted along the vertical axis. Each point on the curve indicates the likelihood that the correct database entry is within the top N matches.

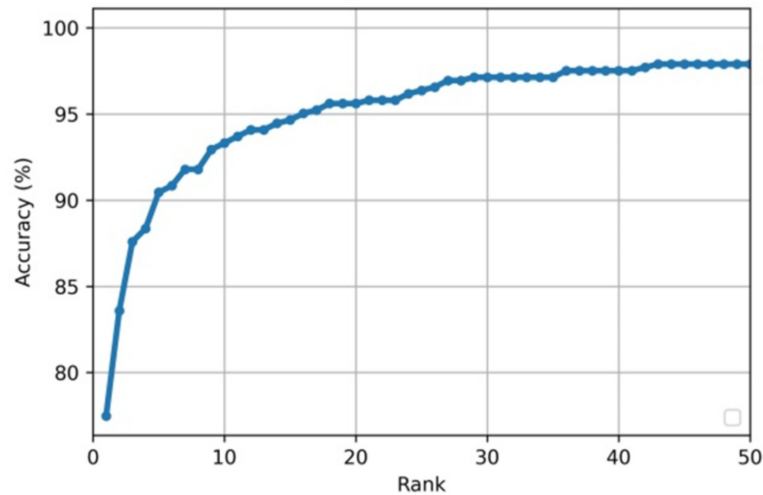


Figure 1.4: CMC curve representation.

## 1.7 Biometric modalities

When discussing biometric modalities, it is important to differentiate between physiological and behavioural human characteristics. A physiological characteristic refers to a stable and inherent physical attribute of a person, such as a fingerprint, iris pattern, or blood vessel pattern located in the rear of the eye. This form of measurement remains constant and immutable until subjected to tremendous pressure. On the other hand, a behavioural feature is an expression of an individual's psychological composition, while physical attributes, such as size and gender, have significant effect. Examples of behavioural attributes often used for person identification include keystroke dynamics, as well as voice identification and/or verification. Table 1.1 summarises and classifies the most often used biometric modalities.

Table 1.1: Most of used biometric modalities.

Physical Biometrics	Behavioural Biometrics
<ul style="list-style-type: none"> <li>- <b>Fingerprint:</b> Analysing fingertip patterns.</li> <li>- <b>Face recognition:</b> Measuring facial characteristics.</li> <li>- <b>Hand Geometry:</b> Measuring the shape of the hand.</li> <li>- <b>Iris scan:</b> Analysing features of coloured ring of the eye.</li> <li>- <b>Retinal scan:</b> Analysing blood vessels in the eye.</li> <li>- <b>Vascular patterns:</b> Analysing vein patterns.</li> <li>- <b>DNA:</b> Analysing genetic makeup.</li> <li>- <b>Ear recognition:</b> Measuring the shape of the ear.</li> </ul>	<ul style="list-style-type: none"> <li>- <b>Speaker/voice recognition:</b> Analysing vocal behaviour.</li> <li>- <b>Signature/handwriting:</b> Analysing signature dynamics.</li> <li>- <b>Keystroke/patterning:</b> Measuring the time spacing of typed words.</li> <li>- <b>Gait Biometrics:</b> Analysing the walking style.</li> </ul>

### 1.7.1 Fingerprint

Fingerprint recognition is the most established and widely recognised biometric authentication method. It is an automated, digitised version of the antiquated paper-and-ink system that law enforcement agencies employed for identification. It operates by analysing the characteristics of a given individual’s fingerprint in order to identify them. The fingerprint of every person is different and fixed, and the basic features do not change with the passage of time. Different fingerprints of similar twins can be identified. Moreover, fingerprints are different on the fingers of the same person’s two hands [14].

A fingerprint consists of raised ridges and sunken furrows. The distinctiveness of a fingerprint is determined by the arrangements of ridges, furrows, and minutiae points on the finger. The three fundamental categories of ridge patterns are loops, whorls, and arches (Figure 1.5). Fingerprint biometrics uses the fine details and ridge pattern comparison in order to identify individuals.

Fingerprint recognition has various advantages as well as limitations. The fingerprint modality is characterised by its intrinsic advantages: uniqueness, wide distribution range,

and resistance to natural changes over time. It does not require additional or abundant space. However, scars, cuts, or a lack of fingers can pose challenges for identification, highlighting some of the drawbacks in specific situations. Additionally, the fingerprint modality is intrusive, requiring physical interaction with the system and physical presence. Fingerprints can also be easily deceived by counterfeit fingers made from materials like polish.



Figure 1.5: Fingerprint patterns.

### 1.7.2 Face

For generations, humanity has relied on the face as the primary biometric attribute for personal identification. It has become a consistent characteristic in customised papers such as identity cards, driving licenses, passports, and so on. Face recognition, or facial recognition (Figure 1.6), is a technique that uses digital images or video frames to identify or verify persons based on their facial traits. It is a type of technology that uses the distinct features of a person's face, such as the positions and distances between the eyes, nose, and mouth, to identify or authenticate individuals [15]. This technology is often used because to its non-intrusive and contactless process; yet, the achieved accuracy is generally less than that of fingerprint and iris recognition [16].

Face recognition has several advantages, such as its exceptional speed in identification, its capability to operate with cameras of lesser resolution, and its status as the most captivating and advanced technology among biometric systems. A fundamental limitation of this method is that face features might change over time or due to surgical procedures. Moreover, this method proves to be inefficient when the person uses a mask. Ensuring enough lighting is essential, and modifying facial expressions might affect the final image. 3D images provide superior accuracy for authentication purposes owing to their increased degree of information compared to 2D images.

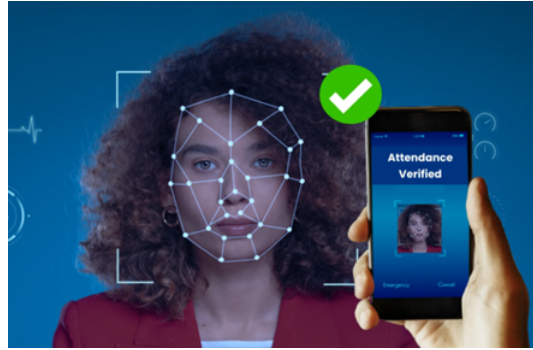


Figure 1.6: Face recognition.

### 1.7.3 Iris

The iris modality is among the limited number of biometric systems that are highly efficient in authentication mode. The iris, which is the pigmented and coloured part of the eye, exhibits multifaceted patterns and contains over 200 distinct eye spots. These patterns transmit a significant amount of information. The main distinction of amazing eye spots is the presence of tissue that gives an appearance of separating the iris into circular patterns, rings, furrows, markings, and a corona [17]. Each individual have a unique iris that varies in terms of its traits, physical attributes, and visual appearance. A representative iris sample is shown in Figure 1.7.

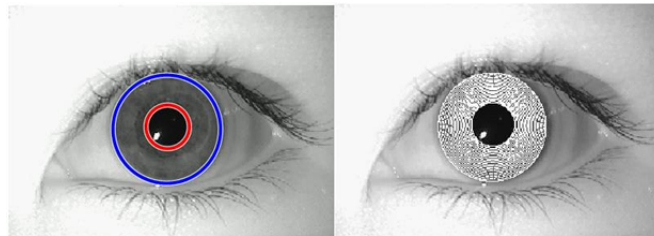


Figure 1.7: Iris sample.

The use of human iris for identification has many benefits, including a low probability of similarity, with only 1 in 10 million persons having a comparable iris. This high level of accuracy makes it a reliable and trustworthy method of identification. The system is extensively secured, with a little incidence of incorrectly accepting unauthorised individuals, and iris recognition delivers outcomes within a brief time frame of 2 to 5 seconds. In addition to its benefits, this approach also has constraints. It has little competition in the market, is susceptible to interference from a significant distance, and relies on user cooperation for optimal scanning.

It is important to note that iris recognition is different from retinal scanning, which focuses on blood vessels in the retina. Iris recognition is exceptional in its ability to uniquely identify individuals based on their irises, making it a powerful tool for secure authentication.

### 1.7.4 Ear

Ear biometrics is an emerging field of biometric identification that leverages the distinct characteristics of a person's ear for the purpose of personal identification and authentication [18]. Ear biometrics has distinct advantages compared to other biometric methods like fingerprints or face identification, due to the consistent and unchanging structure of the ear across a person's lifespan. The ear's shape, size, and contours are very distinctive and challenging to modify or conceal, making it an accurate tool of identification. In addition, ear biometric systems are capable of functioning efficiently in many environmental conditions and do not need the subject's consent, since they may be recorded remotely without any physical touch [19]. The non-intrusive and resilient characteristics of ear biometrics make it a very promising technology for use in security, surveillance, and personal identity applications. Figure 1.8 illustrates the anatomy of the human ear.

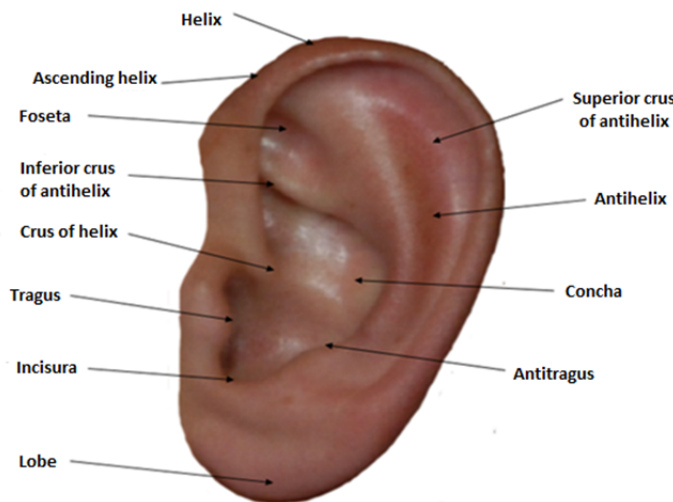


Figure 1.8: Anatomy of the ear.

### 1.7.5 Hand geometry

Hand geometry biometrics involves taking a three-dimensional image of the hand and quantifying its form, finger length, and knuckle dimensions. Hand geometry has been used for several years in diverse applications, mostly for the purpose of access control. The technology



does not attain the utmost levels of accuracy, but it is convenient and expeditious to use. During the capture procedure, the user lays their hand on the reader and aligns their fingers with guidelines that are positioned in a certain way. Cameras located above and on the side of the hand, produce images that are used to obtain measurements at certain spots [20]. Figure 1.9 displays a biometric system based on hand geometry.



Figure 1.9: An example of hand geometry reader.

### 1.7.6 Gait

Gait biometric recognition is a technique used to identify people by analysing their unique walking rhythms. This method examines the synchronised and repetitive sequence of motions in human locomotion. Gait biometrics, a relatively new methodology, has a 90% accuracy rate, which is higher than standard methods such as fingerprint and facial identification [21].

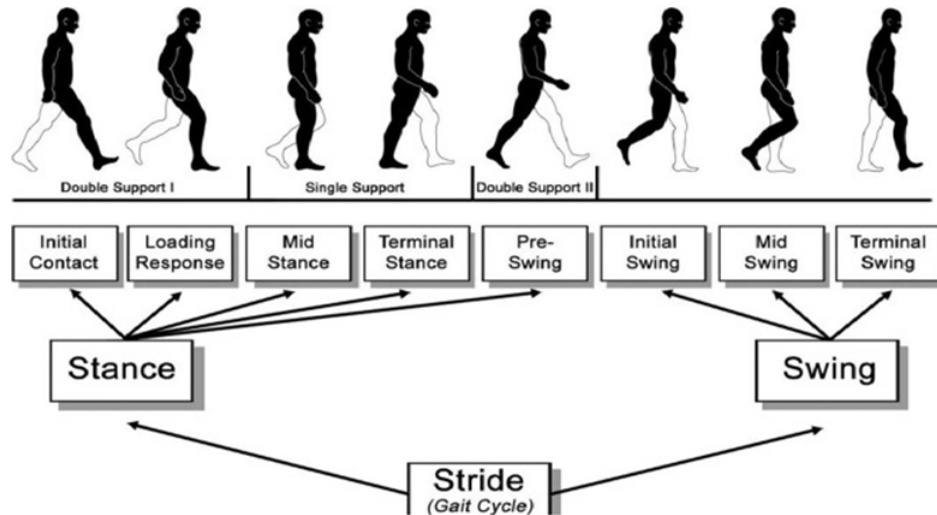


Figure 1.10: Gait cycle.

People may often be identified by common characteristics of their walking style, such as

body alignment, swaying, and the spacing between their feet. Gait biometrics has several appealing characteristics, with one of the most prominent being the ability to record an individual's gait without their active participation or conscious knowledge.

Current research is centred on visually-based systems that use video cameras to assess the movements of each individual part of the body. A gait cycle, or stride, consists of two distinct stages: the stance phase, during which one foot is in contact with the ground, and the swing phase, during which the other foot is not in contact with the ground.

### 1.7.7 Signature

Signature biometrics is a kind of behavioural biometrics that uses the unique pattern of an individual's signature to identify them. The key distinction of this approach is in the manner in which a person performs their signature, placing more emphasis on the behavioural patterns inherent in the act of signing rather than the appearance of the signature itself [22]. There are two main approaches to signature recognition: static and dynamic. In the static, or off-line, approach, the signature is first recorded on paper and then converted into a digital format using either a camera or an optical scanner. Figure 1.11 illustrates the signature biometric.

The technology identifies the signature by assessing characteristics such as velocity, speed, and pressure. The dynamic approach, also known as the on-line method, involves capturing the signature in real-time utilising a computerised tablet. This approach captures several behavioural characteristics such as velocity, force, and direction of strokes, size of signature, and the duration of signing.



Figure 1.11: Online and off-line signatures.

### 1.7.8 Retina

The blood vessels in the retina are illuminated with a low-intensity coherent light source during retinal imaging; the illuminated vessels are subsequently captured on camera. By virtue

of the rapid absorption of infrared radiation by the blood vessels in the retina relative to the adjacent tissue, the retinal blood vessel pattern can be observed and analysed. Retinal scans find applications across diverse domains such as ophthalmology, medicine, and biometric security. In addition to biometric identification and the diagnosis and monitoring of ocular diseases, they can provide vital information. Retinal scans consist of blood vessels arranged in a distinctive pattern, which presents a promising biometric identification technique. Due to its extremely individualised pattern of blood vessels, the retina is an invaluable instrument for security and authentication [23]. A retinal image is shown in Figure 1.12.

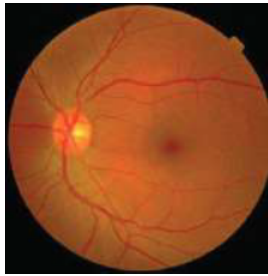


Figure 1.12: A retinal image.

## 1.8 Comparison between biometric modalities

A comparative analysis of various biometric methods, taking into account factors such as universality, uniqueness, collectability, permanence, performance, acceptability, and circumference. In detail, these factors are discussed in Section 1.4. Table 1.2 presents a comparison of biometric characteristics according to various criteria [24].

Table 1.2: Comparison between different biometric modalities.

Modality/criteria	Universality	Uniqueness	Collectability	Permanence	Performance	Acceptability	Circumvention
Face	High	Medium	High	Medium	Low	High	High
Fingerprint	Medium	High	Medium	High	High	Medium	Medium
Hand geometry	High	Medium	High	Medium	Low	High	High
Hand veins	Medium	High	Medium	High	High	Medium	Medium
Iris	High	High	High	High	High	Medium	Low
Retina	High	High	Medium	High	High	Low	Low
DNA	High	High	Low	High	High	Low	Low
Gait	High	Medium	High	Medium	Low	Medium	Medium
Odor	High	High	Low	High	Low	Medium	Low
Ear	Medium	Medium	Medium	High	Medium	High	Medium
Hand Vein	Medium	Medium	Medium	Medium	Medium	Medium	Low
Signature	Low	Low	High	Low	Medium	High	High
Keystroke	Low	Low	Medium	Low	Low	Medium	Medium
Voice	Medium	Low	Medium	Low	Low	High	High

## 1.9 Why this thesis focuses on ear biometrics?

The human ear has attracted interest as a viable biometric modality because of its stability and distinctiveness. Although the ear has great potential, there are now a limited number of software applications available that make use of it for personal identification [25]. The ear is often regarded as one of the most reliable anatomical traits of the human body, which makes it an excellent choice for biometric identification. This constancy is especially remarkable when compared with other modalities, such as the face, which may be affected by things such as cosmetics and haircuts. Moreover, human faces exhibit diverse reactions to expressions and emotions such as fear, sadness, joy, or surprise [18, 25].

Nevertheless, the presence of hair or curls might result in partial occlusion. It is important to emphasise that the general public has a significant degree of approval for the use of ear modality in access control and security applications, such as visa and passport programs. Furthermore, there is no need to make touch with the sensor, which eliminates any concerns about cleanliness. Imaging capture may be done discreetly from a distance without any need for connection between the user and the sensor.

In addition, ear images provide less privacy problems compared to face images due to the difficulty in visually associating an ear image with a specific individual. Most people are unable to recognise their own ear images. Consequently, databases that save ear images are less prone to danger compared to databases that store face images, since the possibility for abuse is higher. Combining ear images with face images may enhance the accuracy of identification [19].

The field of ear recognition has seen significant advancements with the adoption of deep learning techniques, particularly for biometric recognition and identification. Deep learning models have been applied to various aspects of ear recognition, including feature extraction, classification, and multimodal recognition. These techniques have addressed the limitations of traditional machine learning algorithms, showing promise in handling unconstrained databases and achieving high accuracy in ear recognition tasks. Thus, we proposed several promising biometric approaches based on deep learning.

## 1.10 Conclusion

Biometric recognition is the process of verifying the identity of a person by examining their distinct physical or behavioural traits. This chapter clarifies the notion of evaluating architectural design in relation to human well-being, by using several modalities to observe an

individual's biometric characteristics. Biometrics is used in many sectors such as healthcare, banking, government, and education. The latter has successfully used biometrics to achieve positive outcomes.

Biometric identification systems are advanced automated systems designed to address the limitations of traditional identifying methods. Nevertheless, biometric-based systems also include some limitations that hinder the advancement of biometric technologies. This chapter presents a comprehensive introduction to the often used biometric features and their characteristics, various biometric working modes, as well as the limits and weaknesses of the system. Nevertheless, despite the potential risks, biometrics offers compelling security solutions and continues to be an expanding technique of identification, holding great potential for the future. However, it has been shown that human identification is more precise and appealing. Multiple scholarly publications and research articles have examined and summarised the advancements made in cost-effective and sophisticated design methodologies. Exploring several modalities contributes to the development of an exceptional system.

The research and improvements in ear biometrics have shown that ear recognition has the potential to be a reliable and efficient biometric modality for a range of applications, such as, security, and user identification. The integration of deep learning techniques, multimodal biometric systems, and advanced feature extraction methods has enhanced the accuracy, resilience, and efficiency of ear biometric identification systems. Consequently, we selected the ear as a biometric modality. The next chapter will provide a comprehensive examination of the basic principles of the ear as a biometric modality.

## Chapter 2

### Ear Biometrics

## 2.1 Introduction

In this thesis, we have chosen ear recognition as the main topic of our study because to its distinctive characteristics that makes it a favourable choice for identification purposes, unlike other modalities. Currently, this modality is a highly researched topic and the effectiveness of its systems is influenced by the context in which data is collected.

This chapter is organised as follows: Section 2 contains a detailed motivation of our choice. Section 3 introduces the characteristics of the ear, followed by ear recognition system presentation in section 4. In section 5, we introduce and categorise most ear recognition methods, with a particular attention given to deep-learning-based methods. Section 6 is devoted to discuss the open challenges related to ear biometrics. Finally, we conclude this chapter in section 7.

## 2.2 Motivation: Why ear biometric?

The human ear is measured as a new feasible class of biometrics with some additional advantages compared to other biometric modalities, and no doubt the ear presents unique characteristics that make it a promising biometric modality.

Identifying people by their ears has gained significant importance due to several reasons:

- It has a comprehensive and stable structure that does not change considerably over time, and its form does not change by facial expressions.
- It can also be captured from a distance and without help from the user.
- It can be used as a complementary part to the face in a multi-biometric system.
- It will be helpful to directly answer phone calls without unlocking and controlling various features from a distance.
- Ear images can be used for biometric identification without requiring active user participation. Unlike fingerprints and iris scans, which often involve direct interaction, ear recognition can be done passively. The unique shape and features of the ear, such as its contours and ridges, provide a reliable basis for identification. This makes ear-based biometrics an interesting area of research and application[26].

Comparing to other modalities like face, fingerprint, or iris scanning, ear biometric systems eliminate the need for users to perform any physical movements, such as positioning a finger or face. The process becomes more seamless and minimises any disruptions.



The ear exhibits a range of distinctive attributes that remain consistent and unaltered over an individual's lifetime, offering a reliable biometric characteristic that remains unaffected by ageing or emotional states. Unlike face recognition systems, the ear is not easily concealable with items like spectacles, beards, or moustaches throughout the process of acquiring data. Nevertheless, hair or curls may to some extent impede the view.

The on-going worldwide epidemic resulting from the novel coronavirus (COVID-19) has compelled the mandatory utilisation of facial masks in public [27]. Consequently, this has posed a significant challenge to facial recognition. Moreover, the problem is further highlighted in recognition systems, especially in surveillance settings, due to the masks obstructing a significant part of the face. This has increased the significance of research on ear recognition.

Because of these qualities, the interest in ear recognition systems has developed significantly in recent years.

## 2.3 Characteristics of the ear

The human ear develops early during pregnancy and is fully formed by birth. Due to its function as the human hearing organ, the ear has a characteristic structure shared across people. In 1890, French criminologists first discovered that the human ear structure is unique and suggested its use as a biometric modality [28]. Later, after years, it proved practically by investigating 10,000 ear images and found that every ear image features are unique and different from others [29]. The ear's shape remains unchanged over the age of 70 years [30]. Generally, a person's left and right ears are similar to the amount that matches the right ear to the left ear. The human ear has eleven unique and major parts from all thirty-seven ear biometric features. The outer part of the ear is known as a helix. The lower part of the ear, called the lobe, surrounds the ear. The antihelix runs parallel to an outer helix. The area between the inner helix and the lower subdivision of the antihelix forms the concha, which has a shell-like form. The lower part of the concha combines into a sharp inter-tragic notch [31]. The crus of helix is the intersection between the helix and antihelix. A little prominence on the right side of the inter-tragic crash is antitragus. The outer ear dominates by the shape of the helix rim and lobe. The inner ear has many prominent features like antihelix, incisura inter-tropical, concha, triangular fossa, crus of helix, and tragus. Figure 2.1 presents the anatomical structure of the outer ear shape.

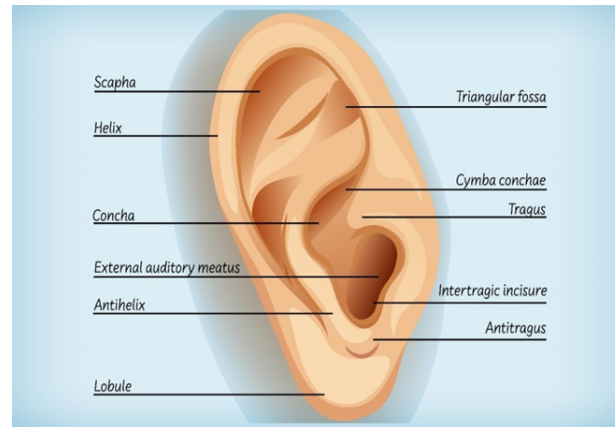


Figure 2.1: Anatomical structure of the outer ear shape.

## 2.4 Ear recognition system

The ear recognition process comprises many essential sub-processes, each of which has an essential role in correctly identifying and authenticating people based on their ear characteristics. The sub-processes include ear detection, pre-processing or normalisation, feature extraction, and classification or identification (Figure 2.2).

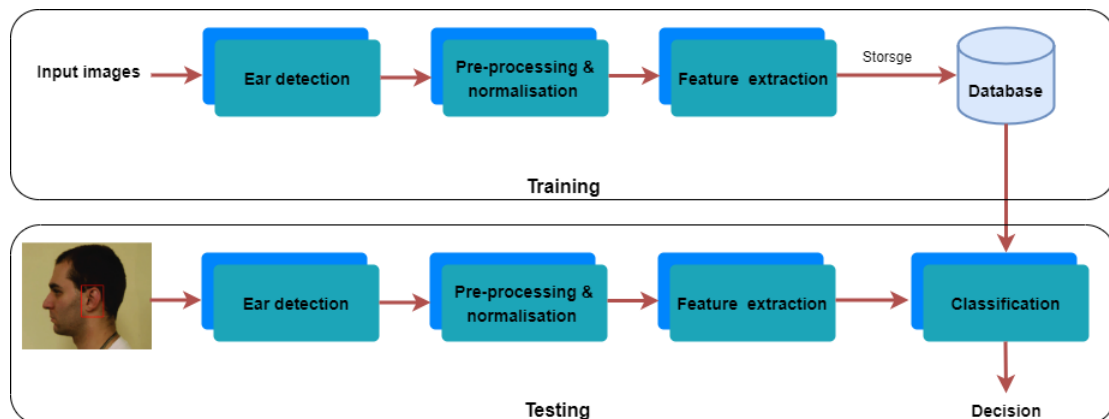


Figure 2.2: Flowchart of an ear biometric recognition system.

### 2.4.1 Ear detection

Ear detection is the process of identifying the location of the ear in images, which may include one or several ears. The objective is to locate the ear box with maximum precision in the images as seen in Figure 2.3. Efficient ear identification algorithms are essential for accurately recognising and extracting ear characteristics for further processing.

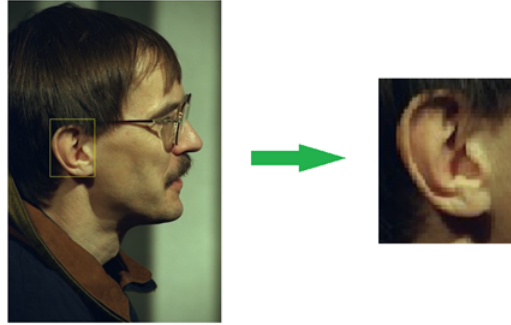


Figure 2.3: An example of ear detection.

### 2.4.2 Pre-processing or normalisation

The pre-processing or normalisation phase is an intermediary step that seeks to improve and streamline the classification process. The process involves eliminating irrelevant data from ear images and rectifying issues such as lighting and occlusions. Implementing efficient pre-processing techniques may enhance the quality of the input data.

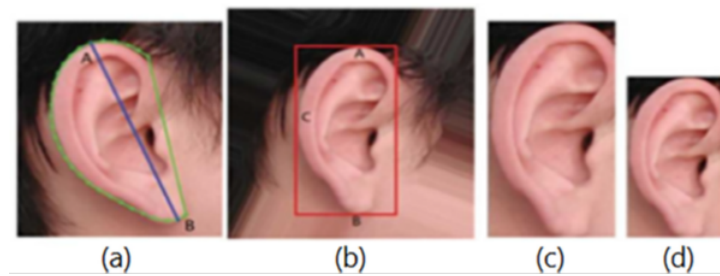


Figure 2.4: Example of ear normalisation [32].

### 2.4.3 Feature extraction

Feature extraction is a crucial stage that has a direct influence on the final performance of the system. The process entails extracting distinctive characteristics from images of the ear. Feature extraction may be achieved by using either handcrafted or deep learning-based features. The selection of the feature extraction approach has a substantial impact on the accuracy and resilience of the ear recognition system.

### 2.4.4 Classification or identification

During the classification or identification step, the extracted features are used to ascertain or authenticate the subject's identity. This stage usually involves employing conventional

classification methods or distance metrics to compare the extracted features with enrolled models for the purpose of recognition.

The efficiency of each sub-process is vital for the overall performance of the ear recognition system. In order to achieve high identification accuracy and reliability, it is important to have efficient ear detection, robust pre-processing techniques, discriminating feature extraction, and accurate classification methods.

## 2.5 Ear recognition methods

Ear images are commonly used in biometric systems to extract relevant features, which are subsequently compared to pre-existing models stored in a database. Ear recognition approaches may be categorized into two main categories: namely, handcrafted and deep-learning-based, based on the type of data and the strategy used for feature extraction.

### 2.5.1 Handcrafted methods

Early efforts in ear biometrics focused on employing handcrafted algorithms for feature extraction, primarily targeted at defining the anatomical structure of the ear. For example, Hassaballah et al. [33] Conducted a comprehensive analysis and comparison of various extensions of the local binary pattern (LBP) descriptor with the goal to enhance ear recognition. In addition, they aimed to enhance performance by implementing an update variant known as averaged local binary pattern (ALBP) using a simple thresholding technique. In a further study, Hassaballah et al. [34] established the robust local oriented pattern (RLOP) using LBP, elucidating enhanced resistance to rotation and noise in unconstrained ear recognition, surpassing the performance of standard LBPs and their variations. Sarangi et al. [35] used the Jaya algorithm to improve the quality of ear images. The researchers used the speeded-up robust features (SURF) descriptor to extract local invariant pose characteristics. Sajadi and Fathi [36] introduced a hybrid method that combines local and global features in the frequency domain to depict ear images. A genetic algorithm was used to optimise feature selection and improve the accuracy of recognition by using the local phase quantisation (LPQ) texture descriptor for local features and the Gabor-Zernike (GZ) operator for global features. Boujnah et al. [37] proposed a novel approach for identifying humans in challenging conditions by using ear-prints. The local and frequency domain properties are combined into a single vector. The dual tree complex wavelet (DTCW) transformation is determined by six rotation angles and five scales, which are used to forecast spectral-saliency-based dual tree

complex wavelets. Statistical properties are computed from the vector created and its derivatives. In addition, they included a set of essential factors derived from the implementation of Harris descriptors to improve the system's effectiveness in handling changes in size, rotation, and translation. Regouid et al. [38] investigated the conversion of two-dimensional ear images into one-dimensional representations. Their inquiry centred on the one-dimensional local binary pattern (1D-LBP) descriptor and its variants as a substitute technique for extracting features.

Although these handcrafted methods have shown impressive accomplishments in ear recognition, their underlying structural limitations restrict the amount of discriminative information that can be extracted from acquired images.

## 2.5.2 Deep learning methods

Recently, the focus in ear biometrics has significantly shifted from traditional feature engineering methods to approaches based on deep learning paradigms. [39]. Alshazly et al. [40] performed a comparative study that examined the differences between ear recognition models based on handcrafted techniques and convolutional neural networks (CNNs). Their assessment, which focused on unconstrained ear recognition, included seven well-established handcrafted descriptors and four CNN models based on variants of AlexNet. CNNs demonstrated higher recognition accuracy in comparison to handcrafted features. Subsequently, Alshazly et al. [41] proposed a novel ear recognition technique based on ensembles of deep CNNs models. CNNs were specifically trained on visual geometry group (VGG)-based network topologies to extract deep characteristics. After conducting tests on several VGG representations, it was determined that the VGG-13-16-19 configuration was the most optimum choice. Priyadharshini et al. [31] developed a specialized six-layer deep CNN designed for the purpose of identifying ears. Their investigation included methodical testing of activation functions, learning rates, kernel sizes, and epochs to evaluate the suggested model. CNNs have shown the capacity to learn the most effective features directly from input images, outperforming conventional descriptors. Khaldi and Benzaoui [42] tackled the issue of test pictures that lack colour information for a system that was trained on colour images. The researchers used a deep convolutional generative adversarial network (DCGAN) model to add colour and enhance greyscale and dark images. Subsequently, They deployed a CNN-based classification model to classify ears, resulting in considerable increases in recognition performance. In a recent study, Alshazly et al. [43] improved recognition performance by developing and integrating several ensembles of ResNet architectures with different depths.

Additionally, they investigated the use of a support vector machine (SVM) classifier as an alternative to the conventional dense classifier. Omara et al. [44] introduced a new ear identification technique based on Mahalanobis distance learning, using deep CNN features extracted from VGG and ResNet models. Sharkas [45] proposed a two-stage ear recognition method. In the first stage, the discrete Curvelet transform was used to extract key ear features. In the second stage, end-to-end deep learning network ensembles were employed for classification. Xu et al. [46] introduced a method for ear recognition that uses channel features and dynamic convolution (CFDCNet). CFDCNet modifies the DenseNet-121 model by including dynamic convolution to improve the extraction of ear features. This modification enhances the combination of features within samples belonging to the same class and the distribution of features across different samples. In addition, a channel attention mechanism is used to prioritise significant ear characteristics and suppress irrelevant ones, hence enhancing the discriminative power of the feature representation. Aiadi et al. [47] introduced an effective unsupervised lightweight network that uses gradient magnitude and direction alongside responses of data-driven filters (MDFNet) for ear print recognition, showcasing its simplicity, robustness, and high performance compared to existing methods in the field. The experimental results validate the effectiveness of MDFNet in achieving accurate and robust ear recognition results, emphasising the importance of ear alignment in the process.

Tables 2.1 and 2.2 provide a summary of the research discussed in this chapter, organising them based on their categories, the datasets employed, and the experimental protocols.

Table 2.1: Handcrafted approaches comparative summary.

Paper	Method	Employed dataset			Evaluation protocol
		Name	Sub	Img	
Hassaballah et al. (2019)	LBP's variants	IITD-1	125	493	5-fold cross-validation
		IITD-2	221	793	
		WPUT	474	3348	
		AMI	100	700	
		AWE	100	1000	
Hassaballah et al. (2020)	(RLOP)	IITD-1	125	493	5-fold cross-validation
		IITD-2	112	793	
		AMI	100	700	
		AWE	100	1000	
Sarangi et al. (2020)	Jaya algorithm + SURF	IITD-2	221	793	442 Train 221 Test
		UND-E	114	464	228 Train 114 Test
		UND-J2	415	2413	546 Train 273 Test
Sajadi and Fathi . (2020)	LPQ + GZ	USTB-1	60	180	5-fold cross-validation
		IITD-1	125	493	
		IITD-2	221	793	
		AWE	100	1000	
Boujnah et al. (2020)	DTCW+ Harris descriptors	USTB-1	60	180	2 img/sub Train remaining Test (3 perm)
Regouid (2022)	1D-LBP	AMI	100	700	60% Train 40% Test
		USTB-1	60	180	
		USTB-2	77	308	
		AWE	100	1000	

Ear recognition is the most promising candidate in multi-pose face recognition. In today's era, ear detection, and recognition systems have reached a certain level of maturity [48]. However, the constraints of ear biometrics resulting from inaccurate ear-detecting methods and the difficulties linked to the decline in performance of ear images when there are changes in posture and image conditions are significant factors to be taken into account in the area of ear identification. The following section outlines some unresolved research issues and potential study areas that are deemed crucial and need more investigation in future research. In addition, we provide our perspectives on new concepts to contribute to future study.

Table 2.2: Deep learning approaches comparative summary.

Paper	Method	Employed dataset			Evaluation protocol
		Name	Sub	Img	
Alshazly et al. (2019)	AlexNet (Fine Tuning)	AMI CVLE	100 16	700 804	60% Train 40% Test
Alshazly et al. (2019)	VGG-13-16-19ensembles	AMI WPUT	100 474	700 3348	60%Train 40% Test
Priyadharshini et al. (2020)	CNN	IITD-2	221	793	490 Train303 Test
Khalidi and Benzaoui (2021)	DCGAN+ VGG-16	AMI AWE	100 100	700 1000	60%Train 40% Test
Alshazly et al. (2021)	ResNet ensemble	AMI AMIC WPUT AWE	100 100 474 100	700 700 1960 1000	60%Train 40% Test
Omara et al. (2021)	Learned Maha-lanobis distance from deep CNN features	AWE AMI WPUT USTB-II	100 100 475 77	1000 700 1957 308	60% Train 40% Test
Sharkas (2022)	DCT different CNNs	AMI IITD-1	100 125	700 493	70% Train 30% Test
Xu et al. (2023)	CFDCNet	AMI AWE	100 100	700 1000	Hold-outcross-validation
Aiadi et al. (2023)	MDFNet	AWE AMI IITD-2	100 100 221	1000 700 793	60% Train 40% Test



## 2.6 Open challenges

Ear recognition is a very nascent field in comparison to other biometric methods. There are several challenges and research concerns that have been thoroughly examined in other domains, but need more investigation in the subject of ear recognition. This presents potential for future study. This section provides a concise overview of the key unresolved issues and the most favourable directions for study in this domain.

- *Ear symmetry (right & left right ears)*: The level of bilateral symmetry between the left and right ears is still uncertainly. Several studies have investigated ear symmetry by analysing either the form or structure [49]. Approximately 90% of persons have bilateral symmetry between their right and left ears. However, the performance of symmetry in these individuals is not deemed sufficient. The analysis of human ear symmetry aims to assess its influence on the effectiveness of ear recognition systems. Further empirical research and investigations are necessary to accurately identify the level of symmetry and use it in automated ear-based recognition systems.
- *Stability of the ear over time*: The ear structure's age-invariance and lack of large changes are still debated, with few studies done. Ear recognition is less affected by age-related changes than other biometric systems like the face, even though ears get larger with age. However, experimental studies are needed to determine how age-related changes and the gap between ears images affect ear-based identification. A significant constraint is the lack of suitable long-term datasets to examine age-invariant ear recognition. Therefore, long-term data collection is necessary.
- *Exposure to spoofing attempts*: For biometric identification, the ear offers several advantages and has been widely studied. Displayed or printed graphics may be utilised to fool ear recognition systems. Presentation, adversarial, and template attacks compromise ear biometric security since these systems cannot always distinguish between fake and real ear images [50]. Due to a dearth of study on ear biometric fraud detection, powerful algorithms are needed to detect and avoid these security concerns. Unfortunately, anti-spoofing databases are lacking in this sector. Suggested detection methods need extensive counterfeit ear image libraries to reliably recognise fake ears in varied scenarios.
- *Ear detection and localisation*: Effective ear recognition systems must recognise and locate the ear in an image or video. These automated identification methods rely on

the ear detecting mechanism's strength and reliability [51]. Academic papers have proposed many methods for ear detection, advancing the discipline. Most ear detectors come from optical object identification advances. However, some approaches only detect ear images in controlled environments. Recent advances in deep learning have enabled detection methods that account for ear image variability. Different domain-specific methods have been developed to identify the ear's unique visual properties [48, 28]. These detection approaches have performed well on moderate-sized ear datasets, but alternative ear localisation processes must be investigated to produce reliable detection methods for large-scale ear datasets gathered in unconstrained conditions. Additionally, most recommended detectors focus on 2D images. However, 3D ear recognition and segmentation remain difficult.

- *Kinship verification from ear images*: Analysing kinship using visual data is a difficult scientific topic with significant practical implications. Although previous studies have mostly focused on examining facial images, there is a possibility to explore additional physical attributes of individuals for the purpose of kinship verification. A recent research examined the issue of kinship verification using ear images to discover whether distinctive visual features can be derived from ear data for this purpose. This study emphasises the possibility of using ear images as a practical option for verifying family relationships, expanding the range of biometric analysis beyond only face images. The results indicate that noticeable physical traits related to familial relationships may be obtained from ear data, highlighting the potential of ear biometrics for kinship verification [52].

## 2.7 Conclusion

This chapter provides an overview of the structure of the ear, the mechanism used for recognising ears, and the many components involved in this process. Several approaches and concepts provide satisfactory outcomes in limited settings. The identification ability tends to decline mostly due to varied conditions. We conducted a comprehensive analysis of the most significant cutting-edge approaches and procedures in ear identification. Specifically, we focused on the extraction of features and the collecting of data for the purpose of recognition. Two main approaches have been developed to achieve complete automation in ear recognition: handcrafted techniques and deep learning.

The performance of unconstrained databases is affected by difficult situations, leading

to a variance in their performance. The findings demonstrate that there is still potential to develop novel models for unconstrained ear identification in order to achieve improved performance and facilitate commercial implementation.

Deep learning approaches shown superior performance compared to conventional approaches, particularly in unconstrained situations. Due to the appearance of many unconstrained ear datasets, conventional machine learning methods saw a decline in performance. Several of these obstacles have been successfully addressed by using deep learning methodologies.

Hence, we made the decision to embark on a research endeavour using deep learning techniques. The next chapters give a thorough analysis and comparison of our suggested techniques with the most advanced methods currently available.

# Chapter 3

## Review of Deep Learning

## 3.1 Introduction

Deep learning, a subset of machine learning, draws inspiration from the neural networks seen in the human brain. The present surge in popularity of this technology may be attributed to recent technological improvements and the widespread availability of vast amounts of data. Deep learning algorithms have the ability to surpass regular machine learning approaches when provided with such data. This technology is now causing a revolution in several scientific domains that intersect with each other, such as computer vision, natural language processing (NLP), voice recognition, and others.

The rest of the chapter is organised as follows: Sections 3.2 and 3.3 represent a background of artificial intelligence and artificial neural networks. Section 3.4 discusses the approaches of deep learning. Section 3.5 outlines convolutional neural networks architectures. Section 3.6 presents some applications of deep learning. Next, we review the standing challenges in deep learning. Last, in section 3.8, we conclude this chapter.

## 3.2 Artificial intelligence, machine learning and deep learning

*Artificial intelligence (AI)* refers to the cognitive abilities shown by machines. Computer science AI research is primarily concerned with the investigation of "intelligent agents," which are machines capable of perceiving their surroundings and taking activities to optimise their likelihood of attaining certain objectives [53]. Artificial intelligence is a phrase often used to describe the replication of cognitive capabilities, like as learning and problem-solving, that are normally associated with human brains. AI has a range of skills such as comprehending human speech, demonstrating exceptional performance in strategic games, operating vehicles without human intervention, enhancing the efficiency of content delivery networks, performing war simulations, and analysing intricate data [54].

*Machine learning (ML)* is founded on the concept that a machine can acquire knowledge from data, recognise patterns, and make decisions with little human involvement [55]. This is the field of research that focuses on algorithms and statistical models that enable computer systems to accomplish certain tasks without relying on explicit instructions, inference, or patterns. Machine learning algorithms construct a mathematical model using a set of sample data and then use this model to make decisions.

Supervised learning and unsupervised learning are two very used machine learning tech-

niques. The majority of machine learning, around 70%, consists of supervised learning. Unsupervised learning comprises 10 to 20%. Semi supervised and reinforcement learning are sometimes used as alternative methods.

**Deep learning**, often referred to as deep machine learning, is the field of research that focuses on artificial neural networks and associated machine learning algorithms that include a number of hidden layers [56]. These deep nets:

- Employ a series of many layers of non-linear processing units to extract and transform features. Each subsequent layer utilises the output from the preceding layer as its input. The algorithms may be categorised as either supervised or unsupervised. They are used for many purposes, such as pattern analysis (unsupervised) and classification (supervised).
- They rely on the unsupervised learning of several hierarchical layers of features or representations of the data. A hierarchical representation is formed by deriving higher level characteristics from lower level ones.
- Belong to the wider domain of machine learning, specifically focused on acquiring data representations.
- Acquire knowledge about several tiers of representations that correspond to distinct degrees of abstraction; these levels create a structured system of ideas arranged in a hierarchical manner.

With the availability of advanced computing power and extensive datasets at their fingertips, deep learning becomes a powerful technology that has revolutionised many manual tasks in different fields. It has overcome the limitations of traditional machine learning techniques by learning representations of data through multiple layers of non-linear transformations. It has successfully resolved several problems of traditional machine learning algorithms, including feature extraction, over-fitting, and scalability [57]. Deep learning automates the process of feature extraction, enabling the system to automatically extract new characteristics from raw data. Figure 3.1 provides an excellent illustration of this idea.

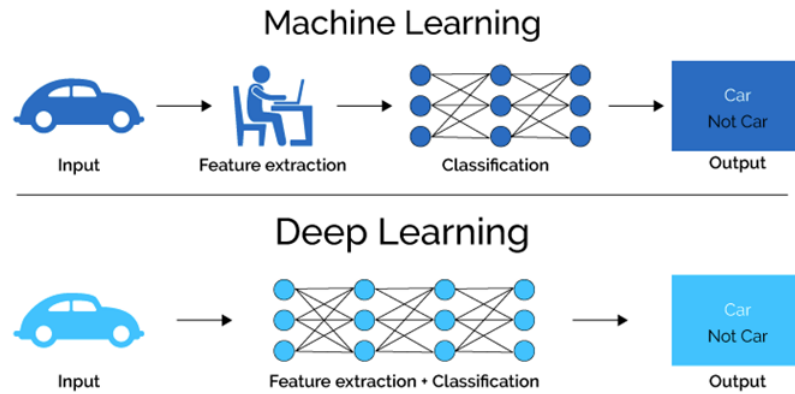


Figure 3.1: Machine learning vs. deep learning feature extraction.

### 3.3 Artificial neural network

An artificial neuronal network (ANN) is a computer model that draws inspiration from the neuronal structure of the human brain. The system is composed of linked nodes, also known as artificial neurons, which are arranged in layers. Input is sent across these nodes, and the network adapts the connection strengths (weights) through training in order to acquire knowledge from the input. This enables the network to identify patterns, make predictions, and successfully complete a range of tasks in the fields of machine learning and artificial intelligence. The structure may vary from a single layer to several layers of linked nodes (neurons) arranged in a hierarchical manner [58].

The perceptron is regarded as the most basic kind of feed-forward neural network, including a single layer of input nodes and one output node. It has the ability to acquire knowledge and make choices depending on input characteristics. The perceptron's learning process enables it to adapt its weights by analysing the input data, hence facilitating predictions or conclusions about the input [59].

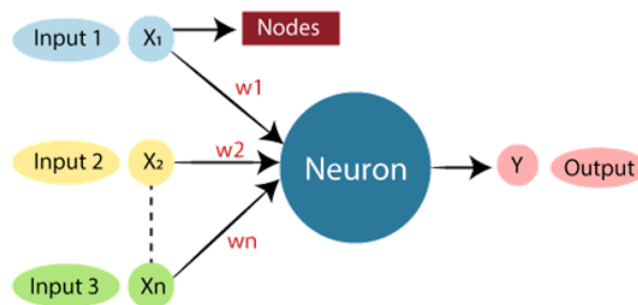


Figure 3.2: The structure of a perceptron.

### 3.3.1 Architecture of an artificial neural network

The structure of an artificial neural network has three primary kinds of layers: the input layer, the hidden layer (which may include many layers), and the output layer. The layers are organised in a certain order, and a standard feed-forward network handles information in a unidirectional manner, moving from input to output. The architecture is often known as the multi-layer perceptron (MLP) [60].

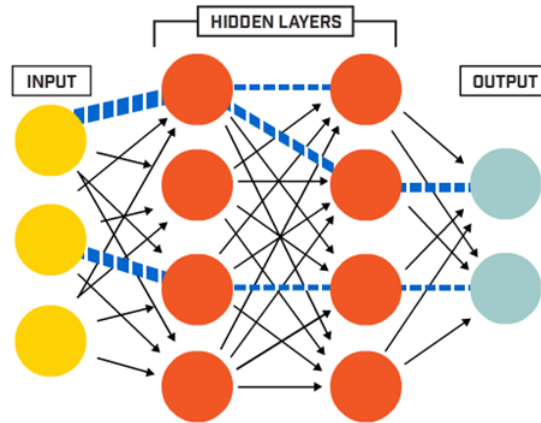


Figure 3.3: Architecture of an ANN.

- **Input layer:** It is the first layer of the neural network. The system is composed of input nodes, with each node representing an input parameter for the model. The quantity of input nodes is established by the quantity of features or input variables in the dataset. Every input node receives a feature as input and transmits it to the neurons in the hidden layer.
- **Hidden layer:** The hidden layer is responsible for the processing of incoming data. The neural network is composed of many layers of interconnected neurons, with each neuron being linked to every neuron in the preceding and succeeding layers. The hidden layers have the task of collecting characteristics from the input data and acquiring intricate patterns and correlations within the data. The quantity of hidden layers and the quantity of neurons in each hidden layer are crucial variables in ascertaining the network's capacity to acquire knowledge and extrapolate from the input.
- **Output layer:** It represents the last layer of the neural network. The network generates the output by using the input data and the acquired patterns. The quantity of nodes in the output layer is depending upon the nature of the issue that the network is



intended to address. In a classification assignment involving several classes, the number of output nodes would correspond to the number of classes, and the output would provide the anticipated probabilities for each class.

### 3.3.2 Activation functions

An activation function is a mathematical function that transforms the input of a neuron and generates an output, which is then transferred via hidden layers. Every neuron has a weight, and the product of the input number and the weight yields the output of the neuron, which is then sent to the next layer [61]. Subsequently, it may help in standardising the output of every neuron to a range of either 1 to 0 or -1 to 1 etc. (depending upon the function).

Activation functions play a crucial role in neural networks by introducing nonlinearity, which enables the network to learn complex patterns and carry out nonlinear calculations. They are mainly classified into three types: binary step function, linear activation function, and nonlinear activation function.

1. **The binary step function** is an activation function that is based on a threshold (Figure 3.4). That means whether it exceeds or falls below a certain threshold. If the output of the neuron exceeds the threshold, the activation neuron will transmit an identical signal to the following layer, and vice versa.

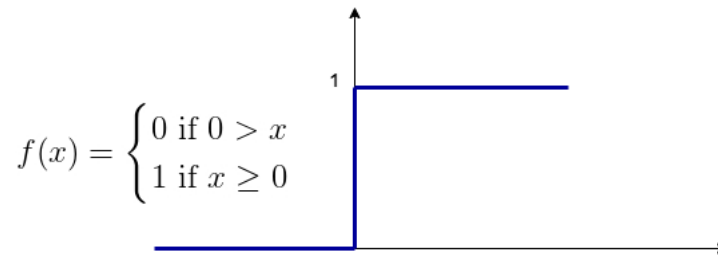


Figure 3.4: A binary activation function.

However, as it is still binary, it is not suitable for solving multi-classification issues. The trigger function serves just as a stimulus and does not have the ability to modify any of the data received from the preceding layer [62].

2. **The linear activation function** is a function that directly outputs the input without any alteration or modification (Figure 3.5). Regardless of its simplicity to implement and calculate, it has significant limitations. It lacks the capability to add nonlinearity

into the neural network, hence restricting its learning capacity. Additionally, it suffers from the issue of gradient saturation, when the gradient diminishes to zero or an extremely tiny value, resulting in a slow or inefficient learning process [62].

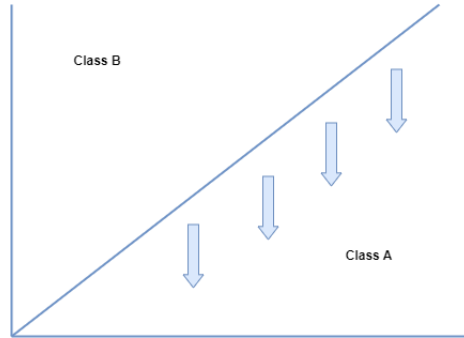


Figure 3.5: A linear activation function.

**3. Nonlinear activation functions:** Nowadays, many neural network models employ nonlinear activation functions. Complicated mappings between inputs and outputs are crucial for learning and modelling complicated data, including images, video, audio, and nonlinear or high-dimensional datasets. They address issues with linear activation functions [61]. The current neural network models use several forms of nonlinear activation functions, including sigmoid/logistic, tanh/hyperbolic tangent, ReLU (rectified linear unit), leaky ReLU, softmax, and swish.

**a) Sigmoid or logistic activation function:** A non-linear function known as a sigmoid activation function uses a smooth S-shaped curve (Figure 3.6) to convert the input into a value between 0 and 1. Take  $f(x) = 1/(1 + \exp(-x))$  as a mathematical representation. Binary classification issues benefit from the output of a sigmoid activation function, which provides a probability value. Nevertheless, there are a few downsides as well. The learning process becomes slow or useless due to the gradient vanishing issue, which occurs when the gradient becomes very tiny for big positive or negative inputs. The output is always positive since it is not zero-centred, which might lead to irregular gradient updates.

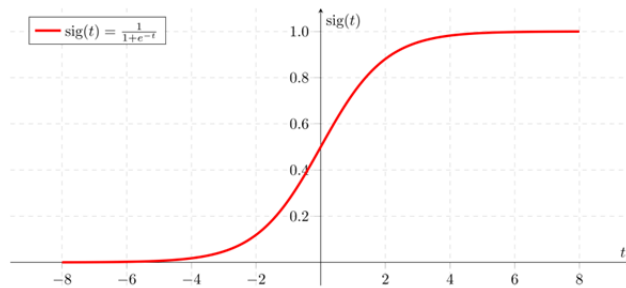


Figure 3.6: Sigmoid activation function curve.

- b) **Tanh function:** It is a Hyperbolic Tangent similar to the sigmoid function, however it exhibits symmetry around the origin (Figure 3.7). This leads to varying signs of outputs from the preceding layers, which will be used as input for the next layer. The function may be described as follows:  $f(x) = 2\text{sigmoid}(2x) - 1$ . Tanh is both continuous and differentiable, with its values ranging from -1 to 1. The gradient of the tanh function is steeper than that of the sigmoid function. The Tanh function is favoured over the sigmoid function because to its unrestricted gradients, which can vary in any direction, and its zero-centred nature.

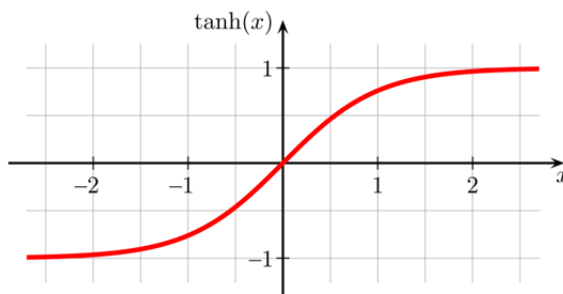


Figure 3.7: Tanh activation function curve.

- c) **Rectified linear unit function (ReLU):** The rectified linear unit function, often known as ReLU, is a non-linear function (see Figure 3.8). The ReLU function's derivative is dependent on  $x$ , making it suitable for back-propagation. The function is defined as  $f(x) = \max(0, x)$ . This formulation of the function implies that not all neurons are simultaneously activated; resulting in improved computing efficiency compared to the previous activation functions [59]. The ReLU function is widely used in neural networks, especially in convolution neural networks (CNNs), because to its several advantages. It is simple to implement, since it does not need any complex mathematical procedures. Additionally, it is not affected by the issue

of gradient vanishing, since the gradient is limited to either 0 or 1. This characteristic may enhance the speed and effectiveness of the learning process. Additionally, it induces sparsity within the neural network by deactivating some neurons, hence potentially mitigating over-fitting and enhancing generalisation. Nevertheless, it also has some disadvantages. The issue of gradient saturation arises, when the gradient diminishes to zero for negative inputs, resulting in a slow or inefficient learning process. Additionally, it is not centred on zero, resulting in the output always being positive. This might lead to irregular gradient updates.

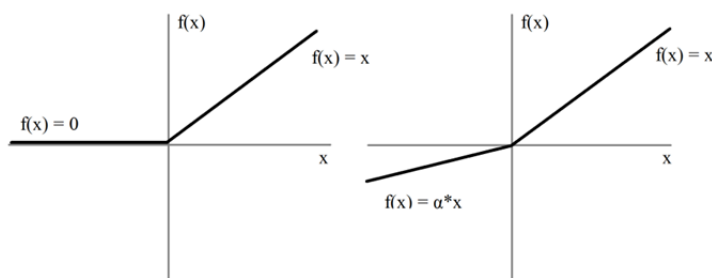


Figure 3.8: Illustration of ReLU and Leaky ReLU activation functions.

- d) **Leaky ReLU function:** Leaky ReLU is an enhanced variant of the ReLU function. In Leaky ReLU, instead of setting the value of the function to zero for negative values of  $x$ , it is specified as an extremely small linear element in  $x$ . The mathematical expression for this function is as follow:  $f(x) = x, for x \geq 0, f(x) = ax, for x < 0$
- e) **Softmax activation function:** The softmax activation function is a non-linear function that transforms the input into a probability distribution over a given set of outcomes, using an exponential function. For instance, consider the function  $f(x) = exp(x)/sum(exp(x))$ , where the sum is taken across all possible outcomes. The softmax activation function can be useful for multi-class classification tasks since it may provide a probability value for each class. Furthermore, it has a distinct explication as the probability of a neuron being classified into a certain category [59]. Nevertheless, it also has some disadvantages. The computational cost is high due to the use of an exponential function and a normalizing component. Additionally, it suffers from the issue of gradient vanishing. These are some of the most common and widely used activation functions, but there are many others that can be utilised in neural networks, such as ELU, SELU, Swish, Mish, and more. The choice of activation function depends on the specific problem, the

architecture of the neural network, and the performance and behaviour of the learning process.

### 3.3.3 Model training

Training a neural network involves adjusting the model's parameters to minimize the loss function and enhance its performance on a particular task. These parameters are functions that are depending on the values of the weight vectors and biases of each individual node. The subsequent stage is to ascertain the method of modifying those parameters in order for the artificial neural network (ANN) to accurately calculate predetermined values of the function. This involves using a sequence of input-output pairs to determine how the weight vectors and biases should be adjusted to achieve the desired outcome. A conventional approach involves defining an error function  $E$  that operates on a set of pairings  $X$ . Therefore, the process of training the ANN involves the task of minimizing the error function  $E$  in relation to the parameters.

Due to the complicated nature of the error function  $E$ , which is a combination of many non-linear functions based on the output of the ANN, it is often impossible to determine its minimum analytically. Fortunately, there exists a widely applicable technique called gradient descent that may be used to minimize differentiable functions.

#### 3.3.3.1 Gradient and back propagation

Gradient descent is a method that calculates the gradient of a function  $f$  at a certain value  $x$  (which represents the parameters of an ANN) and then adjusts that value by moving in the opposite direction of the gradient. This procedure continues until a local minimum is identified, or the gradient reaches a level of convergence that is less than a specified threshold. The learning process for an ANN usually begins by randomly initializing the parameters, which include the weight vectors and biases. These parameters are then updated iteratively using gradient descent until the error function  $E$  reaches convergence.

Back propagation is the fundamental technique used to train artificial neural networks. An algorithm is used to enable the network to modify its internal parameters, known as weights and biases, in order to decrease mistakes and enhance its performance on certain tasks [63].

- **The forward pass:** It involves the sequential propagation of information across the network, beginning at the input layer and progressing towards the output layer. At each layer, the data undergoes a process of transformation using mathematical functions that include weights and biases.

- **Error calculation:** After the output is produced, the discrepancy between the network's prediction and the intended result (ground truth) is computed. The discrepancy is referred to as the error.
- **Backward pass:** This is the stage when the most important computations take place. Subsequently, the error is sent in reverse order via the network, progressing through each layer. At each layer, the algorithm computes the individual contributions of each weight and bias to the total error. Gradient descent is a mathematical approach used to do this.
- **Weight update:** The back propagation algorithm modifies the weights and biases based on their contribution to the error, resulting in a reduction in the total error. This effectively adjusts the network's internal decision-making mechanism.
- **Iteration:** The process of the forward pass, error computation, backward pass, and weight update cycle repeated several times. During each iteration, the network constantly learns and improves its performance.

### 3.4 Convolutional neural networks (CNNs)

Deep learning has achieved impressive results in recent years across several domains, including image identification, audio recognition, and natural language processing. Convolutional neural networks (CNNs) have been the subject of substantial research among several forms of deep neural networks. The study on convolutional neural networks has rapidly evolved and obtained state-of-the-art results on numerous tasks, due to the significant increase in annotated data and the improved capabilities of graphics processing units [64].

A CNN (also known as a ConvNet) is a class of deep neural networks, primarily used for image analysis that assigns importance (in the form of weights and learnable biases) to different components of the image and discriminates between them [65]. Compared to other classification methods, ConvNets require less pre-processing. While primitive methods require handcrafted filters, ConvNets can learn these filters/features with sufficient training data.

A basic ConvNet consists of a series of layers, as depicted in Figure 3.9, where each layer converts one set of activations into another using a differentiable function. The main components of a CNN are:

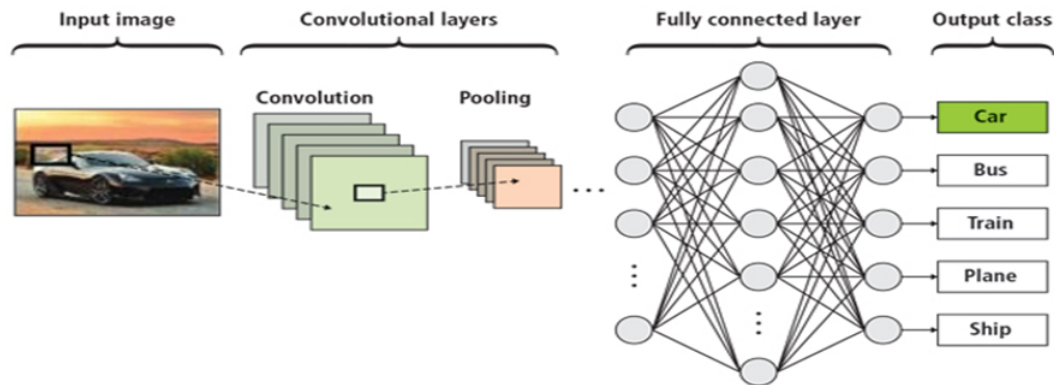


Figure 3.9: A simple CNN architecture.

### 3.4.1 Convolutional layers

A convolutional layer serves as the fundamental component of a CNN. The system includes a set of filters (also known as kernels), the parameters of which are acquired via the training process. The dimensions of the filters are often less than those of the original image. Every filter applies a convolution operation on the image, resulting in the creation of an activation map. In the process of convolution, the filter is moved across the height and width of the image, and at each point, the dot product is computed between each element of the filter and the corresponding input element [66].

Figure 3.10 is a demonstration of the convolution process. The first activation map, shown by the red marking in Figure 3.10, is obtained by convolving the filter with the corresponding blue section of the input picture. The activation map is produced by iteratively applying this technique to each element of the input picture. The output volume of the convolutional layer is obtained by concatenating the activation maps of each filter along the depth axis.

Each element of the activation map may be seen as the result of a neuron. Consequently, every neuron is linked to a small local region in the input image, with the dimensions of the region matching the dimensions of the filter. Each neuron in an activation map is characterised by a set of shared parameters. The convolutional layer's local connection requires the network to learn filters that provide the highest response to a specific area of the input. The early convolutional layers capture the low-level characteristics, such as lines, in images, whereas the subsequent layers extract the high-level information, such as forms and individual objects [67].

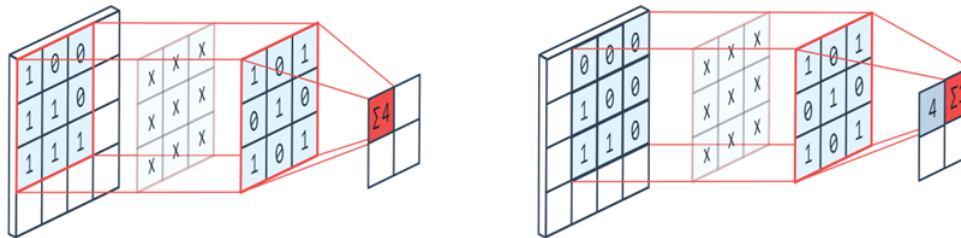


Figure 3.10: A graphical example of the convolution process.

Some of the important concepts that one should know while defining a convolutional layer are as follows:

- **Filter/Kernel:** The convolutional kernel (filter) is a small matrix that slides over input data (typically an image) to extract features. The output value is calculated by element-wise multiplying the local receptive field (a subset of the input data) and summing the results. Kernels are learnable parameters that capture input patterns like edges, textures, and shapes.
- **Feature map:** A feature map is the result of applying a filter to the input data. Each filter produces a feature map that emphasises the existence of a particular pattern or feature in the input.
- **Stride:** The stride determines how much the kernel moves horizontally and vertically during each step of convolution. A larger stride reduces the spatial dimensions of the output feature map, while a smaller stride preserves more spatial information. Common stride values are 1 (no overlap), 2 (skipping every other pixel), and 3 (skipping every two pixels).
- **Padding:** It involves adding extra pixels around the input data before applying convolution. It helps maintain spatial dimensions and prevents information loss at the edges. Common padding types include “valid” (no padding) and “same” (adding padding to keep output size the same as input).

The output volume size may be calculated using the following formula: given an input size of  $W \times W \times D$  and a  $D_n$  number of kernels with a spatial dimension of  $K$  with stride  $S$  and amount of padding  $P$ :

$$W_{out} = \frac{W - K + 2P}{S} + 1 \quad (3.1)$$

This will yield an output volume of size  $W_{out} \times W_{out} \times D_n$ .



### 3.4.1.1 Pooling layers

Pooling layers are often used in CNNs to reduce the size of the feature maps generated by the convolutional layers. Down-sampling aids in decreasing computing complexity and extracting higher-level representations [68]. There are two prevalent forms of pooling:

**Max pooling:** is a process where the largest value inside a sliding window is chosen and the other values are discarded. This technique maintains the most salient characteristics inside each window, preserving crucial information.

**Average pooling:** is a process that computes the mean value inside a sliding window. It has the potential to reduce noise and provide a more smooth representation of the data.

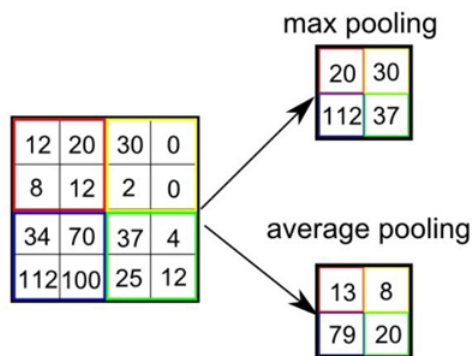


Figure 3.11: Pooling processes.

### 3.4.1.2 Fully connected layer

Fully connected (FC) layer, usually referred to as dense layers, are the conventional layers located at the terminal of neural networks. Contrary to convolutional and pooling layers which only process certain areas, fully connected layers establish connections between every neuron in the previous layer and every neuron in the current layer [69]. This is their basic role:

**Neuron activation:** In a fully connected layer, each neuron gets inputs from all neurons in the preceding layer and performs an activation function to generate an output. This output signifies the activity or responsiveness of the neuron to the inputs.

**Classification or regression:** Fully connected layers are often used for classification or regression tasks, in which the network associates the acquired characteristics with distinct output classes or numerical predictions.

**Final output:** The final predictions of the network are obtained by either feeding the output of the last fully connected layer into a softmax function for classification or leaving it as is for regression.

## 3.5 CNN architectures

The earliest CNNs were developed in the 1980s by Kunihiko Fukushima [70]. These early CNNs were known as Neocognitron and were designed to recognise handwritten characters. Neocognitron used a series of convolutional and pooling layers to learn increasingly complex features in the data. However, these early architectures were limited by the need for more data and computing power available at the time.

### 3.5.1 AlexNet

During the 1990s, Yann LeCun [71] and his colleagues created the LeNet-5 architecture, which was among the first (CNNs) to attain notable success in a practical context. LeNet-5 was specifically developed to accurately identify handwritten digits and was used by banks for the automated recognition of handwritten checks. However, CNNs continued to evolve, with researchers exploring ways to improve their performance on larger datasets. One such architecture was the AlexNet, introduced by Krizhevsky et al. [69] in 2012, marked a significant milestone in the field of deep learning and computer vision. It played a pivotal role in popularising deep CNNs and demonstrated their effectiveness in image classification tasks. As depicted in Figure 3.12, AlexNet consists of eight layers: The first five are convolutional layers, some followed by max-pooling layers. The last three are fully connected layers. The network is split into two copies, each running on a separate GPU. AlexNet's success demonstrated the power of CNNs and the importance of GPUs for deep learning. It spurred a wave of research, leading to many more papers employing CNNs and accelerating deep learning.

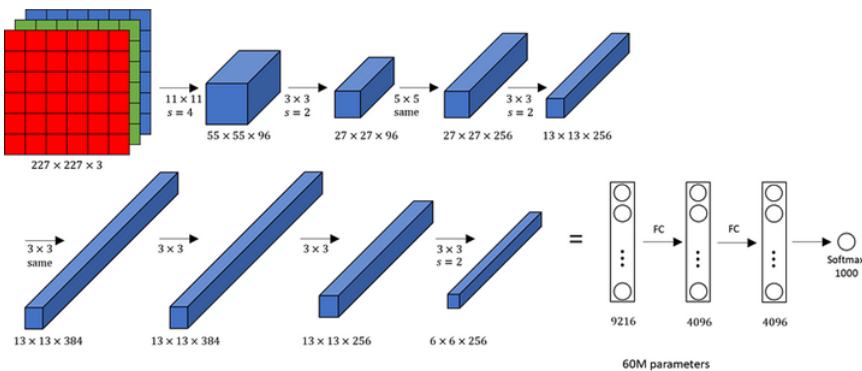


Figure 3.12: AlexNet architecture.

### 3.5.2 Visual geometry group (VGG)

The VGG network, developed by Karen Simonyan and Andrew Zisserman [72], was unveiled in 2014. During that time, the network was regarded as very complex and intricate. The primary significance of this study was to demonstrate that the depth of the network is a crucial factor in improving the accuracy of recognition or classification in CNNs. The VGGNet, as seen in Figure 3.13, utilises  $3 \times 3$  filters. The authors provide the rationale that using two consecutive  $3 \times 3$  filters results in an effective receptive field of  $5 \times 5$ , whereas three consecutive  $3 \times 3$  filters provide a receptive field of  $7 \times 7$ . The quantity of filters in the architecture increases twofold after each max-pooling process.

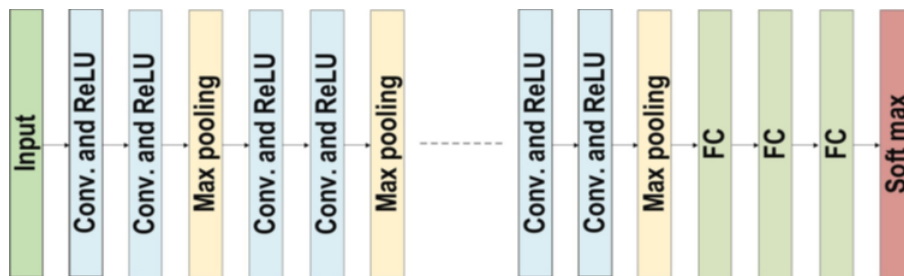


Figure 3.13: VGG architecture.

### 3.5.3 GoogleNet

The network that emerged as the winner of ILSVRC competition in 2014 is GoogleNet (also known as Inception-V1) [73]. GoogleNet is recognised as the pioneering implementation of modern CNN architectures. Unlike traditional CNNs that just rely on successive convolution and pooling layers, GoogleNet introduced the concept of inception architecture, which

incorporates a network inside a network (NIN) approach [74]. The inception module creates a mini-module inside the network by skipping connections, and this module is then reproduced across the network (see Figure 3.14). The inception module significantly decreased the amount of parameters in the network. GoogleNet employs 9 inception modules and replaces all fully connected layers with average pooling, reducing the dimensions from  $7 \times 7 \times 1024$  to  $1 \times 1 \times 1024$ . This reduces a significant amount of parameters that seem to have little impact. To enhance the data, various crops of the same image were generated and used to train the network. Additionally, there are also several follow up versions to the GoogleNet, most recently Inception-v4.

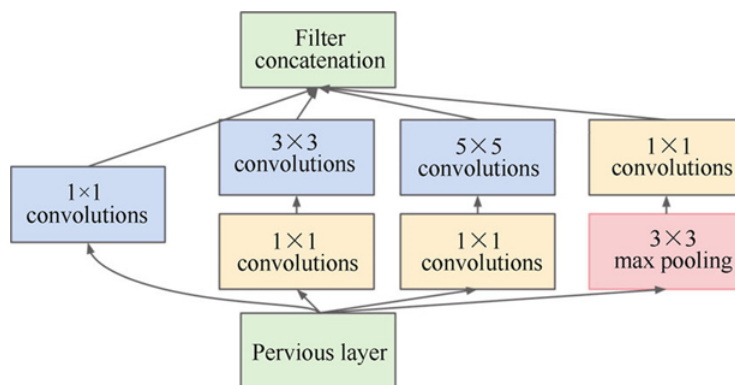


Figure 3.14: Inception module of GoogleNet.

### 3.5.4 ResNet

ResNet (Residual Network), designed by He et al. [75], emerged as the champion of ILSVRC 2015. Their goal was to create a deep neural network that does not suffer from the problem of vanishing gradients, unlike earlier networks. Various types of ResNet were created, each distinguished by the number of layers used. The predominant kind was ResNet-50, consisting of 50 convolutional layers with total number of network weights was 25.5 million. As shown in Figure 3.15, the basic concept of a ResNet is an “identity shortcut connection” that skips more than one layer. A residual block, which is repeated throughout the network, is the basic building block of a ResNet. Therefore, the output can be expressed as follows:  $H(x) = F(x) + x$ , where  $F(x)$  is the residual function. The weight layers are used to learn a kind of residual mapping, which is expressed as follows:  $F(x) = H(x) - x$ . Even if the gradient of the weight layers disappears, we still have the identity  $x$  to transfer to the previous layers. The residual function  $F(x)$  can contain either two layers, as in ResNet-18 and ResNet-34, or three layers for deeper networks (i.e., ResNet-50 and above). These two main block types

form the basis of ResNet.

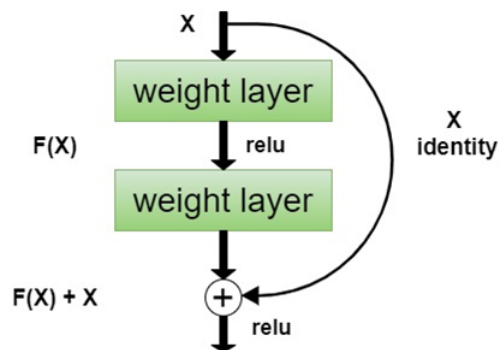


Figure 3.15: A residual block.

ResNet introduced shortcut connections inside layers to facilitate cross-layer interconnection, unlike the highway network. These connections are both parameter-free and data-independent. It is important to understand that the layers represent non-residual functions only when a gated shortcut is blocked in the highway network. In contrast, the individualistic shortcuts in ResNet remain open indefinitely, allowing remaining information to be continuously sent. Moreover, ResNet has the capability to mitigate the issue of gradient vanishing by using shortcut connections (residual links) that expedite the convergence of the deep network. ResNet emerged as the winner in the 2015-ILSVRC competition, with a remarkable depth of 152 layers. This is eight times deeper than VGG and twenty times deeper than AlexNet. Compared to VGG, it has a reduced computational cost, even when the depth is increased.

### 3.5.5 Inception

Inception-V3, V4, and Inception-ResNet are enhanced versions of Inception V1 and V2, as stated by Szegedy et al. [73, 76, 77]. The objective of Inception-V3 was to decrease the computational expense of deep networks while maintaining their generalization capabilities. In order to achieve this objective, Szegedy et al. [77] substituted the usage of big size filters ( $5 \times 5$  and  $7 \times 7$ ) with smaller and asymmetric filters ( $1 \times 7$  and  $1 \times 5$ ). Additionally, they used a  $1 \times 1$  convolution as a bottleneck before to the large filters. The simultaneous positioning of a  $1 \times 1$  convolution with a big size filter enhances the typical convolution process, making it resemble a cross-channel correlation. In a recent study, Lin et al. [74] examined the capabilities of  $1 \times 1$  filters in the NIN architecture. Szegedy et al. [77] smartly used the same notion. The Inception-V3 model employs a  $1 \times 1$  convolutional operation to

transform the input data into smaller spaces, often 3 or 4 in number. These smaller spaces are then processed using regular convolutions ( $3 \times 3$  or  $5 \times 5$ ) to capture correlations. In the Inception-ResNet model, the authors integrated the benefits of residual learning with the inception block, as described by He et al. [75] and Szegedy et al. [73]. The process included replacing filter concatenation with the residual connection. Furthermore, the authors conducted an experimental study that demonstrated that Inception-V4 with residual connections (Inception-ResNet) had equivalent generalization capabilities as plain Inception-V4, but with greater width and depth. However, it was shown that Inception-ResNet achieves convergence at less time compared to Inception-V4. This indicates that including residual connections in the training process significantly speeds up the training of Inception networks. Inception Residual unit's fundamental block diagram is shown in Figure 3.16.

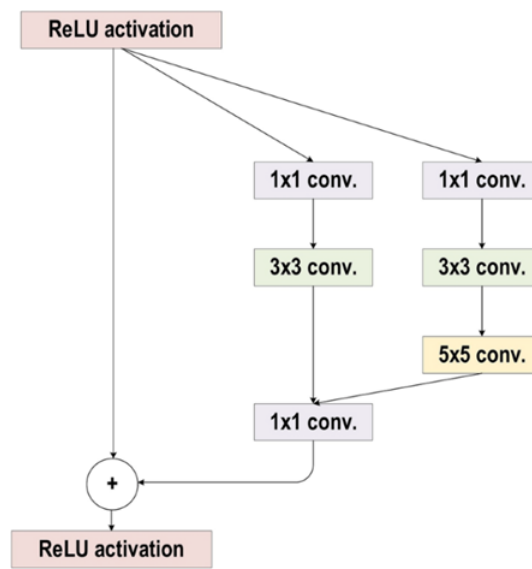


Figure 3.16: A basic schematic for the Inception Residual unit.

### 3.5.6 DenseNet

DenseNet was designed to overcome the vanishing gradient issue like Highway Networks and ResNet. ResNet expressly protects information via additive identity transformations; therefore many layers may provide little or no information. DenseNet modified cross-layer connection to overcome this issue [75, 78]. DenseNet feed-forwards feature-maps from previous layers to the next layer, enabling cross-layer connectivity by concatenating layer information before assigning it to a new transformation layer. This creates direct connections in DenseNet, unlike typical CNNs [79].

Since DenseNet concatenates characteristics from the preceding layer instead of adding

them, the network may be able to distinguish between added and conserved information. With more feature-maps, DenseNet becomes parametrically costly despite its small layer structure. Improving information flow in the network involves giving each layer direct access to gradients via the loss function. Direct gradient admission prevents over-fitting on tasks with smaller training sets. DenseNet network's block is shown in Figure 3.17.

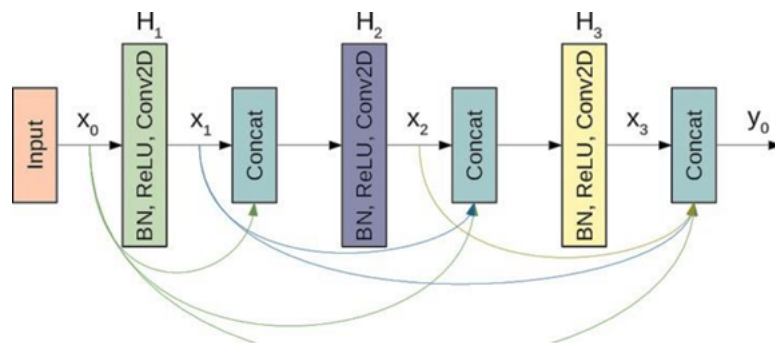


Figure 3.17: A schematic illustration of a 3-layer dense block used in the DenseNet.

### 3.5.7 ResNeXt

ResNeXt, also referred to Aggregated Residual Transform Network, is an enhanced version of the Inception Network. Xie et al. [80] effectively used the split, transform, and combine approach by introducing the word "cardinality" in a simple and impactful manner. Cardinality is a supplementary aspect that pertains to the magnitude of the collection. The Inception network has enhanced both the learning capacity of traditional CNNs and the efficiency of network resources. However, since several spatial embedding's used (such as  $3 \times 3$ ,  $5 \times 5$ , and  $1 \times 1$  filters) in the transformation branch, each layer must be designed individually.

ResNeXt used VGG's deep homogeneous topology and standardized GoogleNet architecture by limiting spatial resolution to  $3 \times 3$  filters in split, transform, and merge blocks. It improved deep and wide network convergence with residual learning. Figure 3.18 illustrates the ResNeXt's building block. To reduce ResNeXt's complexity, low embedding filters ( $1 \times 1$ ) were used before  $3 \times 3$  convolution, whereas skip connections enhanced training [81].

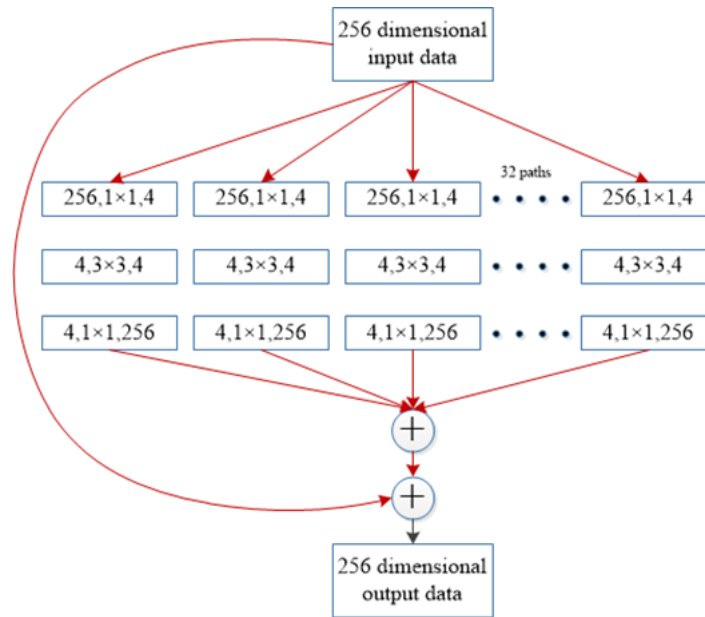


Figure 3.18: ResNeXt building block.

### 3.5.8 Wide ResNet

The main drawback of deep residual networks is the feature reuse issue, where some feature changes or blocks may contribute little to the learning process. The issue was resolved by Wide ResNet [82] as proposed by Zagoruyko and Komodakis. The main learning capacity of deep residual networks comes from the residual units, while the depth of the network has an additional effect. The Wide ResNet approach maximized the potential of residual blocks by increasing the width of the ResNet architecture instead of its depth. Wide ResNet augmented the width by including an extra parameter, which controls the width of the network. This yield effectively to improve the performance. Authors found that although Wide ResNet has twice the number of parameters than ResNet, it can train more effectively than deep networks.

### 3.5.9 Recent advancements in convolutional neural networks

This section provides a comprehensive examination of the latest advancements and discoveries in CNNs throughout the last few years. Similarly, these improvements exemplify the cutting-edge of CNN technology, demonstrating their continuous growth and capacity to handle complex tasks.

**Attention mechanisms:** They have emerged as a crucial notion in CNNs), enabling



networks to focus on certain regions of input data, hence enhancing their capacity to acquire relevant information. These systems have been crucial in performing many tasks, including picture captioning, language translation, and image segmentation.

***Capsule networks:*** They are also known as CapsNets; they provide a novel approach of acquiring features. They excel at collecting hierarchical connections between features, which make them very successful in tasks involving picture recognition. Capsule networks exhibit resilience to fluctuations in object views and provide a more profound comprehension of characteristics.

***Self-supervised learning:*** It signifies a fundamental change in the way convolutional neural networks are trained. Instead of largely depending on labelled data, this strategy employs unlabelled data for training the model. The cost-effectiveness and data efficiency of this technology have made it widely used in the recent years, with its applications ranging from image analysis to natural language processing.

## 3.6 Deep learning applications

Currently, a broad range of deep learning applications are prevalent worldwide. These applications include several fields such as healthcare, social network analysis, audio and voice processing (including detection and enhancement), visual data processing techniques (such as multimedia data analysis and computer vision), and NLP (including translation and sentence classification) [83], among others are illustrated in Figure 3.19.

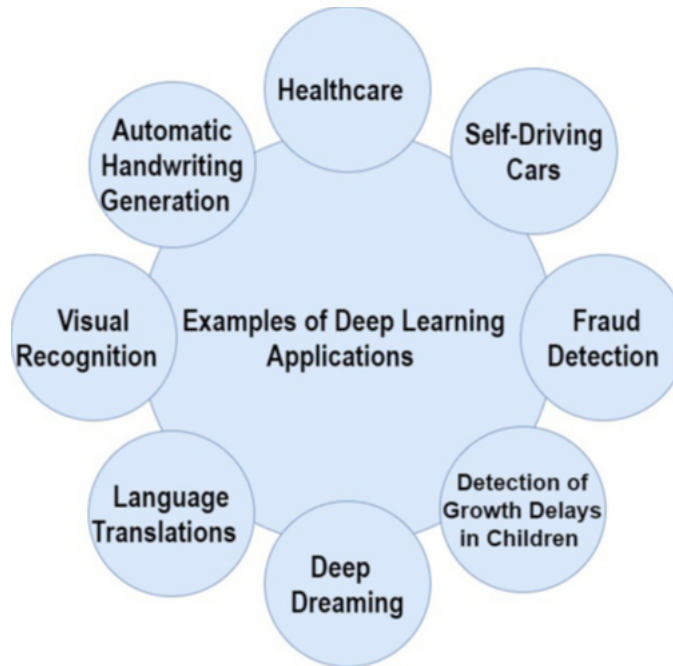


Figure 3.19: Examples of DL applications.

### 3.7 Challenges in deep learning

Up until now, progress in deep learning has been made by studying architectures that have been empirically confirmed. However, there has been a lack of focus on understanding the causes and mechanisms underlying how these deep learning models are able to create such extraordinary outcomes. The choice of structural characteristics and optimal adjustment of model hyper-parameters is now a subject of on-going discussion. Although several computing units have been constructed based on their distinct mathematical capabilities, the present study on their selection primarily relies on experimental methods [84]. Establishing a robust theoretical framework is essential for refining and assessing these networks using real-world data.

Addressing the difficulty of developing regularisation approaches, together with establishing theoretically supported tuning and performance evaluation procedures for DL models using empirical data, is crucial. The assessment of unsupervised learning requires a specialized technique to address the significant challenge posed by structure-based deep learning networks that dynamically change or adapt [83].

Deep learning models also encounter dataset diversity as a challenge. Datasets found on the Internet may be categorized into three types: semi-structured, structured, or unstruc-

tured. This results in a significant level of diversity and inconsistency in terms of data types and formats. Data combining is a feasible approach to address diversity, and it is anticipated to provide superior deep learning models compared to using single-modality data [84]. An additionally challenge lies in the multitude of dimensions that big data includes. The International Data Corporation (IDC) predicted that the global data sphere will expand to 175 zettabytes (ZB) [85].

Other notable challenges comprise inherent trends, theoretical comprehension, vision at a level comparable to humans, training with restricted data, temporal complexity, and more potent models. Despite the encouraging results achieved by deep learning approaches in computer vision applications, the theoretical foundations are still not well comprehended, and there is a lack of clarity regarding which architectures are expected to yield superior performance. It is challenging to ascertain the optimal structure, number of layers, and number of nodes per layer required for a certain application [86]. Additional information on hyper-parameters such as learning rate, decay rate, batch size, and optimizer is necessary.

Over-fitting problem must be addressed. Three separate types of deep learning techniques are identified to address the over-fitting problem. The first class includes popular approaches including weight decay, batch normalization, and dropout, which impact model architecture and parameters [86, 87]. The second class addresses model inputs, such as data corruption and augmentation [83]. A lack of training data might lead to over-fitting, causing the learnt distribution to vary from the true one. The phrase "data augmentation" refers to expanding training data. Compared to marginalized data corruption, data augmentation offers distinct solutions. The last class covers model output. A new technique penalizes excessive trust in model regularization [88]. This approach can regularize RNNs and CNNs.

## **3.8 Conclusion**

This chapter focused on deep learning, a prominent subfield of artificial intelligence that is now gaining momentum. Deep learning is a subset of machine learning that relies on artificial neural networks. In recent years, the field of deep learning has undergone significant expansion and advancement, driven by improvements in computer capacity, increased data accessibility, and creative algorithmic techniques.

Artificial neural networks serve as the foundation of deep learning, and their capacity to acquire knowledge from data has facilitated significant advancements in several domains. CNNs that specialize in image and video recognition have achieved success in a range of

applications, such as autonomous cars, facial recognition, and medical image analysis. CNN architectures have seen fast evolution, with each new model specifically targeting the limits and bottlenecks of its predecessors. CNNs have experienced a notable improvement in their ability to effectively and precisely process intricate visual information.

Deep learning has found widespread applications in several fields such as healthcare, finance, robots, military, and more. Deep learning has the capacity to transform several fields, including drug discovery, fraud detection, and autonomous navigation, by creating systems that are more efficient and effective. Despite the significant advancements in deep learning, there are still many challenges that persist. An essential issue is the requirement for an adequate amount of annotated data to effectively train the models. Data privacy and security are major considerations in several applications. Deep learning models are frequently regarded as black boxes, which poses difficulties in comprehending their decision-making process and the rationale behind it.

## Chapter 4

# Ear Recognition using Mean-Class activation Maps and Convolutional Neural Networks

## 4.1 Introduction

Ear recognition is a challenging research topic in biometrics, with the primary goal being the efficient identification of individuals from images of their ears under uncontrolled conditions. Convolutional neural networks (CNNs) have shown exemplary performance in several real-world applications. However, the performance of deep ear recognition systems is still immature due to the lack of large-scale datasets and it is affected by the presence of noise, blur, or occluded regions. In addition, it is difficult to determine how CNNs learn and which parts of the image are most effective for activating a desired class. To address the above problems, this chapter proposes a practical approach called Mean-class activation maps with CNNs, or simply Mean-CAM-CNN. The Mean-CAM forces the CNN to focus attention on the relevant information: it extracts and considers only the discriminative regions of the entire image. Specifically, a guided mask is first proposed to crop the relevant region of the image by focusing on the Mean heat maps. The cropped region is then used to train a CNN and perform discriminative classification.

The rest of the chapter is organised as follows: Section 4.2 describes the materials and methods employed in this work. Section 4.3 presents the experiments analysis and a comparison with the state-of-the-art. Last, in section 4.4, we conclude this chapter.

## 4.2 Materials and methods

In this section, we first describe the materials and methods employed in this work and then present the principle of our proposed method in detail.

### 4.2.1 ResNet-50

Considering the excellent performance of ResNet-50 in pattern recognition, image segmentation, and object detection, we decided to use it as the backbone network of our method. ResNet-50 is a 50-layers deep trained model on millions of images offered from the ImageNet collection [89]. The model includes approximately 23 million trainable parameters, demonstrating a sophisticated design advancing picture identification. ResNet-50's network structure is illustrated in Table 4.1. It consists of five modules: the initial one is a convolution realized with 64 different kernels, and each kernel has a size of  $7 \times 7$  with a stride size of 2. The remaining four modules are convolution and identity block. Each block has three convolution layers. The final average pool of ResNet-50 converts each feature map into a

Table 4.1: ResNet-50 structural parameters.

<b>Name of the layer</b>	<b>Size of the output</b>	<b>50-layer</b>
Conv1	$112 \times 112$	$7 \times 7 \times 64$ , stride 2
		$3 \times 3 \times$ max pool, stride 2
Conv2-x	$56 \times 56$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
Conv3-x	$28 \times 28$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
Conv4-x	$14 \times 14$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
Conv5-x	$7 \times 7$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
Avg-Pool	$1 \times 1$	Average pooling
FC	1000	Fully Connected, Softmax

single feature. Thus, the pooled field’s size equals the feature map’s size.

## 4.2.2 Class activation map (CAM)

The interpretation of neural network decisions is an active research topic and a critical notion to grasp. Because neural networks are employed in the real world, we cannot approach them as black boxes; we must discover what they interpret, how they solve, and what information each layer/channel in a neural network has learned. In the paper presented by Zhou et al. [90], the authors proposed the class activation mapping (CAM) technique, which indicates the discriminative image regions that can be used to interpret what neural networks are looking for in images while making a decision, as well as how it can help us better understand neural network decisions. Figure 4.1 illustrates the CAMs of several pictures of different classes. The

network consists of many convolutional layers, with global average pooling (GAP) applied immediately before the final output layer. The acquired features are fed to an FC layer with softmax activation, which generates the required output. The significance of picture regions can be assessed by projecting the output layer’s weights onto the convolutional feature maps produced from the previous layer.

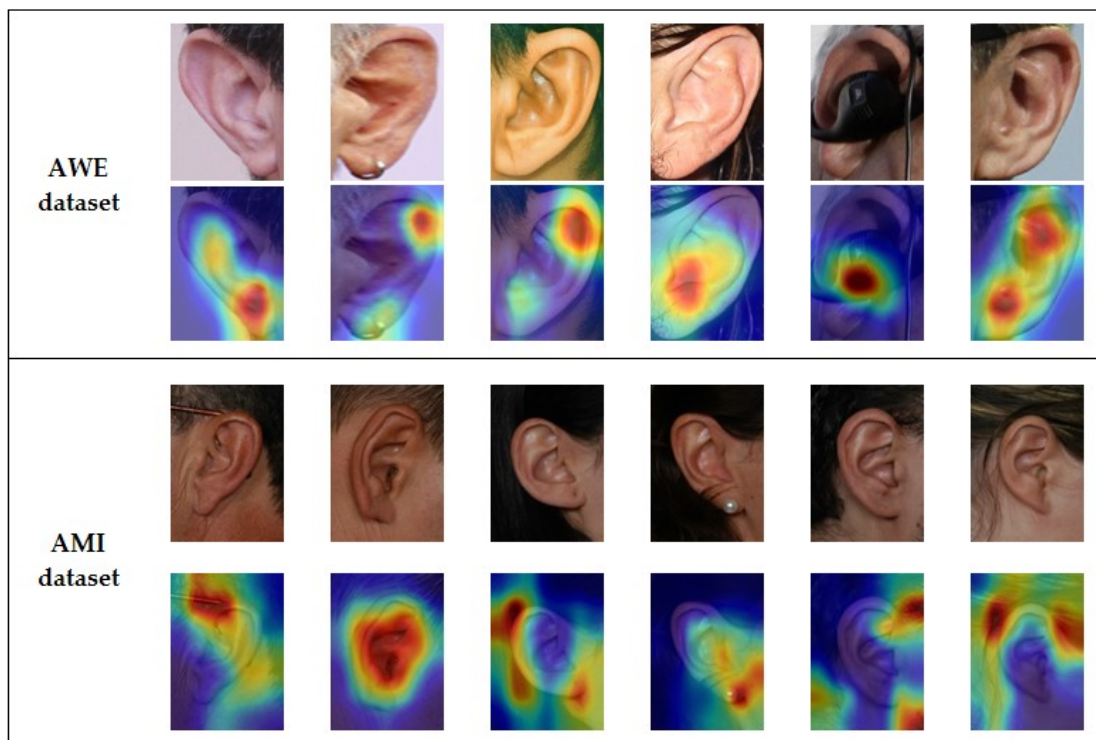


Figure 4.1: CAMs of several images of different classes from AWE and AMI data sets. The maps indicate the relevant image areas employed for image classification.

As illustrated in Figure 4.2, the GAP is recommended over the global max-pooling (GMP) layer because GAP layers help identify the object’s entire extent. In contrast, GMP layers identify just one discriminative part. In GAP, we take average overall activation, which aids in locating all discriminating areas, while the GMP layer analyses the most discriminative one. The GAP layer produces the spatial average of the features map of each unit in the final convolutional layer, which is then weighted and summed to give the final output. Similarly, we calculate the weighted sum of the last convolutional layer to generate our class activation map. The mathematical equations that define CAM are as follows:

Let  $f_k(x, y)$  be the activation map of unit  $k$  in the latest convolutional layer at spatial location  $(x, y)$ . The outcome of GAP, symbolized by  $F_k$ , is represented as:



$$F_k = \frac{1}{(x \times y)} \sum_{i=1}^x \sum_{j=1}^y f_k(x, y) \quad (4.1)$$

For a class  $c$ , an input  $S_c$  to the softmax will be:

$$S_c = \sum_{k=1}^K W_k(c) F_k \quad (4.2)$$

where  $W_k(c)$  is the weight related to a class  $c$  for unit  $k$ , specifying the  $F_k$  significance for a class  $c$ , and  $K$  represents the number of activation maps in the latest convolutional layer.

The output  $P_c$  of the softmax layer is:

$$P_c = \frac{\exp(S_c)}{\sum_{k=1}^K \exp(S_c)} \quad (4.3)$$

where  $c \in \{1, 2, \dots, C\}$ ,  $C$  refers to the total number of classes, and  $S_c$  indicates the input vector's elements to the softmax function.

Therefore, the final equation  $H_c(x, y)$  for an activation map corresponding to class  $c$  would be:

$$H_c(x, y) = \sum_{k=1}^K W_k(c) f_k(x, y) \quad (4.4)$$

where,  $H_c(x, y)$  highlights the significance of the activation at spatial grid  $(x, y)$ , leading to classifying an image as a class  $c$ ;  $k \in \{1, 2, \dots, K\}$ ,  $K$  represents the number of activation maps in the latest convolutional layer,  $W_k(c)$  is the weight related to a class  $c$  for unit  $k$ , and  $f_k(x, y)$  is the activation map of unit  $k$ .

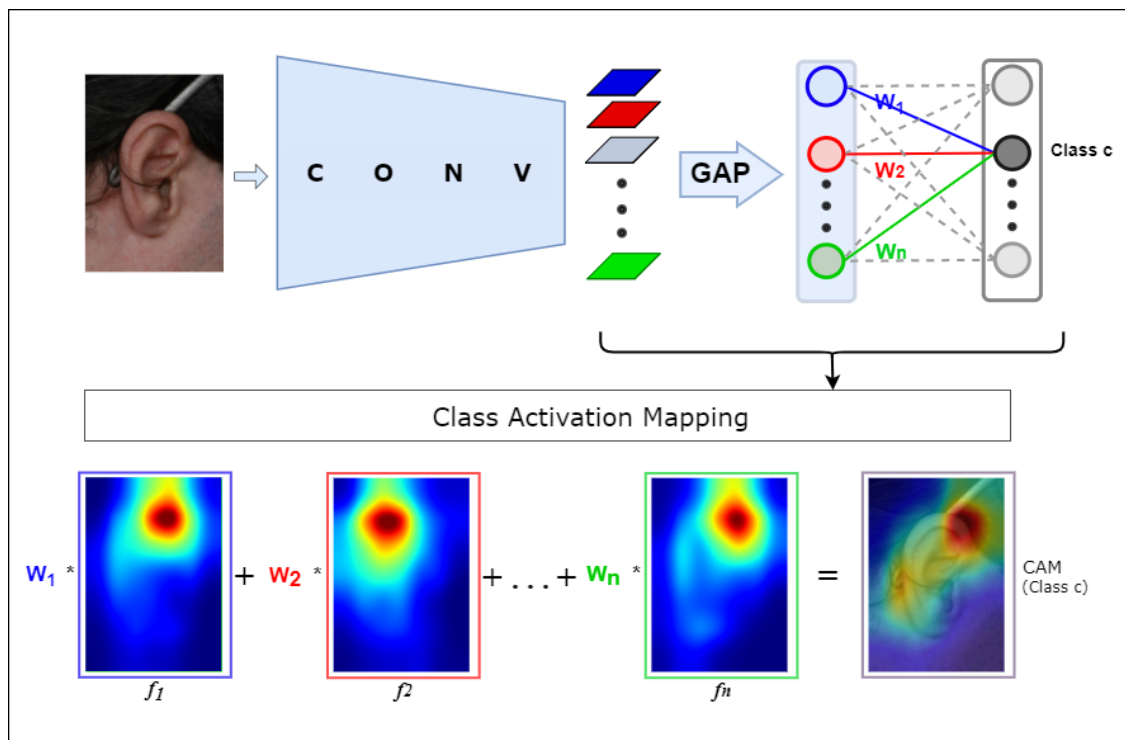


Figure 4.2: Class Activation Mapping: The anticipated class score is projected back to the preceding convolutional layer (CAM) to construct the class activation maps. CAM emphasises the relevant areas that are distinctive to each class.

There is one more reason why we have chosen ResNet-50 for our work since its top layer is a GAP layer, followed by a fully-connected layer with a softmax activation function; this is basically what the CAM technique requires [90], thus no more adjustments to its architecture are needed.

### 4.2.3 Data augmentation

CNN models trained on limited training data are prone to "overfitting". Overfitting occurs when a CNN model performs well on training data but not with new data [91]. There are various strategies to avoid overfitting, including tuning the model's architecture and tweaking hyper-parameters. Nevertheless, in the end, providing more high-quality data to the training data set is the most effective way to combat overfitting. CNN will misclassify images if it does not access a significant and varied training set for image classification. CNN can be taught to better recognise things in the real world by exposing it to images taken from various perspectives and lighting conditions. However, acquiring more training examples is difficult, costly, or impossible. This task becomes considerably more challenging

in supervised learning systems, where human experts classify training samples. Adding new information to an existing data set may provide more training opportunities by making tiny changes to existing data. It is referred to as "data augmentation" [92]. There are three main approaches for data augmentation: (1) deep learning, (2) feature transformation, and (3) basic image manipulation. The first two approaches provide data augmentation depending on the data set's feature space. The last approach, extensively used, conducts augmentation directly in the input space, such as geometry transformations, colour transformations, and noise injection.

## 4.2.4 Proposed approach

This subsection explains the process of the proposed Mean-CAM-CNN framework for ear image classification. More specificity, Section 4.2.4.1 illustrates the architecture of the Mean-CAM-CNN network, while Sections 4.2.4.2 and 4.2.4.3 discuss the Mean-CAM and the mask inference process for discriminative area detection.

### 4.2.4.1 Mean-CAM-CNN framework

Figure 4.3 highlights the architecture of the Mean-CAM-CNN. It is made up of two branches, namely the global stage and the local stage. Two CNNs are used for the global and local stages. We feed the whole images of the same class into the first CNN to obtain their corresponding class activation maps in the global stage. The average CAM is determined from all the CAMs of the images belonging to the same class. Then, using the average CAM, we extract the important (i.e., relevant) region of each global (i.e., original) image and use it for the training/classification process at the local stage.

Figure 4.3 illustrates the overall structure of the Mean-CAM-CNN using the ResNet-50 as the backbone. Mean-CAM-CNN includes two main stages: global and local. Both stages consist of five convolutional blocks with batch normalisation and ReLU. They are then connected to a max-pooling layer, an FC layer, and a softmax layer. In contrast to the global stage, the input to the local stage consists of local images that have been clipped by the mask generated by the global stage. The input image is overlaid on the heatmap for visualisation purposes.

**Global and local stages** The global stage offers insights into the relevant information derived from the entire input image. In this work, we have used the ResNet-50 [93] as the backbone model on the global stage. It consists of five blocks, followed by a GAP layer and

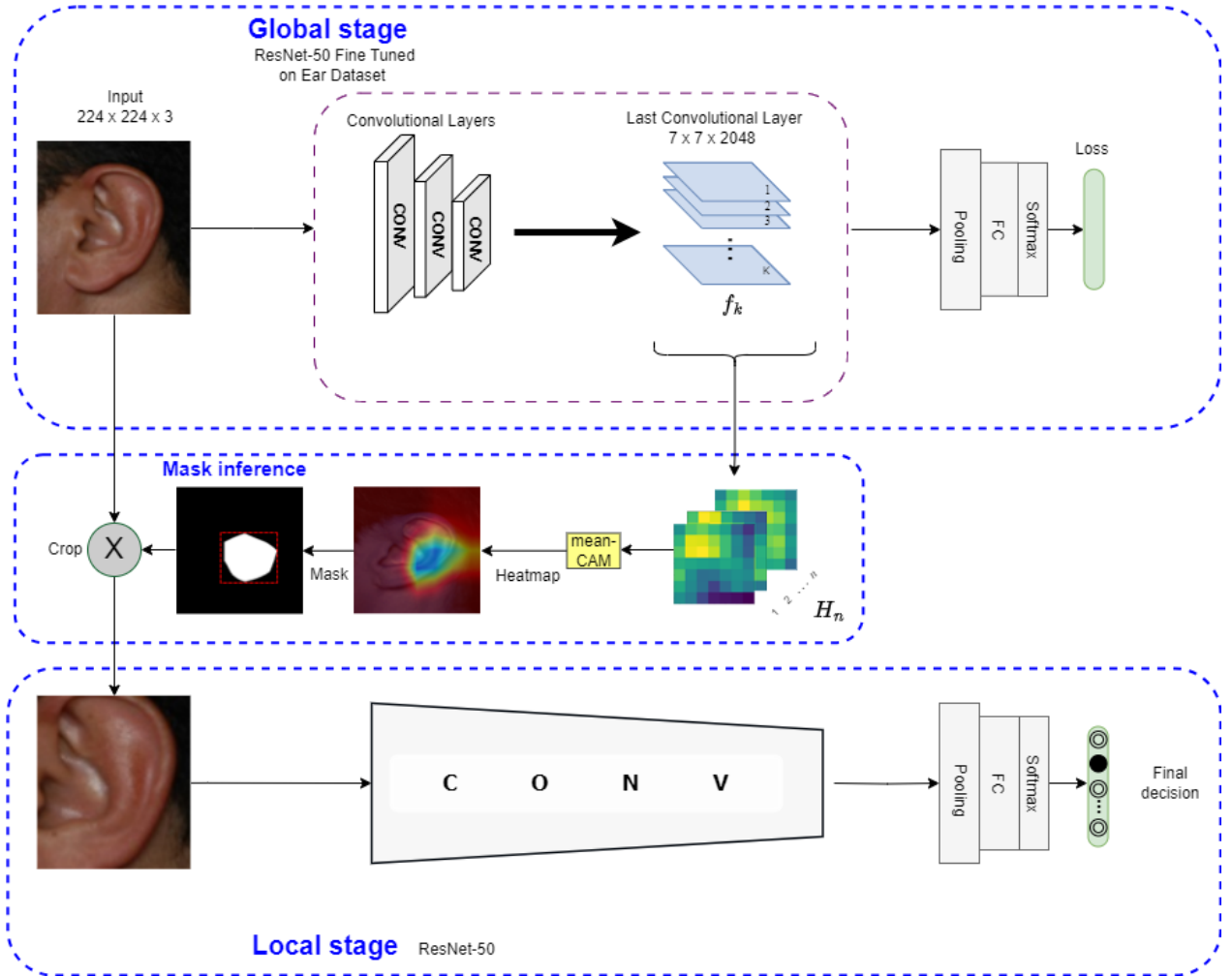


Figure 4.3: A graphical flowchart of the proposed Mean-CAM-CNN framework.

a FC layer for classification. Finally, the output vector  $S_g(c|I)$  is normalised using a softmax layer:

$$P_g(c|I) = \frac{\exp(S_g(c|I))}{\sum_{c=1}^C \exp(S_g(c|I))} \quad (4.5)$$

where  $I$  represents the global image,  $P_g(c|I)$  is the probability score of  $I$  fitting to the  $c^{th}$  class,  $c \in \{1, 2, \dots, C\}$ ,  $C$  is the total number of classes, and  $S_g$  is the input vector elements to the softmax function.

In contrast, the local stage focuses on the local area and deliberately mitigates some of the disadvantages of the global stage. The global and local stages have the same convolutional network structure (i.e., ResNet-50). However, they do not share weights due to their different functions. The probability score of a local stage is calculated using equation (4.5), expressed

as  $P_l(c|I_{crop})$ . In this case,  $I_{crop}$  refers to the input image of the local stage. As with the global stage, normalisation is performed in the same way.

#### 4.2.4.2 Mean-CAM

The proposed CNN model utilises GAP layer followed by a densely connected softmax layer. For each image  $I$ , the model can generate the final convolutional feature map  $f$  of each image:  $f = F(I)$ , where  $F$  refers to a sequence of operations that are performed by the CNN. These operations include convolution, pooling, and activation. Besides,  $f_k$  represents the  $k^{th}$  channel of the feature map, and  $f_k(x, y)$  indicates the value at spatial location  $(x, y)$ . As presented by Zhou et al. [90], the recommended CAM retrieves the class activation map and specifies the object region, as shown in Figure 4.2. Theoretically, the CAM retrieves the class activation map from the ground truth, which is the aggregate of each channel of the weighted feature map. For a given global image, let  $f_k(x, y)$  denote the activation of a spatial location  $(x, y)$  in the  $k^{th}$  channel of the output of the final convolutional layer, where  $k \in \{1, 2, \dots, K\}$ . For ResNet-50, this layer consists of 2048 activation maps, each with  $7 \times 7$  dimensions. The following average pooling layer diminishes the size of the previous layer to  $1 \times 1 \times 2048$  by considering the average of each feature map. Here, the weights  $W_k^c$  connect the dominant predicted class and the  $k^{th}$  node in the flattening layer, which represents the strength of each activation map for the predicted class. Inspiring by equation (4.4), we compute the CAM of each image belonging to the same class (e.g., the second row in Figures 4.5 and 4.6). Hence:

$$H_c(x, y) = \sum_{k=1}^K W_k^c f_k(x, y) \quad (4.6)$$

where  $H_c$  represents the class activation map of class  $c$  and  $W_k^c$  is the  $k^{th}$  weight of the softmax layer of class  $c$ . Also, we express  $f_k$  as the activation map, as presented in Figure 4.2.

For a class  $c$  containing a certain number of images  $I_n^c$ ,  $n \in \{1, 2, \dots, N\}$ , the mean class activation map is obtained by averaging the class activation maps of all images of the class  $c$  ( see Figure 4.4), which can be presented as:

$$meanH_c(x, y) = \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K W_k^n f_k^n(x, y) = \frac{1}{N} \sum_{n=1}^N H_n \quad (4.7)$$

where  $meanH_c$  is the mean class activation map (Mean-CAM); it specifies the common importance of each pixel in the initial images of a class  $c$ .  $W_k^n$  refers to the weights related

to the  $n^{\text{th}}$  image for the class  $c$  with the  $k^{\text{th}}$  channel.  $H_n$  is defined in relation (4.6).

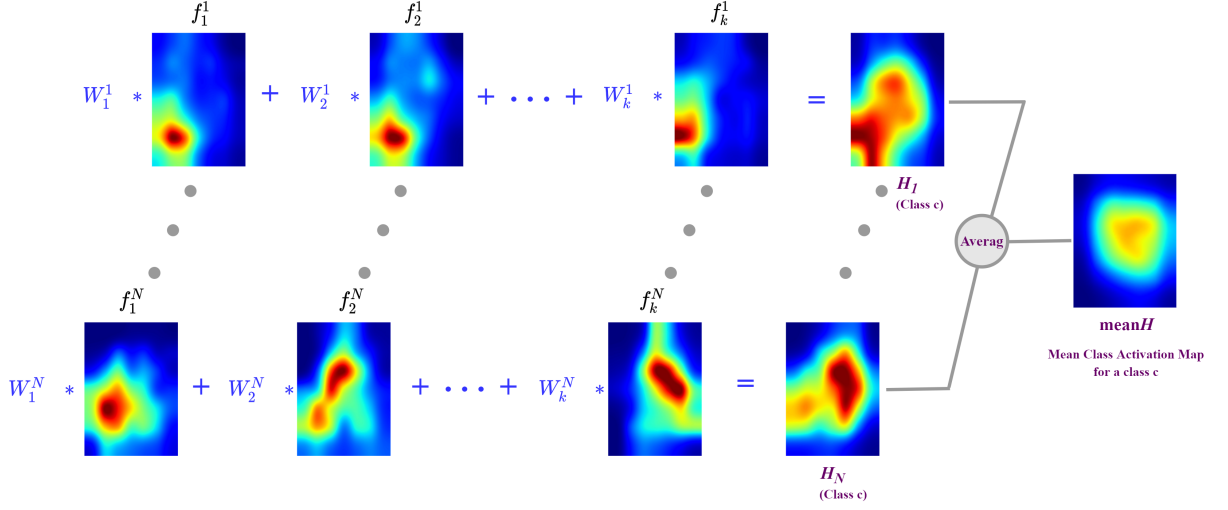


Figure 4.4: A graphical flowchart of the Mean-CAM process.

#### 4.2.4.3 Mask inference

The Mean-CAM represents the ensemble of important pixels in an original image of a given class. In this work, we developed a binary mask to determine the relevant regions in the global image for classification. In the mask inference process, the interesting region is identified by applying a threshold as follows:

$$M_{(x,y)} = \begin{cases} 1, & \text{if } \text{mean}H_c(x, y) > \tau \\ 0, & \text{Otherwise} \end{cases} \quad (4.8)$$

where  $\tau \in ]0, 1[$  is a threshold value that defines the size of the attended region.

A significant value of  $\tau$  results in a smaller area and vice versa. We create a maximum connected region around the mask  $M$ , which includes the discriminative points in  $M$ . The maximum connected region is given by the horizontal and vertical coordinates  $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ ; in this case, any  $(x, y)$  inside the connected region has a value of 1, otherwise it is 0. Finally, the local discriminative region  $I_{\text{crop}} = M.I$  is cropped from the original image  $I$  and resized to the input image size. Figures 4.5 and 4.6 show the bounding boxes and the cropped region with a threshold value of  $\tau = 0.5$ : (Top) Original ear images belonging to the same class for the global stage. (Top-down) The visual output of the mask inference utilising the CAM technique, where the main features are red. (Bottom-up) The

visual output of the mask inference utilising the Mean-CAM technique. Discriminative region is shown by Green bounding boxes. (Bottom) Cropped and resized images as the input image size from the green bounding boxes, used as input to the local stage.

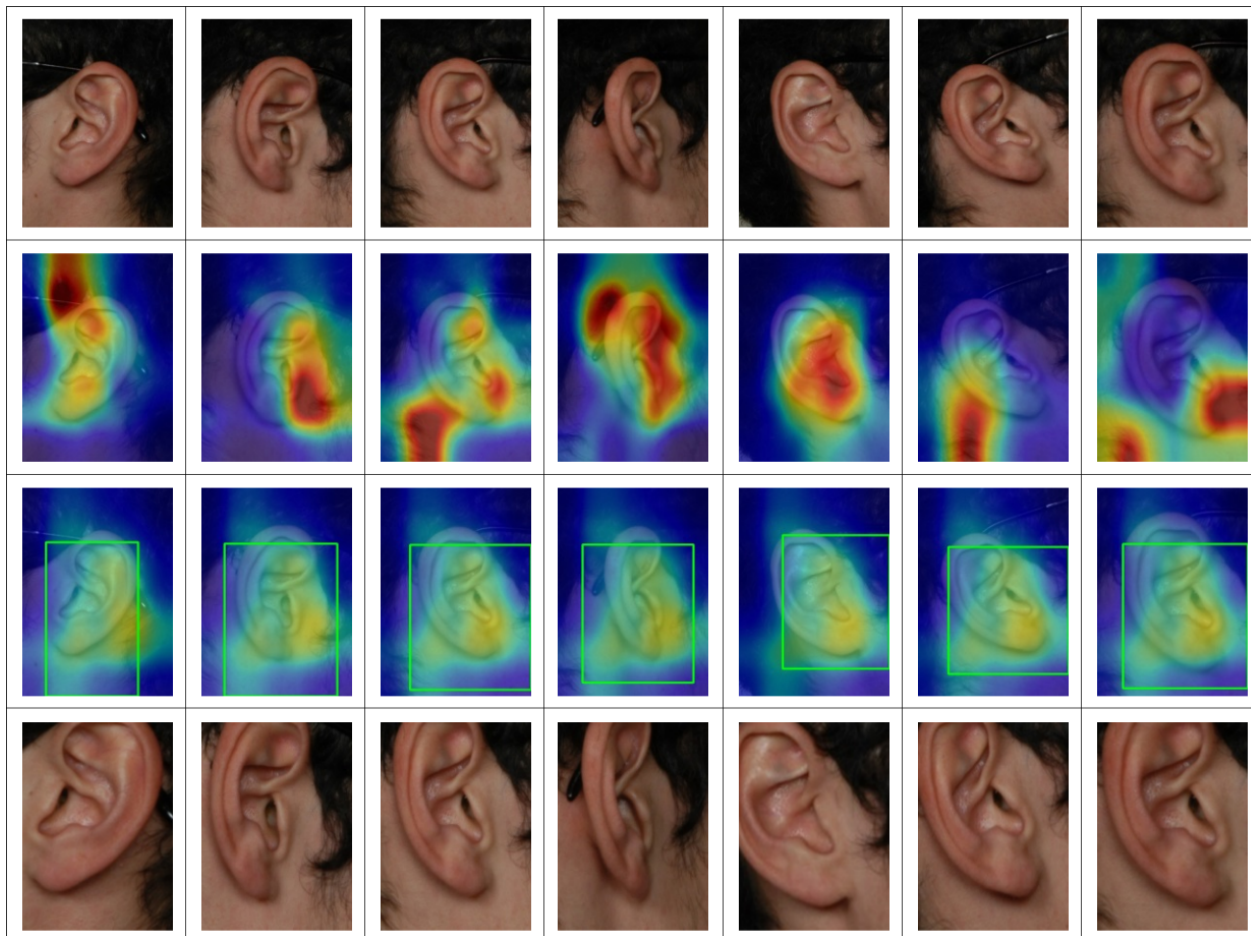


Figure 4.5: The process of generating discriminative regions for the AMI dataset.



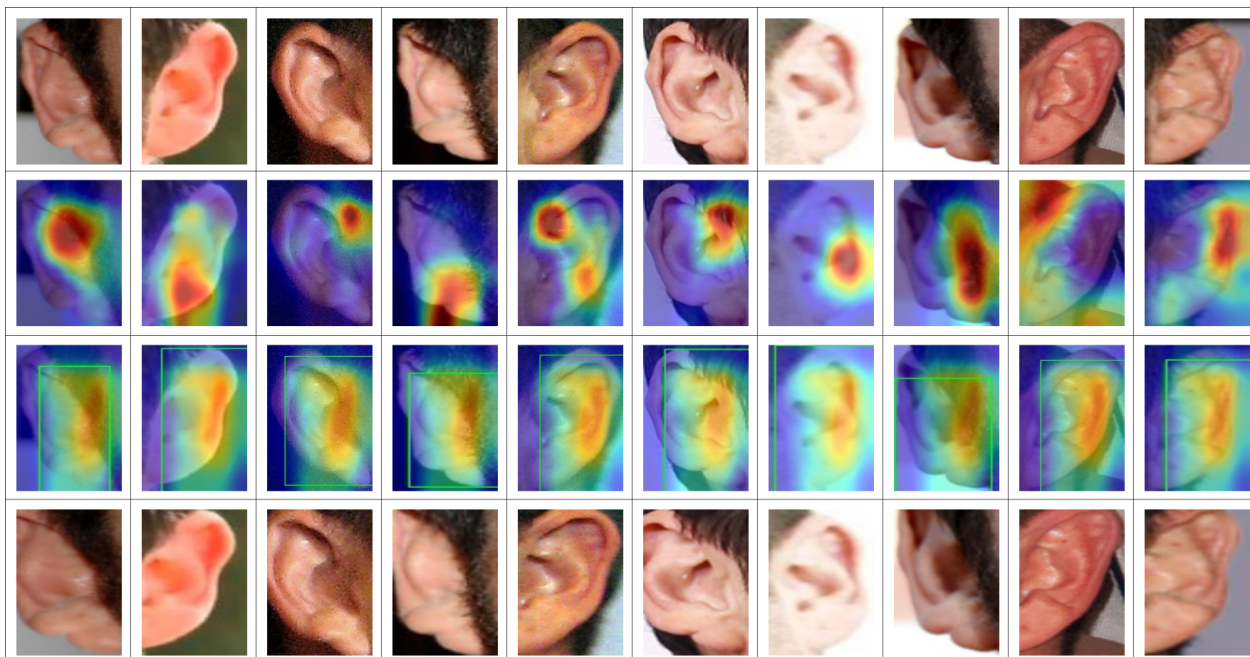


Figure 4.6: The process of generating discriminative regions for the AWE dataset.

Figures 4.5 and 4.6 (row 3) demonstrate that our local branch model enhances the representation of ear images by directing the model’s attention on the most significant common area across all images of the same individual. This area represents the most relevant features.

Algorithm 1 summarises the process of our proposed Mean-CAM-CNN for ear recognition.

---

**Algorithm 1** EAR RECOGNITION BASED ON MEAN-CAM-CNN

---

**Input:** Input image  $I$  ; Label vector  $L$ ; Threshold  $\tau$ .

**Output:** Predicted class.

**Initialization:** The weights of global and local stages (Random).

1. Compute  $P_g(c|I)$  optimized by cross-entropy ( $CE$ ) loss for the global stage;
  2. Compute the CAM for each image based on the global stage;
  3. Compute the Mean-CAM of the images belonging to the same class;
  4. Find the optimal mask and the bounding box coordinates  $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ ;
  5. Crop out  $I_{crop}$  from  $I$ ;
  6. Resize  $I_{crop}$  to the exact size of  $I$ ;
  7. Compute  $P_l(c|I_{crop})$  optimized by cross-entropy ( $CE$ ) loss for the local stage;
  8. Final decision.
-



## 4.3 Experimental analysis

To investigate the applicability of the proposed approach, we evaluated its performance using two unconstrained ear imaging data sets.

In this section, we performed several experiments to assess the performance of the proposed approach using two learning strategies (i.e., feature extraction and fine-tuning) without and with data augmentation. First, we describe the data sets used in this work. Then, the effects of different parameters of our formulation are evaluated using the following performance metrics:

- 1 . Rank-1 and Rank-5 recognition rates.
- 2 . Cumulative match-score curves (CMC).
- 3 . Area under the CMC curve (AUC).

Finally, we compared the performance of our proposed approach to current state-of-the-art ear recognition methods to evaluate its effectiveness.

### 4.3.1 Data sets

- **AMI:** The mathematical analysis of images (AMI) ear dataset was established by Esther González [94]. The dataset consists of 700 images obtained from a sample of 100 people aged between 19 and 65 years. For each person, a total of seven images are taken and gathered, consisting of one picture of the left ear and six pictures of the right ear. All images were taken indoors and under the same lighting conditions. These pictures are in JPEG extension and have a resolution of  $492 \times 702$  pixels.
- **AWE:** The annotated web ears (AWE) data set [95] covers 1000 images (left and right) of 100 persons, with ten shots per person. In contrast to the most available ear data sets that acquired images under supervised laboratory conditions, AWE ear images were collected from the net under unregulated conditions to ensure large intra-class variability. PNG is the format in which all images are stored, and their sizes vary from  $15 \times 29$  to  $473 \times 1022$  pixels, with an average size equal to  $83 \times 160$  pixels. The database has the following perturbations: occlusion, head rotation, race, gender, illumination, and blurring, making it a very challenging data set.

Examples of images of the databases previously cited (i.e., AMI and AWE) are shown in Figure 4.1.

### **4.3.2 Experimental protocols and setups**

During all experiments, the experimental evaluation protocol of the AMI database consists of choosing four samples from each person for the training set and the lasting samples for the testing set. Thus, we obtain 400 (i.e.,  $100 \times 4$ ) training samples and 300 (i.e.,  $100 \times 3$ ) testing samples. On the other hand, the experimental evaluation protocol of the AWE database consists of selecting six samples per person for the training set and the lasting samples for the testing set. Therefore, we achieve 600 (i.e.,  $100 \times 6$ ) training samples and 400 (i.e.,  $100 \times 4$ ) testing samples.

Experiments without and with data augmentation are established on both data sets for feature extraction and fine-tuning strategies. We have merged multiple transformation steps to change the original images and produce other variations. The image transformations applied to each training image in the gallery are:

- Resize the image to fit the model input ( $224 \times 224 \times 3$ ).
- Normalise the image using the mean and standard deviation.
- Random image rotation from -20 and +20 degrees.
- Image blurring using a Gaussian blur filter.
- Change the image's brightness, contrast, saturation, and hue by giving the desired range value.
- Flip the image horizontally with 50%.

### **4.3.3 Ablation analysis**

In this subsection, we evaluated the performance of the proposed Mean-CAM-CNN and the traditional ResNet-50 under two feature representation strategies, feature-extraction and fine-tuning, using the previously cited metrics. In the first case (i.e., feature-extraction strategy), when a model is trained from an extensive and mixed database like ImageNet, characteristics derived from network layers may be transmitted to other tasks, such as ear identification. While fine-tuning principle consists of updating the pretrained model's parameters for our new task. In this work, the available filters learned from the ImageNet database are adapted to the ear identification issue. We present and discuss the results obtained under the two modes of feature representation in the following subsections. Finally, the ResNet-50 was

chosen in this work as an experimental CNN model to explore if Mean-CAM-CNN can improve the performance of using only a classical CNN.

### 4.3.3.1 Feature extraction results

In the first part of the experiments, we assessed the performance of the suggested Mean-CAM-CNN framework by evaluating the representation capacity of their different thresholding values using the feature extraction-based strategy. We conducted our experimentations as defined in subsection 4.3.2. Besides, we examined the influence of data augmentation on the ResNet-50 and Mean-CAM-CNN performance for these experiments. Tables 4.2-4.3 present the Rank-1, Rank-5, and AUC results using the feature extraction mode based-strategy for models trained without and with augmented versions of the AMI and AWE databases. Bold values highlight the best result for each considered metric. Furthermore, Figures 4.7-4.8 display the CMC curves that highlight the features performance of the traditional ResNet-50 and the proposed Mean-CAM-CNN framework (with different thresholding values of the parameter  $\tau$ ) on both employed databases without and with data augmentation.

Table 4.2: Rank-1, Rank-5, and AUC results (%) using the feature extraction strategy for the AMI data set.

AMI Data set							
Strategy	Model	Without Data Augmentation			With Data Augmentation		
		R1	R5	AUC	R1	R5	AUC
Feature Extraction	ResNet-50	94.33	<b>99.66</b>	99.98	95.00	<b>100</b>	<b>99.99</b>
	Mean-CAM-CNN with $\tau = 0.1$	93.33	<b>99.66</b>	99.98	94.00	99.66	<b>99.99</b>
	Mean-CAM-CNN with $\tau = 0.2$	94.33	99.33	99.96	93.67	<b>100</b>	<b>99.99</b>
	Mean-CAM-CNN with $\tau = 0.3$	<b>96.33</b>	99.33	99.97	94.97	<b>100</b>	<b>99.99</b>
	Mean-CAM-CNN with $\tau = 0.4$	<b>96.33</b>	99.33	<b>99.99</b>	94.97	99.66	<b>99.99</b>
	Mean-CAM-CNN with $\tau = 0.5$	96.00	99.33	99.97	95.67	99.33	<b>99.99</b>
	Mean-CAM-CNN with $\tau = 0.6$	94.33	99.00	99.98	96.33	99.00	99.98
	Mean-CAM-CNN with $\tau = 0.7$	93.00	98.33	99.97	<b>97.00</b>	99.66	<b>99.99</b>
	Mean-CAM-CNN with $\tau = 0.8$	91.33	98.66	99.96	93.00	99.00	99.97
	Mean-CAM-CNN with $\tau = 0.9$	94.00	99.00	99.96	93.33	99.33	<b>99.99</b>

Table 4.3: Rank-1, Rank-5, and AUC results (%) using the feature extraction strategy for the AWE data set.

AWE Data set							
Strategy	Model	Without Data Augmentation			With Data Augmentation		
		R1	R5	AUC	R1	R5	AUC
Feature Extraction	ResNet-50	18.00	38.00	83.23	20.25	43.50	83
	Mean-CAM-CNN with $\tau = 0.1$	17.25	41.00	83.46	20.25	41.75	83.19
	Mean-CAM-CNN with $\tau = 0.2$	18.00	41.50	83.31	20.25	44.00	83.24
	Mean-CAM-CNN with $\tau = 0.3$	19.00	38.50	81.90	22.25	46.50	84.68
	Mean-CAM-CNN with $\tau = 0.4$	19.75	42.25	84.53	27.75	51.75	86.74
	Mean-CAM-CNN with $\tau = 0.5$	21.75	48.25	87.44	25.50	52.25	89.26
	Mean-CAM-CNN with $\tau = 0.6$	25.00	54.50	89.84	28.75	<b>60.50</b>	<b>91.73</b>
	Mean-CAM-CNN with $\tau = 0.7$	24.75	53.50	<b>91.60</b>	29.75	56.75	90.72
	Mean-CAM-CNN with $\tau = 0.8$	<b>26.00</b>	<b>55.75</b>	91.25	<b>33.75</b>	58.25	91.64
	Mean-CAM-CNN with $\tau = 0.9$	18.00	38.00	83.23	19.50	42.75	83.27

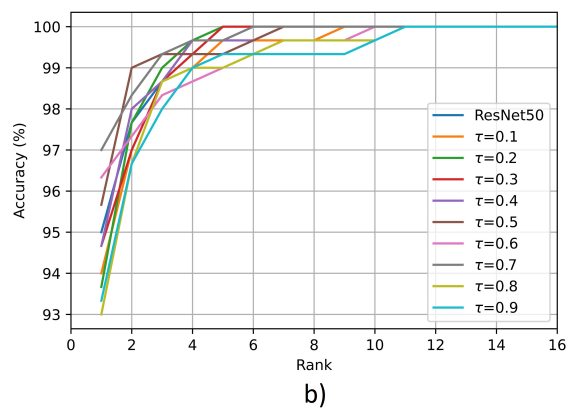
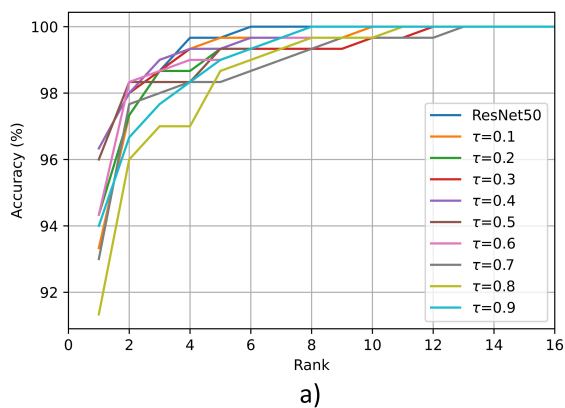


Figure 4.7: CMC curves produced on the testing set from the AMI database using the feature extraction strategy: a) Without data augmentation. b) With data augmentation.

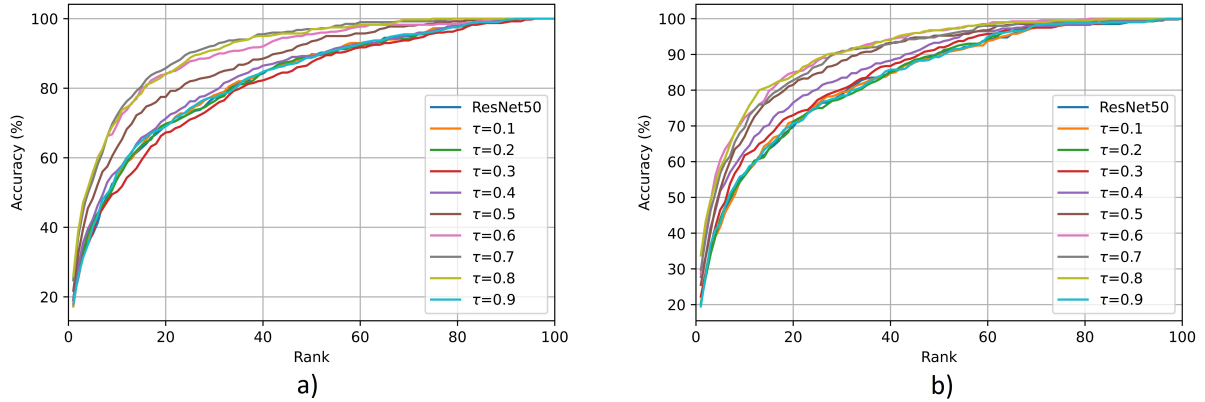


Figure 4.8: CMC curves produced on the testing set from the AWE database using the feature extraction strategy: a) Without data augmentation. b) With data augmentation.

As a result, the Mean-CAM-CNN framework and the traditional ResNet-50 perform better on AMI than in AWE due to the low number of images for the AMI database and the complex and unconstrained nature of the AWE’s images. Besides, data augmentation has a positive impact on improving accuracy by an average of 0.3% for AMI, 3.1% for AWE, and 1.7% for their mixture. The influence of data augmentation is more significant on the AWE database than in the AMI database; the sizable intra-class variability can explain this for the AWE database.

For the AMI database, the results of the ResNet-50 and Mean-CAM-CNN (with different values of the parameter  $\tau$ ) are almost all similar and close (with some marginal variations) either without or with data augmentation; this can be explained by the low intra-class variability (i.e., variability between images of the same person) and the high inter-class variability as well as by the higher learning capacity of the two trained models used in this experiment (i.e., ResNet-50 and Mean-CAM-CNN). However, for the AWE database, the results of the Mean-CAM-CNN with the different parameters are almost all the better than the ResNet-50. In addition, data augmentation has a very positive effect in improving results for the three employed metrics.

Overall, for the AMI database, the results are acceptable, with a best achieved Rank-1 recognition rate equal to 97%. However, the results are not satisfactory for the AWE database, noting that the best Rank-1 recognition rate achieved is equal to 33.75%. AMI is a simple and not challenging database, and in contrast, AWE is considered among the most challenging databases in the topic of ear recognition. In the feature extraction mode and for all tested cases using two databases without and with data augmentation, no particular

configuration can be considered a generalized case, i.e., that gives a better result and performs well for the two employed databases.

### 4.3.3.2 Fine-tuning results

In the second part of the experiments, we evaluated the performance of the suggested Mean-CAM-CNN framework by assessing the representation capacity of their different threshold values using the fine-tuning-based strategy. The protocols of evaluation conducted in these experiments are described in subsection 4.2. As we did in the previous experiments, we assessed the impact of augmenting data on the ResNet-50 and Mean-CAM-CNN performance. Tables 4.4 and 4.5 show the Rank-1, Rank-5, and AUC results using the fine-tuning-based strategy for models trained without and with expanded versions of the AMI and AWE databases. Bold values highlight the best results for each specific metric. In addition, Figures 4.9 and 4.10 plot the CMC curves that indicate the features performance of the ResNet-50 and the proposed Mean-CAM-CNN framework (with different thresholding values of the parameter  $\tau$ ) on both employed databases without and with data augmentation.

Table 4.4: Rank-1, Rank-5, and AUC results (%) using the fine-tuning strategy for the AMI data set.

AMI Data set							
Strategy	Model	Without Data Augmentation			With Data Augmentation		
		R1	R5	AUC	R1	R5	AUC
Fine Tuning	ResNet-50	98.33	<b>100.00</b>	<b>100.00</b>	98.66	99.66	<b>100.00</b>
	Mean-CAM-CNN with $\tau = 0.1$	97.00	<b>100.00</b>	99.98	96.67	<b>100.00</b>	<b>100.00</b>
	Mean-CAM-CNN with $\tau = 0.2$	97.67	99.66	99.99	98.00	<b>100.00</b>	99.99
	Mean-CAM-CNN with $\tau = 0.3$	98.67	<b>100.00</b>	<b>100.00</b>	98.66	99.66	99.99
	Mean-CAM-CNN with $\tau = 0.4$	99.33	<b>100.00</b>	<b>100.00</b>	99.33	99.33	99.99
	Mean-CAM-CNN with $\tau = 0.5$	99.00	<b>100.00</b>	99.99	<b>99.67</b>	<b>100.00</b>	<b>100.00</b>
	Mean-CAM-CNN with $\tau = 0.6$	<b>99.67</b>	99.67	<b>100.00</b>	<b>99.67</b>	<b>100.00</b>	<b>100.00</b>
	Mean-CAM-CNN with $\tau = 0.7$	99.00	99.67	99.99	98.33	<b>100.00</b>	<b>100.00</b>
	Mean-CAM-CNN with $\tau = 0.8$	98.00	99.67	99.99	98.66	99.66	99.99
	Mean-CAM-CNN with $\tau = 0.9$	98.67	99.67	99.99	96.00	99.00	99.99

Table 4.5: Rank-1, Rank-5, and AUC results (%) using the fine-tuning strategy for the AWE data set.

AWE Data set							
Strategy	Model	Without Data Augmentation			With Data Augmentation		
		R1	R5	AUC	R1	R5	AUC
Fine Tuning	ResNet-50	29.00	50.25	86.69	57.75	79.00	96.54
	Mean-CAM-CNN with $\tau = 0.1$	25.00	46.25	85.45	62.25	79.50	96.70
	Mean-CAM-CNN with $\tau = 0.2$	26.75	51.50	85.98	58.25	78.75	97.02
	Mean-CAM-CNN with $\tau = 0.3$	26.00	47.50	83.49	68.50	85.75	97.99
	Mean-CAM-CNN with $\tau = 0.4$	31.00	56.75	89.15	68.75	83.25	98.56
	Mean-CAM-CNN with $\tau = 0.5$	36.25	61.25	91.15	<b>74.50</b>	<b>89.50</b>	<b>98.93</b>
	Mean-CAM-CNN with $\tau = 0.6$	46.00	69.50	<b>94.59</b>	69.25	87.00	98.56
	Mean-CAM-CNN with $\tau = 0.7$	<b>47.00</b>	<b>71.25</b>	94.43	67.50	87.25	98.34
	Mean-CAM-CNN with $\tau = 0.8$	45.50	69.25	93.31	62.00	85.25	97.88
	Mean-CAM-CNN with $\tau = 0.9$	22.00	44.00	84.78	60.00	78.75	96.81

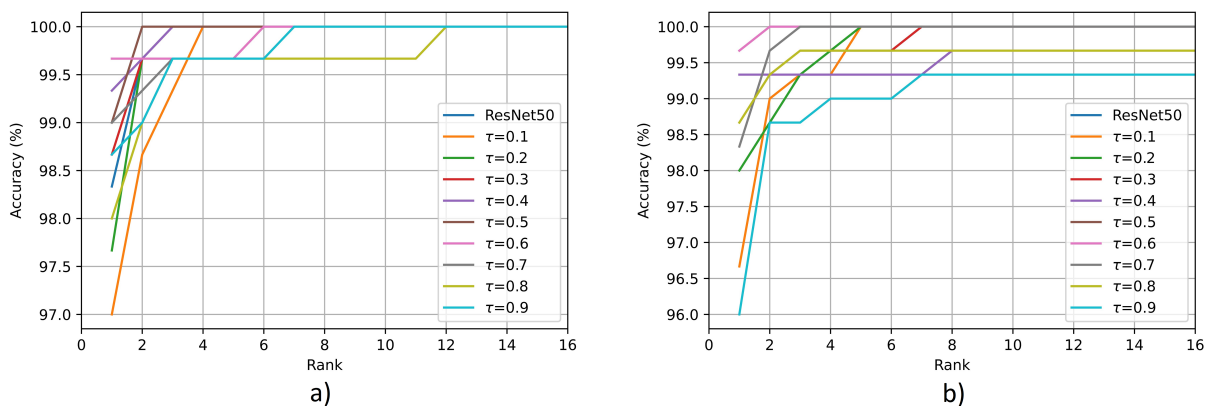


Figure 4.9: CMC curves produced on the testing set from the AMI database using the fine-tuning strategy: a) Without data augmentation. b) With data augmentation.

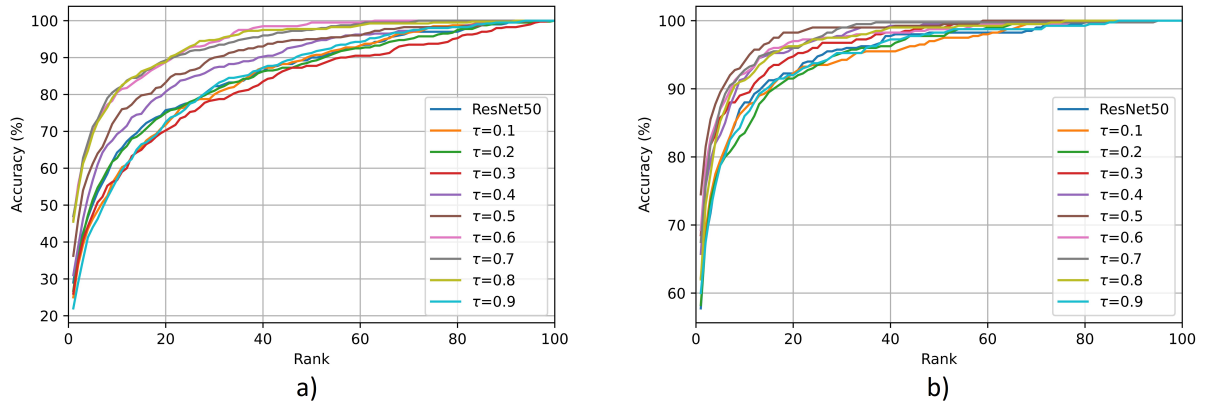


Figure 4.10: CMC curves produced on the testing set from the AWE database using the fine-tuning strategy: a) Without data augmentation. b) With data augmentation.

Similar to the first part of the experiments, we observe that the Mean-CAM-CNN framework and the traditional ResNet-50 perform better on AMI than in AWE. For the AMI database, the results of the ResNet-50 and Mean-CAM-CNN (with different values of the parameter  $\tau$ ) are almost all similar and close (with some marginal variations) either without or with data augmentation. However, for the AWE database, the results of the Mean-CAM-CNN with the different parameters are almost all the better than the ResNet-50.

Nevertheless, for the second part of the experiments, we can see that data augmentation has marginally decreased the accuracy by an average of 0.09% for the AMI database and vastly improved it by 22.34% for the AWE database. In addition, the average accuracy of combining both data sets has been improved by 11.22%. As the AMI database is characterized by lower intra-class variability, the presence of additional intra-class images leads to more avenues for error, which justify the marginal decrease in the recognition performance. In contrast, generating more intra-class images for the AWE data set has significantly improved the recognition performance. As large intra-class variability characterizes the AWE data set, generating more intra-class images has reduced the distance between images of the same person.

The highest Rank-1 recognition rates achieved are 99.67% and 74.50% for AMI and AWE data sets. These best Rank-1 recognition rates are achieved using the configuration: Mean-CAM-CNN with the parameter  $\tau = 0.5$  and data augmentation. This configuration is the generalized case because it achieves the highest performance for both data sets and under all employed performance metrics. Compared to the feature extraction-based strategy, the best Rank-1 recognition rates using the fine-tuning-based strategy have increased by 2.67% and



41.75% for the AMI and AWE data sets, respectively. Compared to the classical ResNet-50 model, the best Rank-1 recognition rates using the proposed Mean-CAM-CNN model have increased by 1.01% and 16.75% for the AMI and AWE data sets, respectively. Finally, the power of the Mean-CAM-CNN, especially with the challenging data set AWE, can be justified by exploiting only the valuable and pertinent information for recognition and eliminating the irrelevant information (without neglecting the vital role of using data augmentation). In fact, Mean-CAM allows the CNN to focus only on the region of interest containing the valuable information during the features-extraction/classification phase.

### 4.3.4 Comparison

For a comprehensive analysis, we compared the results of the proposed Mean-CAM-CNN with some recently published state-of-the-art ear recognition approaches. In this comparison, the Rank-1 recognition rate was used as an assessment measure with both data sets AMI and AWE. The comparison results using the previously defined protocols of each data set are presented in Table 4.6. We can see that most of the compared papers have achieved satisfactory Rank-1 recognition rates with the AMI data set, like [96] and [97, 98, 99, 100, 101, 102], with recognition rates surpassing 93%. Similarly, our approach has the highest and most competitive performance, with a Rank-1 recognition rate of 99.67%. The good results achieved with most papers using the AMI data set can be justified by its images' simple and constrained nature. By contrast, the results of the majority of competing papers, handcrafted or deep learning-based, are shallow and near to 50% with the challenging AWE data set, except the work of Dodge et al. [103], which has achieved 68.50%; this can be justified by the complex and unconstrained nature of the AWE images. Nevertheless, the results of our proposed Mean-CAM-CNN appear very acceptable and competitive, with a Rank-1 recognition rate of 74.50%, because the Mean-CAM method allows the CNN to exploit only the valuable information for recognition and eliminates irrelevant information.

Table 4.6: Comparing Mean-CAM-CNN Rank-1 recognition rate with several competing approaches using AMI and AWE ear recognition data sets.

Approach	Year	Publication	Method	AMI	AWE
<b>Handcrafted</b>	2017	Emeršič et al. [95]	POEM	--	49.60
	2018	Chowdhury et al. [104]	Tunable Filter Bank	70.58	--
	2019	Hassaballah et al. [33]	LBP's variants	73.71	49.60
	2020	Hassaballah et al. [34]	RLQP	72.29	54.10
	2020	Khaldi and Benzaoui [105]	Pix2Pix-GAN + BSIF	--	44.53
	2020	Sarangi et al. [35]	Jaya algorithm + SURF	--	44.00
	2020	Sajadi and Fathi [36]	GZ + LPQ	--	53.50
	2021	Lavanya et al. [96]	Gaussian filter+Canny edge detector	96.60	--
<b>Deep-learning</b>	2018	Dodge et al. [103]	ResNet18 Fine-tuned	--	68.50
	2018	Zhang et al. [106]	VGG-face	--	50.00
	2019	Alshazly et al. [97]	AlexNet (Fine Tuning)	94.50	--
	2019	Alshazly et al. [98]	VGG-13-16-19-ensembles	97.50	--
	2019	Zhang et al. [99]	MAML + CNN	93.96	--
	2020	Priyadharshini et al. [100]	CNN	96.99	--
	2021	Khaldi and Benzaoui [101]	DCGAN + VGG16	96.00	50.53
	2021	Khaldi et al. [102]	VGG16 + DUAL	98.33	51.25
	<b>2023</b>	<b>Our Proposed Method</b>	<b>Mean-CAM-CNN</b>	<b>99.67</b>	<b>74.50</b>

## 4.4 Conclusion

This chapter proposes an original and practical framework based on Mean-CAM-CNNs in the challenging context of unconstrained image classification and recognition. The recognition process consists of two main phases: a global phase and a local phase. In the first phase, the proposed Mean-CAM method extracted a region of interest from the image based on the Mean heat maps derived from the images of the same class. In the second phase, we trained a CNN using only the extracted areas to classify and recognise the images. Further experiments were conducted extensively using the AMI and AWE unconstrained ear recognition databases to prove the effectiveness of the proposed approach. Four different scenarios were

implemented and compared in an ablation analysis involving two training strategies, namely feature extraction and fine tuning, without and with data augmentation. The best Rank-1 recognition rates obtained are 99.67% and 74.50% for AMI and AWE, respectively, using fine tuning with data augmentation. In addition, the Mean-CAM-CNN network has improved recognition performance compared to the use of the conventional CNN. ResNet-50 was used as the experimental CNN model, and the improvement was 1.01% and 16.75% for the AMI and AWE data sets, respectively. These significant improvements confirm the effectiveness of the proposed approach and make it more efficient than current ear recognition methods. The added value of the proposed framework is justified by the superior learning ability of the CNNs, in particular the ResNet-50 model, the elimination of irrelevant information and the extraction of only information useful for recognition using the Mean-CAM cropping strategy.

In the next chapter we will employ deep convolutional generative adversarial networks (DCGANs) to enhance ear images, in order to boost the performance of our approach.

## Chapter 5

# Boosting the Performance of Deep Ear Recognition Systems Using Generative Adversarial Networks and Mean Class Activation Maps

## **5.1 Introduction**

This chapter delves into a two-step method of ear recognition. Deep convolutional generative adversarial networks (DCGANs) are implemented in the initial phase for enhancing ear images. This entails the colourization of greyscale images and the improvement of dark hues, thereby resolving visual flaws. Mean-CAM-CNN is subsequently introduced as a feature extraction and classification technique. Mean-class activation maps are employed in conjunction with CNNs in this approach. The Mean-CAM approach directs the CNN to concentrate on pertinent information, extracting and evaluating only significant regions within the entire image. The procedure entails the use of a mask to selectively crop the relevant region of the image. The cropped region is subsequently employed to train a CNN for discriminative classification.

The rest of the chapter is organised as follows: Second section introduces the proposed approach. Experimental analysis is carried out in the third section. Section four conducted a comparative analysis. Finally, Section five concludes the chapter..

## **5.2 Proposed approach**

This section outlines the approach used in this work to address the complex task of identifying ears. The approach involves two fundamental processes: first, image preprocessing using DCGAN, followed by feature extraction and classification utilising Mean-CAM and a fine-tuned CNN model. In the following sub-sections, we provide comprehensive information about each of the two phases.

### **5.2.1 Preprocessing**

In the examined datasets, a significant part of images show poor visual quality, frequently appearing dark or monochromatic in the AWE dataset. In an earlier study [42], we presented a framework using a generative model for the purpose of adding colour to greyscale and enhancing dark ear images. This resulted in enhanced performance when evaluated by a trained classifier. We used deep convolutional generative adversarial networks (DCGANs) for colourisation. A typical U-Net framework was developed for the generative model to provide colourised ear representations, as seen in Figure 5.1.

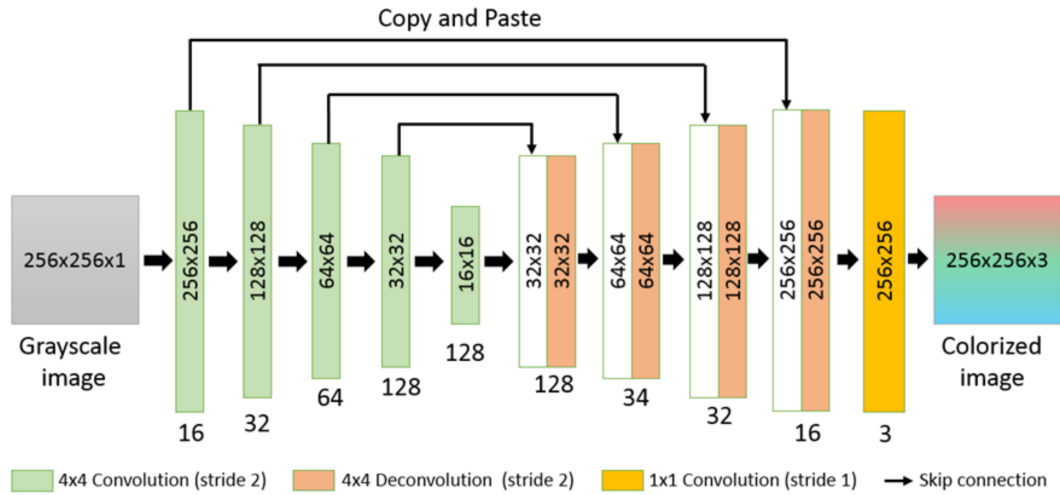


Figure 5.1: U-Net design for the generative paradigm.

The discriminative model’s structural design, as shown in Figure 5.2, has a more simpler architecture compared to that of the generator. The model consists of five convolutional layers that use  $4 \times 4$  filters with a stride of two. The model also incorporates batch normalization and Leaky-ReLU activation. The final layer utilises a single  $4 \times 4$  filter with a stride of one and is activated using the sigmoid function. This activation produces a scalar value that represents the authenticity of the resulting image. The binary classification model produces a probability output ranging from 0 to 1, denoted as  $P$  in Figure 5.2.

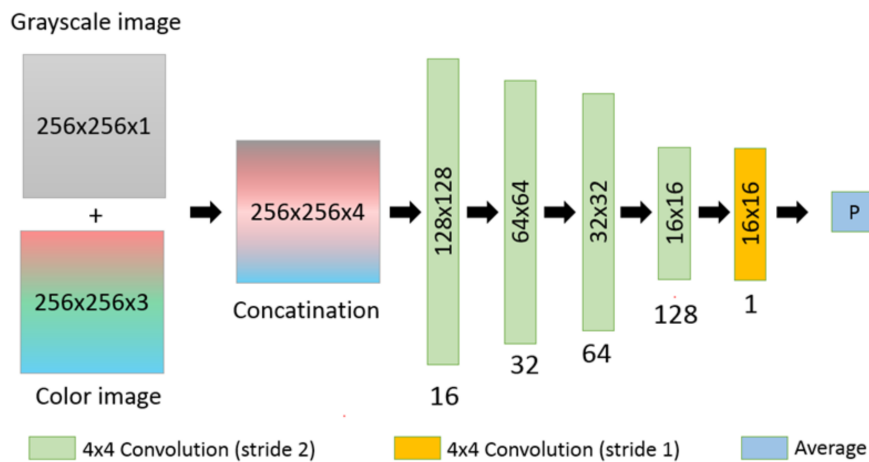


Figure 5.2: Design of the discriminative model.

The discriminator receives a  $256 \times 256 \times 4$  input, achieved by combining a grayscale image with either a real or artificially created colour image. The main objective is to maximize the probability of correctly distinguishing between created and authentic images, represented as

$(\log D(y|x))$ . On the other hand, the generator endeavours to reduce the opposite statement,  $1 - \log D(G(\theta_z|x))$ . The final cost function, expressed as  $V$ , is formally represented by Equation (5.1). [107].

$$\min_G \max_D V(G, D) = \mathbb{E}_x[\log D(y|x)] + \mathbb{E}_z[1 - \log D(G(O_z|x))] \quad (5.1)$$

Where the greyscale image is represented as  $x$ , the ground truth as  $y$ , and  $G(O_z|x)$  denotes the mapping function that captures the generator's output colour image obtained from the input image  $x$ . The discriminator is represented by the mapping function  $D(y|x)$ , which produces a scalar value between 0 and 1. This value indicates the probability of the input being created or not. The symbol  $\mathbb{E}_x$  denotes the average value taken over all actual colour pictures, whereas  $\mathbb{E}_z$  indicates the average value calculated over all created colour images.

Equation (5.2) defines the goal of training the adversarial model, which is to minimize the average Euclidean distance between colourised images ( $\theta$ ) and ground truth images ( $y$ ) at the pixel level. The formulation displays a greyscale picture as  $x$ , with  $e$  as the channel index, and  $i, j$  as indices pointing to individual pixels.

$$Distance(x, \theta) = \frac{1}{3uv} \sum_{e=1}^3 \sum_{i=1}^u \sum_{j=1}^v \left\| x_{i,j}^e - y_{i,j}^e \right\|_2^2 \quad (5.2)$$

Figure 5.3 clearly depicts the results that occur once the preprocessing is applied. The findings indicate that the images were successfully colourised and that consistent brightness and intensity equalization were achieved, along with standardization of image dimensions. The suggested DCGAN model was used to produce these transformations.

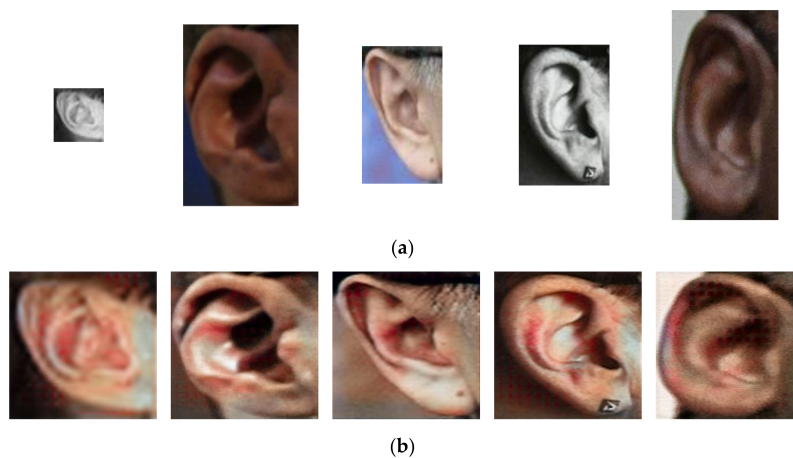


Figure 5.3: The proposed DCGAN model is used to colour, enhance, and resize ear images.: (a) original images; (b) enhanced images.

## **5.2.2 Feature extraction/classification**

This phase signifies the primary contribution of the current work. Figure 5.4 illustrates the architectural structure of the proposed feature extraction/classification approach, referred to as Mean-CAM-CNN. It consists of two separate branches: the global stage and the local stage. Every step incorporates a specialized CNN. In a global stage, the first CNN takes in all images from the same class and produces CAMs for each image. The average CAM is calculated by aggregating the CAMs of images belonging to the same class. The average CAM is used to extract and use the important area of each initial image in the training and classification processes at the local stage. To summarize, the use of Mean-CAM improves the ability to make predictions by guiding the focus of the CNN towards the most important common area, thereby including the most significant characteristics.

Figure 5.4 presents a detailed representation of the Mean-CAM-CNN architecture, in which the ResNet-50 He et al. (2016) is used as the main backbone model. The Mean-CAM-CNN consists of two main stages: global and local. The architecture of each stage is composed of five convolutional blocks, which include batch normalization and rectified linear unit (ReLU) activation functions. The blocks are seamlessly linked to global max-pooling (GMP), a fully connected (FC) layer, and a softmax layer.



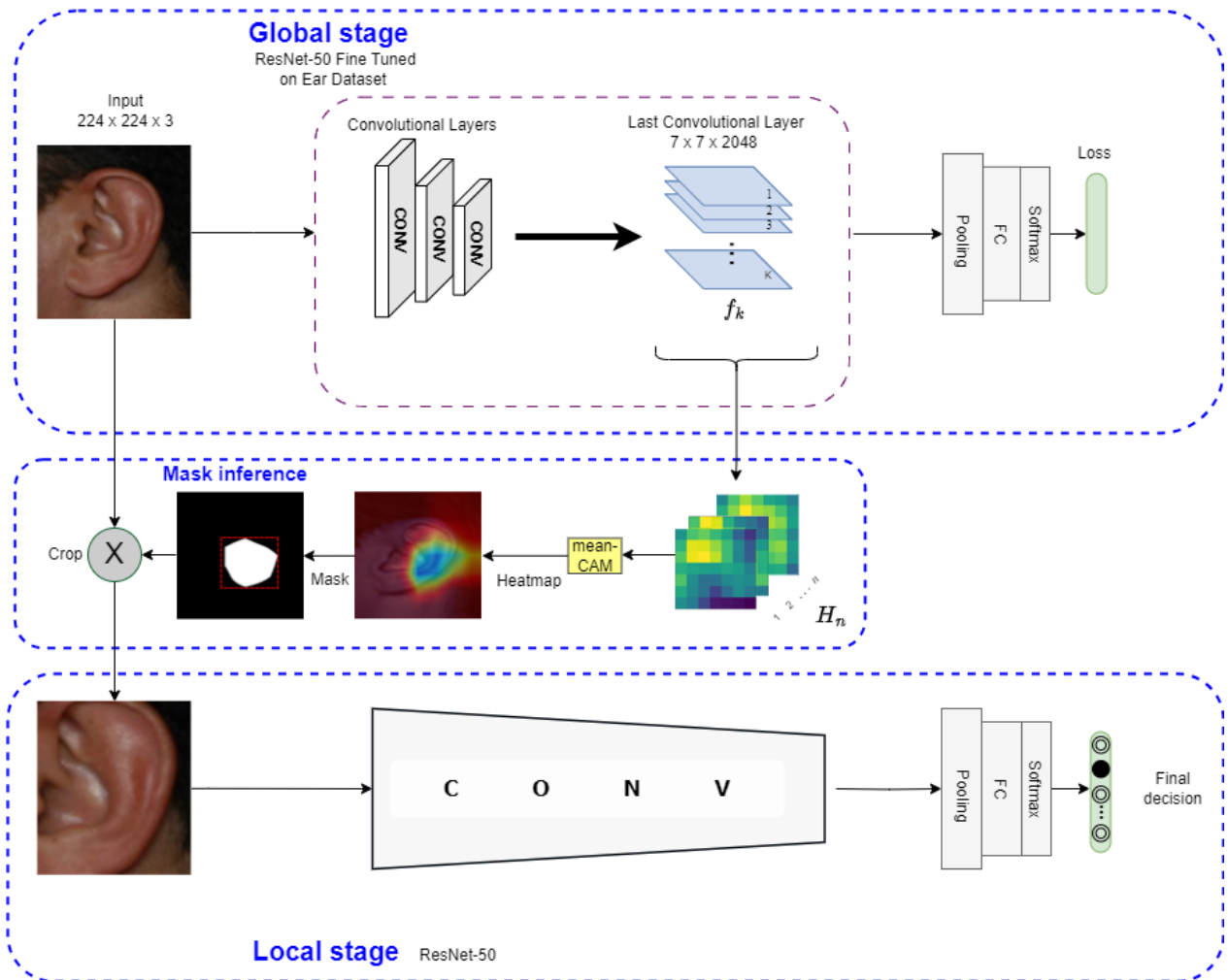


Figure 5.4: An illustration of the suggested mean-CAM-CNN framework’s flowchart.

Unlike the global stage, the local stage is defined by a set of local images that are limited by the mask created in the global stage. To provide a clear example, the input image is overlaid into the heatmap.

The processes of the proposed Mean-CAM-CNN approach is well explained in Chapter 4, specifically in sections 4.2.2 and 4.2.4, with clear explanations and mathematical expressions.

### 5.3 Experimental analysis

This section includes a series of experiments carried out to evaluate the effectiveness of the suggested method. The assessment of the suitability of the suggested method was conducted on two different unconstrained ear image datasets. The following description provides an overview of the datasets used in this study. Afterwards, we analyse the effect of different

factors in our formulation by using the following performance metrics:

1. Rank-1 and Rank-5 recognition rates.
2. Cumulative match score curves (CMCs).
3. Area under the CMC curve (AUC).

Finally, the new strategy is compared to previously published ear recognition methods to determine its usefulness.

A comprehensive discussion of the two datasets in chapter 4, paragraph 4.3.1 and 4.3.2, including the evaluation protocol used.

### **5.3.1 Evaluation protocols and setup**

In these experiments, we followed the same protocols as those used in Chapter 4. The experiments were performed using a laptop computer equipped with an Intel(R) Core(TM) i7-10750H CPU, 16 GB of RAM, and an Nvidia RTX 2060 graphics card. The experiments were conducted using the Anaconda framework as the simulation environment, with PyTorch version 1.9.0.

The CNN models were trained using the RMSprop optimizer, which used cross-entropy loss with momentum and a decay value of 0.9. Significantly, a weight decay of 0.0001 was selected to reduce over-fitting, and this value was consistently used in all experimental configurations. The initial learning rate was set at 0.005 and gradually decreased to 0.00005 during the training procedure. The model was trained for 200 epochs on the AMI dataset, whereas it took 320 epochs to attain complete convergence on the AWE dataset.

### **5.3.2 Experiment 1**

During the first experiment, we conducted training sessions using various CNN architectures with the goal of identifying ear images. We used two datasets, namely AMI and AWE. The datasets were used as input for four pre-trained CNN architectures. AlexNet [69], VGG-16 [72], VGG-19 [72], and ResNet-50 [75]. Pre-trained CNNs were initialized by adjusting their weights to match those of the ImageNet models. [108]. Subsequently, the CNNs were fine-tuned on the ear datasets, using reduced learning rates to allow the networks to modify their weights according to the unique features of ear images. Data augmentation methods were used to improve the training of the CNNs.

The main goal of this experiment was to determine the most resilient structure among the models being investigated, specifically designed for the purpose of ear classification. It is important to emphasise that in this experiment, neither the recommended pre-processing nor the Mean-CAM procedures were used. The findings of this experiment are shown in Table 5.1, with the best outcomes emphasised in bold.

Table 5.1: Comparative evaluation of fine-tuned CNN architectures for ear recognition.

Metric	Architecture	AMI	AWE
Rank-1 (%)	AlexNet	88.50	30.25
	VGG-16	95.50	47.25
	VGG-19	91.50	40.25
	ResNet-50	<b>98.66</b>	<b>57.75</b>
Rank-5 (%)	AlexNet	95.50	50.25
	VGG-16	99.50	73.75
	VGG-19	98.50	66.75
	ResNet-50	<b>99.66</b>	<b>79.00</b>
AUC (%)	AlexNet	92.11	56.54
	VGG-16	94.56	75.41
	VGG-19	93.91	72.44
	ResNet-50	<b>100.00</b>	<b>96.54</b>

The results shown in Table 5.1 clearly demonstrate that ResNet-50 performs better than other competing architectures in all three performance metrics used. The performance of ResNet-50 on the AMI dataset is remarkable, with a Rank-1 recognition rate of 98.66%. However, the Rank-1 rate for the AWE dataset was much lower, measuring at 57.75%. The AMI dataset is characterized by its simplicity and lack of complexity, whereas the AWE dataset is widely regarded as one of the most challenging datasets for ear recognition. This disparity highlights the need for a different method that goes beyond just increasing of the training set and adjusting the weights.

### 5.3.3 Experiment 2

The objective of the second experiment was to suggest a different approach to improve the performance of the fine-tuned ResNet-50 model. Our recommendation is to combine the Mean-CAM framework with the ResNet-50 model. As shown in Section 5.2, Mean-CAM

Table 5.2: Ear recognition results using ResNet-50 and Mean-CAM-CNN across different threshold values of the  $\tau$  parameter.

		AMI			AWE		
		Rank-1 (%)	Rank-5 (%)	AUC (%)	Rank-1 (%)	Rank-5 (%)	AUC (%)
<b>ResNet-50</b>		98.66	99.66	<b>100.00</b>	57.75	79.00	96.54
Mean-CAM-CNN	$\tau = 0.1$	96.67	<b>100.00</b>	<b>100.00</b>	62.25	79.50	96.70
	$\tau = 0.2$	98.00	<b>100.00</b>	99.99	58.25	78.75	97.02
	$\tau = 0.3$	98.66	99.66	99.99	68.50	85.75	97.00
	$\tau = 0.4$	<b>99.33</b>	99.33	99.99	68.75	83.25	98.56
	$\tau = 0.5$	<b>99.67</b>	<b>100.00</b>	<b>100.00</b>	<b>74.50</b>	<b>89.50</b>	<b>98.93</b>
	$\tau = 0.6$	<b>99.67</b>	<b>100.00</b>	<b>100.00</b>	69.25	87.00	98.56
	$\tau = 0.7$	98.33	<b>100.00</b>	<b>100.00</b>	67.50	87.25	98.34
	$\tau = 0.8$	98.66	99.66	99.99	62.00	85.25	97.88
	$\tau = 0.9$	96.00	99.00	99.99	60.00	78.75	96.81

improves the representation of ear images by directing the model’s attention to the most important common region, which represents the most relevant attributes. The project focuses on evaluating whether Mean-CAM-CNN may enhance recognition performance compared to using a traditional CNN alone.

The performance of the Mean-CAM-CNN architecture was assessed by examining the suitability of different threshold values for the  $\tau$  parameter. The evaluation strategies used in these experiments are described in Section 5.3.1. Table 5.2 displays the Rank-1, Rank-5, and AUC outcomes for the AMI and AWE datasets. The best results for each metric are emphasized with bold values. Furthermore, Figure 5.5 displays the CMCs, which illustrate the performance of ResNet-50 features and the proposed Mean-CAM-CNN framework. The figure considers various threshold values of the  $\tau$  parameter and includes data from two independent datasets.

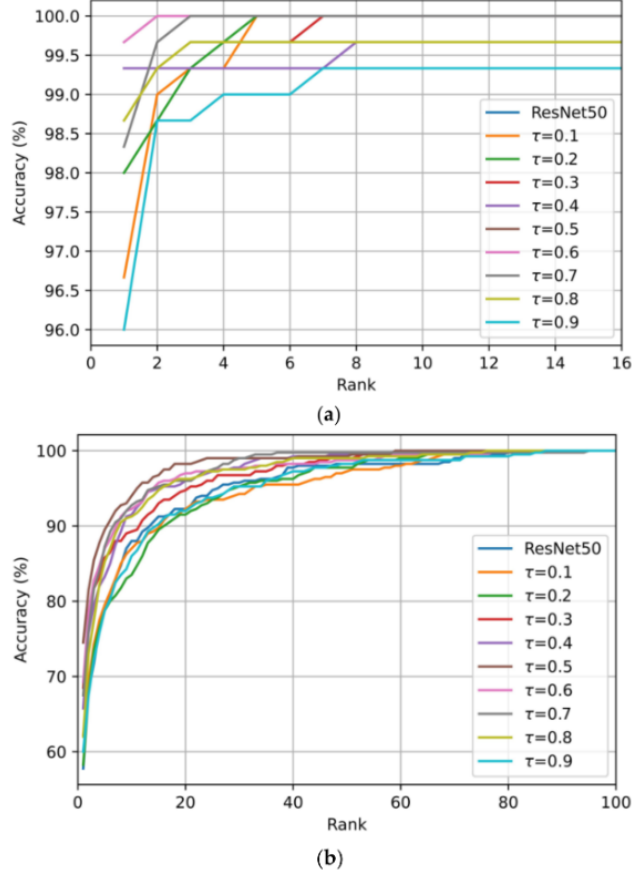



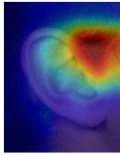
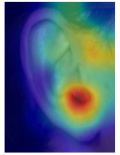

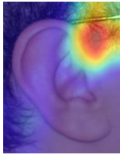
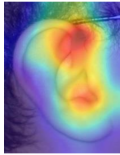
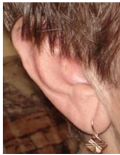
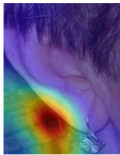
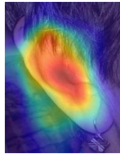

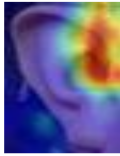
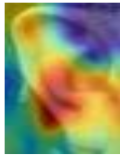
Figure 5.5: CMCs for the testing sets of (a) AMI and (b) AWE datasets: comparative analysis of ResNet-50 and Mean-CAM-CNN with varied threshold values ( $\tau$ ).

In this experiment, we find that the performance outcomes of ResNet-50 and Mean-CAM-CNN, using different values of the  $\tau$  parameter, are quite similar for the AMI dataset, showing little differences. In contrast, when it comes to the AWE dataset, Mean-CAM-CNN frequently achieves better results than ResNet-50. The AMI dataset scored a Rank-1 identification accuracy of 99.67%, whereas the AWE dataset obtained a Rank-1 identification accuracy of 74.50%. The most effective setup, which produces the highest Rank-1 recognition rates, is the Mean-CAM-CNN with a  $\tau$  value of 0.5. This setup exhibits exceptional performance on both datasets and all used performance metrics. Compared to the traditional ResNet-50 model, the Mean-CAM-CNN achieved higher Rank-1 recognition rates. Specifically, there was a 1.01% improvement for the AMI dataset and a significant 16.75% increase for the challenging AWE dataset. The remarkable success of Mean-CAM-CNN, especially on the challenging AWE dataset, may be due to its ability to selectively direct the CNN’s attention towards essential information, while ignoring irrelevant aspects..

In order to demonstrate the efficacy of the Mean-CAM approach in directing the CNN to

prioritize relevant input, we provide visual examples in Table 5.3. The results demonstrate the substantial influence of Mean-CAM in mitigating misclassification caused by the baseline model. In addition, Mean-CAM improves the accuracy of predicting the proper class. In the first AWE picture shown in Table 5.3, Mean-CAM improved the accuracy of the target class by removing unnecessary areas. It also increased the projected probability of the right class to 99.14%. In addition, in the first AMI, the baseline model accurately classified the class. However, Mean-CAM increased the prediction probability from 56.69% to 90.10%. This enhancement highlights the effectiveness of our suggested Mean-CAM architecture.

Table 5.3: Visualisation and performance analysis of model predictions: a comparative study using the Mean-CAM technique on ear images (B/L pred denotes baseline prediction, and P represents probability).

Dataset	Origin	Results	B/L visualisation	mean-CAM visualisation
AMI		Input class: 046 B/L pred: 046 P= 56.69% mean-CAM pred: 046 P= 90.10%		
		Input class: 088 B/L pred: 074 P= 33.27% mean-CAM pred: 088 P= 99.83%		
		Input class: 095 B/L pred: 041 P= 92.51% mean-CAM pred: 095 P= 99.14%		
AWE		Input class: 008 B/L pred: 065 P= 79.10% mean-CAM pred: 008 P= 98.96%		

Our proposed approach is based on the use of a global stage to extract salient features. Specifically, the Mean-CAM method is used to delineate a region of interest in the image by using average heatmaps obtained from the images of the same class. Each image within a class plays a crucial role in creating the corresponding RoI by considering the average of all extracted salient maps. Refer to Figures 4.5 and 4.6 for visual representation. The aggregate of all training images of the same class may focus our model’s attention on the

most frequently relevant area and divert it away from the irrelevant regions. As a result, this strategy decreases the probability of incorrect classification by reducing the intra-class variability. This, in turn, makes the process of extracting features less affected by the differences in data inside a class.

### 5.3.4 Experiment 3

In the last experiment, we aimed to determine the possible improvements in recognition performance by using the suggested preprocessing methods based on the trained DCGAN model. The main objective of this experiment was to assess the effectiveness of the preprocessing strategy in enhancing the performance of the Mean-CAM-CNN model. In order to examine this, we created new versions of the AMI and AWE datasets. These datasets were then subjected to preprocessing techniques, including colourisation and enhancement, using the trained DCGAN model. Afterwards, we used the Mean-CAM-CNN architecture on the newly generated datasets. The outputs of recognition, with and without preprocessing, are shown in Table 5.4, demonstrating the findings for the three performance criteria that were previously set. The results demonstrate that the suggested preprocessing has a beneficial effect on the accuracy of recognition, as seen by the enhancements in recognition rates. More precisely, the Rank-1 recognition rates improved from 99.67% to 100.00% for AMI and from 74.50% to 76.25% for AWE.

Table 5.4: Comparative analysis of ear recognition outcomes: Evaluating the influence of preprocessing on Mean-CAM-CNN performance.

	AMI			AWE		
	Rank-1 (%)	Rank-5 (%)	AUC (%)	Rank-1 (%)	Rank-5 (%)	AUC (%)
Without preprocessing	99.67	<u>100.00</u>	<u>100.00</u>	74.50	89.50	98.93
With preprocessing	<u>100.00</u>	<u>100.00</u>	<u>100.00</u>	<u>76.25</u>	<u>91.25</u>	<u>99.96</u>

## 5.4 Comparison

To undertake a thorough evaluation, we performed a comparative study of the created DCGAN + Mean-CAM-CNN methodology in comparison to known ear recognition approaches. The findings of this comparison are elaborated in Table 5.5. The key assessment parameter for both the AMI and AWE datasets was the Rank-1 recognition rate, using preset protocols specific to each dataset.

The comparison demonstrates that the analysed papers, such as [38, 40, 47], achieve Rank-1 recognition rates that are adequate, reaching 93% with the AMI dataset. In contrast, our method demonstrates its highest and most competitive efficiency, with a Rank-1 identification rate of 100.00%. The positive outcomes achieved by several articles using the AMI dataset may be ascribed to the dataset's simple and limited image characteristics. On the other hand, most of the articles being compared, whether they use handcrafted or deep learning approaches, encounter significant challenges when working with the AWE dataset because of its intricate and unconstrained image characteristics. Nevertheless, our suggested method showcases very satisfactory and competitive outcomes, achieving a Rank-1 recognition rate of 76.25% using the AWE dataset. The Mean-CAM technique allows the CNN to concentrate only on significant information for recognition, excluding unimportant aspects. This is achieved by using the colourisation of greyscale images and improving the visibility of dark shades using the trained DCGAN model.



Table 5.5: Comparative analysis of Rank-1 recognition rates: evaluating the DCGAN + Mean-CAMCNN approach against competing methods on AMI and AWE datasets.

Approach	Publication	Year	Method	AMI	AWE
<b>Handcrafted</b>	Hassaballah et al.	2019	LBP’s variants	73.71	49.60
	Hassaballah et al.	2020	RLOP	72.29	54.10
	Sarangi et al.	2020	Jaya algorithm + SURF	/	44.00
	Sajadi and Fathi	2020	GZ + LPQ	/	53.50
	Khaldi and Benzaoui	2020	BSIF	/	44.53
	Regouid et al.	2022	1D multi-resolution LBP	100.00	43.00
<b>Deep-learning</b>	Alshazly et al.	2019	ResNet18 Fine-tuned	93.96	/
	Alshazly et al.	2019	VGG-face	94.50	/
	Priyadharshini et al.	2020	AlexNet (Fine Tuning)	96.99	/
	Khaldi and Benzaoui.	2021	VGG-13-16-19-ensembles	96.00	50.53
	Alshazly et al.	2021	MAML + CNN	99.64	67.25
	Omara et al.	2021	CNN	97.80	/
	Sharkas	2022	DCGAN + VGG16	99.45	/
	Xu et al.	2023	VGG16 + DUAL	99.70	72.70
	Aiadi et al.	2023	MDFNet	97.67	/
	<b>Our proposed method</b>	<b>2024</b>	<b>Mean-CAM-CNN</b>	<b>100.00</b>	<b>76.25</b>

## 5.5 Conclusion

This chapter introduced a practical method that combines DCGAN and Mean-CAM-CNN to tackle the complex task of ear biometric recognition. The recognition process consists of two main steps. A trained DCGAN model is used in the early phase to add colour and improve the quality of ear images. Following that, a Mean-CAM-CNN architecture is shown to extract prominent characteristics and perform effective classification. More precisely, the Mean-CAM approach is used to define a specific area in the image by using average heat maps obtained from images belonging to the same class. Subsequently, a CNN is exclusively trained using the obtained areas to classify and recognize ear images. The suggested technique was tested on unconstrained ear recognition databases, namely AMI and AWE, via comprehensive

experiments to validate its effectiveness. The highest Rank-1 recognition rates obtained were 100.00% for AMI and 76.25% for AWE. Specifically, Mean-CAM improved the Rank-1 recognition rates by 1.01% for the AMI dataset and 16.75% for the AWE dataset. On the other hand, DCGAN increased the Rank-1 recognition rates by 0.33% for the AMI dataset and 1.75% for the AWE dataset. The proposed framework's added value is justified by the superior learning capabilities of CNNs, the effectiveness of the preprocessing technique used, the removal of unnecessary information, and the selective extraction of relevant information for recognition using the Mean-CAM cropping strategy.

# General conclusion

Biometric identification, which relies on biological features such as physiological or behavioural traits, has been widely recognised as the most practical and efficient method for accurately and quickly identifying individuals. However, biometric systems that rely on physiological parameters are known for their great dependability since they are resistant to the negative effects of stress and remain reasonably steady over an individual's lifetime. Scientific studies have revealed novel biometric methods for personal identification, including the analysis of ear shape. Several research have been performed to investigate the distinctive features of human ears as a compelling alternative or supplement to conventional biometrics. Ears, in comparison to traditional biometric methods like faces and fingerprints, possess distinct characteristics and are regarded as a valuable resource for identifying individuals. The human ear possesses an extensive collection of textural attributes, exhibiting durability over extended times, resilience to external factors like ageing and facial emotions, and little intrusiveness and sensitivity for collection. Consequently, we aimed to assess our approach by utilising ear datasets.

In this thesis, our main focus was on providing innovative concepts and strategies to improve recognition tasks. Although the proposed solutions surpassed the current leading methodologies, there are still some complex challenges that need further refinement.

The thesis has focused on the following key issues:

- 1) The exceptional level of recognition that can be achieved with deep CNNs has not prevented them from being criticized as black boxes. In order to achieve greater reliability in recognition, we aimed to explore and understand how deep learning models work, and to determine which part of the image under consideration contains features useful for recognition. In order to ensure meaningful information for training, we proposed a new framework based on class activation maps (CAMs), called Mean-CAMs. It extracts a region of interest (RoI) from the image based on the mean heat maps derived from the same class images. The Mean-CAM framework enables high-resolution visualisations with excellent class-discrimination powers and creates common discriminant points between class images.
- 2) We have proposed a pragmatic method that utilises the integration of DCGAN and Mean-CAM-CNN to tackle the complex task of ear biometric recognition. The recognition process consists of two main steps. At the outset, a proficient DCGAN model is utilised to add colour and improve the quality of ear pictures. Next, a Mean-CAM-CNN architecture is shown to extract prominent features and perform effective classification. More precisely, the Mean-CAM approach is utilised to define a specific area of interest

in the image by using average heat maps generated from images belonging to the same class.

In forthcoming endeavours, our goal in future projects is to improve the effectiveness of our study in reliably identifying persons from ear images taken in uncontrolled conditions. This will be accomplished by thoroughly examining many crucial domains:

- Firstly, we aim to examine several methods for visualizing features, including t-distributed stochastic neighbour embedding (t-SNE). This study will provide us with a more profound understanding of the distinguishing characteristics present in ear images.
- Secondly, our strategy involves simultaneously evaluating several CNN architectures. This comparison analysis will allow us to choose the most efficient CNN architecture for our particular purpose, therefore enhancing the performance of our system.
- Lastly, we will examine the possible synergies that may be achieved by combining deep-learned features with handcrafted features. This involves examining combinations such as local binary patterns, robust local oriented patterns, and local phase quantization. By including these various sorts of characteristics, our objective is to use the advantages of both deep learning and conventional feature engineering methods, leading to a more resilient and precise identification system.

# Bibliography

- [1] Anil K Jain, Patrick Flynn, and Arun A Ross. *Handbook of biometrics*. Springer Science & Business Media, 2007.
- [2] Quang Nhat Tran, Benjamin P Turnbull, and Jiankun Hu. Biometrics and privacy-preservation: How do they evolve? *IEEE Open Journal of the Computer Society*, 2:179–191, 2021.
- [3] Navdeep Kaur et al. A study of biometric identification and verification system. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 60–64. IEEE, 2021.
- [4] Diptadeep Addy and Poulami Bala. Physical access control based on biometrics and gsm. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1995–2001. IEEE, 2016.
- [5] Souhail Guennouni, Anass Mansouri, and Ali Ahaitouf. Biometric systems and their applications. In *Visual impairment and blindness-what we know and what we have to know*. IntechOpen, 2019.
- [6] Marcela Hernandez-de Menendez, Ruben Morales-Menendez, Carlos A Escobar, and Jorge Arinez. Biometric applications in education. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 15:365–380, 2021.
- [7] Hugo Gamboa and Ana Fred. A behavioral biometric system based on human-computer interaction. In *Biometric Technology for Human Identification*, volume 5404, pages 381–392. SPIE, 2004.
- [8] Abhijyot Ahire, Aditya Jambhale, Tanuj Patil, Manish Chavan, Amit Nerurkar, and Rugved V Deolekar. Comparative analysis of biometric systems. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 895–901. IEEE, 2019.

- [9] E Cherrat, R Alaoui, and H Bouzahir. Système d'identification biométrique par fusion multimodale. –, 2019.
- [10] F Barbuceanu and C Antonya. Eye tracking applications. *Bulletin of the Transilvania University of Brasov. Engineering Sciences. Series I*, 2:17, 2009.
- [11] Romain Giot, Mohamad El-Abed, and Christophe Rosenberger. Fast computation of the performance evaluation of biometric systems: Application to multibiometrics. *Future Generation Computer Systems*, 29(3):788–799, 2013.
- [12] T Sabhanayagam, V Prasanna Venkatesan, and K Senthamarai Kannan. A comprehensive survey on various biometric systems. *International Journal of Applied Engineering Research*, 13(5):2276–2297, 2018.
- [13] Tanurup Das, Abhimanyu Harshey, Vindresh Mishra, and Ankit Srivastava. An introduction to biometric authentication systems. In *Textbook of Forensic Science*, pages 805–839. Springer, 2023.
- [14] Neil Yager and Adnan Amin. Fingerprint classification: a review. *Pattern Analysis and Applications*, 7:77–93, 2004.
- [15] Shaveta Dargan and Munish Kumar. A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. *Expert Systems with Applications*, 143:113114, 2020.
- [16] Insaf Adjabi, Abdeldjalil Ouahabi, Amir Benzaoui, and Abdelmalik Taleb-Ahmed. Past, present, and future of face recognition: A review. *Electronics*, 9(8):1188, 2020.
- [17] Shabab Bazrafkan, Shejin Thavalengal, and Peter Corcoran. An end to end deep neural network for iris segmentation in unconstrained scenarios. *Neural Networks*, 106:79–95, 2018.
- [18] Amir Benzaoui, Yacine Khaldi, Rafik Bouaouina, Nadia Amrouni, Hammam Alshazly, and Abdeldjalil Ouahabi. A comprehensive survey on ear recognition: databases, approaches, comparative analysis, and open challenges. *Neurocomputing*, 2023.
- [19] Yacine Khaldi, Amir Benzaoui, Abdeldjalil Ouahabi, Sebastien Jacques, and Abdelmalik Taleb-Ahmed. Ear recognition based on deep unsupervised active learning. *IEEE Sensors Journal*, 21(18):20704–20713, 2021.

- [20] Eric Kukula and Stephen Elliott. Implementation of hand geometry at purdue university’s recreational center: an analysis of user perspectives and system performance. In *Proceedings 39th Annual 2005 International Carnahan Conference on Security Technology*, pages 83–88. IEEE, 2005.
- [21] Munish Kumar, Navdeep Singh, Ravinder Kumar, Shubham Goel, and Krishan Kumar. Gait recognition based on vision systems: A systematic survey. *Journal of Visual Communication and Image Representation*, 75:103052, 2021.
- [22] Marcos Faundez-Zanuy. On-line signature recognition based on vq-dtw. *Pattern Recognition*, 40(3):981–992, 2007.
- [23] Marcos Ortega, C Marino, MG Penedo, M Blanco, and F Gonzalez. Biometric authentication using digital retinal images. In *Proceedings of the 5th WSEAS international conference on Applied computer science*, pages 422–427. Citeseer, 2006.
- [24] Anil K Jain and Ajay Kumar. Biometric recognition: an overview. *Second generation biometrics: The ethical, legal and social context*, pages 49–79, 2012.
- [25] Chiarella Sforza, Gaia Grandi, Miriam Binelli, Davide G Tommasi, Riccardo Rosati, and Virgilio F Ferrario. Age-and sex-related changes in the normal human ear. *Forensic science international*, 187(1-3):110–e1, 2009.
- [26] Oyediran George Oyebiyi, Adebayo Abayomi-Alli, Oluwasefunmi ‘Tale Arogundade, Atika Qazi, Agbotiname Lucky Imoize, and Joseph Bamidele Awotunde. A systematic literature review on human ear biometrics: Approaches, algorithms, and trend in the last decade. *Information*, 14(3):192, 2023.
- [27] Walid Hariri. Efficient masked face recognition method during the covid-19 pandemic. *Signal, image and video processing*, 16(3):605–612, 2022.
- [28] Anika Pflug and Christoph Busch. Ear biometrics: a survey of detection, feature extraction and recognition methods. *IET biometrics*, 1(2):114–129, 2012.
- [29] Mark Burge and Wilhelm Burger. Ear biometrics. *Biometrics: personal identification in networked society*, pages 273–285, 1996.
- [30] Aman Kamboj, Rajneesh Rani, and Aditya Nigam. A comprehensive survey and deep learning-based approach for human recognition using ear biometric. *The Visual Computer*, 38(7):2383–2416, 2022.



- [31] Ramar Ahila Priyadharshini, Selvaraj Arivazhagan, and Madakannu Arun. A deep learning approach for person identification using ear biometrics. *Applied intelligence*, 51(4):2161–2172, 2021.
- [32] Li Yuan and Zhichun Mu. Ear recognition based on gabor features and kfda. *The Scientific World Journal*, 2014, 2014.
- [33] M Hassaballah, Hammam A Alshazly, and Abdelmgeid A Ali. Ear recognition using local binary patterns: A comparative experimental study. *Expert Systems with Applications*, 118:182–200, 2019.
- [34] Mahmoud Hassaballah, Hammam A Alshazly, and Abdelmgeid A Ali. Robust local oriented patterns for ear recognition. *Multimedia Tools and Applications*, 79(41):31183–31204, 2020.
- [35] Partha Pratim Sarangi, Bhabani Shankar Prasad Mishra, Satchidanand Dehuri, and Sung-Bae Cho. An evaluation of ear biometric system based on enhanced jaya algorithm and surf descriptors. *Evolutionary Intelligence*, 13(3):443–461, 2020.
- [36] Shabbou Sajadi and Abdolhossein Fathi. Genetic algorithm based local and global spectral features extraction for ear recognition. *Expert Systems with Applications*, 159:113639, 2020.
- [37] Sana Boujnah, Sami Jaballah, Mohamed Ali Mahjoub, and Mohamed Lassaad Ammari. Ear recognition in degraded conditions based on spectral saliency: smart home access. *Journal of Electronic Imaging*, 29(2):023024–023024, 2020.
- [38] Meryem Regouid, Mohamed Touahria, Mohamed Benouis, Lotfi Mostefai, and Imane Lamiche. Comparative study of 1d-local descriptors for ear biometric system. *Multimedia Tools and Applications*, 81(20):29477–29503, 2022.
- [39] Aicha Korichi, Sihem Slatnia, and Oussama Aiadi. Tr-icanet: A fast unsupervised deep-learning-based scheme for unconstrained ear recognition. *Arabian Journal for Science and Engineering*, 47(8):9887–9898, 2022.
- [40] Hammam Alshazly, Christoph Linse, Erhardt Barth, and Thomas Martinetz. Hand-crafted versus cnn features for ear recognition. *Symmetry*, 11(12):1493, 2019.

- [41] 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.
- [42] Yacine Khaldi and Amir Benzaoui. A new framework for grayscale ear images recognition using generative adversarial networks under unconstrained conditions. *Evolving Systems*, 12(4):923–934, 2021.
- [43] Hammam Alshazly, Christoph Linse, Erhardt Barth, Sahar Ahmed Idris, and Thomas Martinetz. Towards explainable ear recognition systems using deep residual networks. *IEEE Access*, 9:122254–122273, 2021.
- [44] Ibrahim Omara, Ahmed Hagag, Guangzhi Ma, Fathi E Abd El-Samie, and Enmin Song. A novel approach for ear recognition: learning mahalanobis distance features from deep cnns. *Machine Vision and Applications*, 32:1–14, 2021.
- [45] Maha Sharkas. Ear recognition with ensemble classifiers; a deep learning approach. *Multimedia Tools and Applications*, 81(30):43919–43945, 2022.
- [46] Xuebin Xu, Yibiao Liu, Chenguang Liu, and Longbin Lu. A feature fusion human ear recognition method based on channel features and dynamic convolution. *Symmetry*, 15(7):1454, 2023.
- [47] Oussama Aiadi, Belal Khaldi, and Cheraa Saadeddine. Mdfnet: An unsupervised lightweight network for ear print recognition. *Journal of Ambient Intelligence and Humanized Computing*, 14(10):13773–13786, 2023.
- [48] 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.
- [49] Peter Claes, Jonas Reijniers, Mark D Shriver, Jonatan Snyders, Paul Suetens, Joachim Nielandt, Guy De Tré, and Dirk Vandermeulen. An investigation of matching symmetry in the human pinnae with possible implications for 3d ear recognition and sound localization. *Journal of anatomy*, 226(1):60–72, 2015.
- [50] I Toprak and Önsen Toygar. Detection of spoofing attacks for ear biometrics through image quality assessment and deep learning. *Expert Systems with Applications*, 172:114600, 2021.

- [51] Žiga Emeršič, Vitomir Štruc, and Peter Peer. Ear recognition: More than a survey. *Neurocomputing*, 255:26–39, 2017.
- [52] Grega Dvoršak, Ankita Dwivedi, Vitomir Štruc, Peter Peer, and Žiga Emeršič. Kinship verification from ear images: An explorative study with deep learning models. In *2022 International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2022.
- [53] Rene Y Choi, Aaron S Coyner, Jayashree Kalpathy-Cramer, Michael F Chiang, and J Peter Campbell. Introduction to machine learning, neural networks, and deep learning. *Translational vision science & technology*, 9(2):14–14, 2020.
- [54] Niklas Kühl, Marc Goutier, Robin Hirt, and Gerhard Satzger. Machine learning in artificial intelligence: Towards a common understanding. *arXiv preprint arXiv:2004.04686*, 2020.
- [55] Josep Lluís Berral-García. When and how to apply statistics, machine learning and deep learning techniques. In *2018 20th International Conference on Transparent Optical Networks (ICTON)*, pages 1–4. IEEE, 2018.
- [56] Bayya Yegnanarayana. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.
- [57] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [58] AD Dongare, RR Kharde, Amit D Kachare, et al. Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, 2(1):189–194, 2012.
- [59] Bin Ding, Huimin Qian, and Jun Zhou. Activation functions and their characteristics in deep neural networks. In *2018 Chinese control and decision conference (CCDC)*, pages 1836–1841. IEEE, 2018.
- [60] Hisham El-Amir and Mahmoud Hamdy. *Deep learning pipeline: building a deep learning model with TensorFlow*. Apress, 2019.
- [61] Sagar Sharma, Simone Sharma, and Anidhya Athaiya. Activation functions in neural networks. *Towards Data Sci*, 6(12):310–316, 2017.
- [62] Chigozie Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*, 2018.

- [63] Jing Li, Ji-hang Cheng, Jing-yuan Shi, and Fei Huang. Brief introduction of back propagation (bp) neural network algorithm and its improvement. In *Advances in Computer Science and Information Engineering: Volume 2*, pages 553–558. Springer, 2012.
- [64] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.
- [65] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377, 2018.
- [66] Jianxin Wu. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, 5(23):495, 2017.
- [67] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9:611–629, 2018.
- [68] Keiron O’shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.
- [69] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [70] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.
- [71] Yann LeCun, Lawrence D Jackel, Léon Bottou, Corinna Cortes, John S Denker, Harris Drucker, Isabelle Guyon, Urs A Muller, Eduard Sackinger, Patrice Simard, et al. Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective*, 261(276):2, 1995.
- [72] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [73] 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.
- [74] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.
- [75] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [76] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
- [77] 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.
- [78] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [79] Ping Kuang, Tingsong Ma, Ziwei Chen, and Fan Li. Image super-resolution with densely connected convolutional networks. *Applied Intelligence*, 49:125–136, 2019.
- [80] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.
- [81] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*, 2016.
- [82] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- [83] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

- [84] Vasant Dhar. The scope and challenges for deep learning, 2015.
- [85] Brinne B Arpteg A and Bosch J Crnkovic-Friis L. Software engineering challenges of deep learning. In *2018 44th euromicro conference on software engineering and advanced applications (SEAA)*, pages 50–59, 2018.
- [86] Sutskever I Krizhevsky A and Hinton G E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84–90, 2017.
- [87] Wang C Zhang G and Grosse R Xu B. Three mechanisms of weight decay regularization. *Communications of the ACM*, pages 1810–12281, 2018.
- [88] Tucker G Pereyra G, Kaiser L Chorowski J, and Hinton G. Regularizing neural networks by penalizing confident output distributions. *Communications of the ACM*, pages 1701–06548, 2017.
- [89] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, Miami, FL, USA, 2009.
- [90] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, Las Vegas, NV, USA, 2016.
- [91] Q. Zheng, M. Yang, X. Tian, N. Jiang, and D. Wang. A full stage data augmentation method in deep convolutional neural network for natural image classification. *Discrete Dynamics in Nature and Society*, 2020:11, 2020.
- [92] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(60):1–48, 2019.
- [93] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv: 1512.03385*, 2015.
- [94] E. Gonzalez, L. Alvarez, and L. Mazon. [AMI Ear Database](http://www.ctim.es/research_works/ami_ear_database). (2008) Available online: [http://www.ctim.es/research\\_works/ami\\_ear\\_database](http://www.ctim.es/research_works/ami_ear_database)., Accessed on 11 Apr 2022, 2008.
- [95] Žiga Emeršič, Vitomir Štruc, and Peter Peer. Ear recognition: More than a survey. *Neurocomputing*, 255:26–39, 2017.

- [96] B. Lavanya, H. Hannah Inbarani, Ahmad Taher Azar, Khaled M. Fouad, Anis Koubaa, Nashwa Ahmad Kamal, and I. Radu Lala. Particle swarm optimization ear identification system. In *Soft Computing Applications*, V. E. Balas, L. C. Jain, M. M. Balas, S. N. Shahbazova (Eds.), pages 372–384, Cham, 2021. Springer International Publishing.
- [97] H. Alshazly, C. Linse, E. Barth, and T. Martinetz. Handcrafted versus cnn features for ear recognition. *Symmetry*, 11(12):1493, 2019.
- [98] H. Alshazly, C. Linse, E. Barth, and T. Martinetz. Ensembles of deep learning models and transfer learning for ear recognition. *Sensors*, 19(19):4139, 2019.
- [99] Jie Zhang, Wen Yu, Xudong Yang, and Fang Deng. Few-shot learning for ear recognition. In *2019 International Conference on Image, Video and Signal Processing*, pages 50–54, Shanghai, China, 2019.
- [100] R. A. Priyadharshini, S. Arivazhagan, and M. Arun. A deep learning approach for person identification using ear biometrics. *Applied Intelligence*, 51(4):2161–2172, 2020.
- [101] Y. Khaldi and A. Benzaoui. A new framework for grayscale ear images recognition using generative adversarial networks under unconstrained conditions. *Evolving Systems*, 12(4):923–934, 2021.
- [102] Y. Khaldi, A. Benzaoui, A. Ouahabi, S. Jacques, and A. Taleb-Ahmed. Ear recognition based on deep unsupervised active learning. *IEEE Sensors Journal*, 21(18):20704–20713, 2021.
- [103] S. Dodge, J. Mounsef, and L. Karam. Unconstrained ear recognition using deep neural networks. *IET Biometrics*, 7(3):207–214, 2018.
- [104] 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*, 42(1):1–20, 2018.
- [105] Y. Khaldi and A. Benzaoui. Region of interest synthesis using image-to-image translation for ear recognition. In *2020 International Conference on Advanced Aspects of Software Engineering (ICAASE)*, Constantine, Algeria, pages 1–6, 2020.
- [106] Y. Zhang, Z. Mu, L. Yuan, and C. Yu. Ear verification under uncontrolled conditions with convolutional neural networks. *IET Biometrics*, 7(3):185–198, 2018.

- [107] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [108] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.