Ministry Of Higher Education And Scientific Research

جامعة أكلي محند أولحاج - البويرة

Université Akli Mohand Oulhadj - Bouira

Faculty of Sciences and Applied Sciences

Department of Computer Science

Doctoral Dissertation

By: Yacine Khaldi

Laboratoire d'informatique, Mathématique et Physique pour l'agriculture et les Foret (LIMPAF)

Biometric Identification using Deep Ear Features

Dissertation Submitted to the Department of Computer Science in Candidacy for the Degree of "*Doctor*" 3rd Cycle LMD in Computer Science.

Examination Committee

President:	Mourad Amad	Professor, University of Bouira
Supervisor:	Amir Benzaoui	MCA, University of Skikda
Examiners:	Bilal Saoud	MCA, University of Bouira
	Abbas Akli	MCA, University of Bouira
	Aissa Boulmerka	MCA, University of Mila

2021/2022

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

Acknowledgments

In the name of Allah, the Most Gracious, the Most Merciful, the Creator of this universe. I am grateful to him, who has blessed us with the ability to do this thesis work.

I am grateful to my family for their kind support, encouragement and dedication throughout their life for making it possible to pursue postgraduate studies. Certainly, without their moral, emotional and financial support, I would not have been able to complete this work.

Thank you, Dr. Amir Benzaoui! You were a constant source of inspiration during the entire thesis process, and I'd like to take this opportunity to thank you. Thank you for your insightful advice, insightful talks, constructive feedback, and patience.

I also thank all the co-authors for their continued kind cooperation during the various stages of the publication of our articles.

To colleagues, friends and doctoral students.

I would also like to acknowledge our dear professors' jury members and all teachers of the Department of Mathematics and Computer Science.

Dedications

I dedicate this work especially to the world dearest persons to me, who inspired me and gave me courage and hope, to my parents.

I dedicate this work with sincere love to my beautiful fiancée.

I dedicate this work with great pleasure and joy my little adorable kids and my dear brothers and sister.

To all my friends.

To all those who were giving me any kind of support, a sincere smile☺.

Abstract

The human ear is one of the important biometric modalities for identifying individuals. It offers a unique benefit over other biometric models, such as the face or eye, beyond just; only the ear can be employed in some circumstances. Unlike the thumbprint or eye, the ear may be enrolled using a conventional camera; however, this has a serious drawback that requires employing ear detection algorithms before ear identification. In a few recent years, ear biometrics gained considerable attention and was addressed via various studies. Many steps of the ear biometric operation have been explored and solved, from ear detection, preparation, extraction of features to verification and identification.

Machine learning techniques have been proved effective in solving different computer vision tasks such as image classification, object detection, and image segmentation. Recently, Deep Learning is a trend artificial intelligence technique that received much attention due to its superiority in solving problems, especially computer-vision-related tasks. State-of-the-art researches on ear detection, identification, or verification used deep learning to complete those tasks and proved that it yielded a better performance against classic machine learning techniques. Thus, we employed deep learning to tackle all problems we identified during our researches.

We proposed a solid experimental work by introducing new approaches to improve the identification process of the ear. The first issue we addressed was the loss of color information from test images, which might have a detrimental impact on the model's performance. A novel system based on image-to-image translation has thus been suggested that can restore missing data. The second issue we worked on was deleting non-ear pixels from photographs and creating a synthetic region of interest of the ear. Last, we proposed a new ear identification method that uses active unsupervised learning, which means that the classification model can learn new information during testing without the need for manual direction, correction, or decision-making. During the testing phase, new information can be used to improve the model's performance. According to obtained results, our proposed approaches were superior to many existing related works.

Keywords:

Biometrics, Ear recognition, Image-to-Image translation, Active learning, Region of Interest

ملخص

تعد الأذن البشرية إحدى طرق المقاييس الحيوية المهمة لتحديد هوية الأفراد. توفر الأذن مزايا فريدة على النماذج البيومترية الأخرى، مثل الوجه أو العين، أبعد من ذلك؛ قد لا يمكن إلا استخدام الأذن فقط في بعض الظروف. أيضا، على عكس بصمة الإبهام أو العين، يمكن تسجيل الأذن باستخدام كاميرا تقليدية؛ ومع ذلك، فإن هذا له عيب خطير يتطلب استخدام خوارزميات الكشف عن الأذن قبل تحديد الأذن. في السنوات القليلة الماضية، اكتسبت القياسات الحيوية للأذن اهتماماً كبيراً وتمت معالجتها من خلال دراسات مختلفة. تم استكشاف وحل العديد من خطوات عملية المقاييس الحيوية للأذن، بدءاً من اكتشاف الأذن، والتحضير، واستخراج الميزات إلى التحقق وتحديد الهوية.

أثبتت تقنيات التعلم الآلي فعاليتها في حل مهام الرؤية الحاسوبية المختلفة مثل تصنيف الصور واكتشاف الأشياء وتجزئة الصور. في الآونة الأخيرة، يعد التعلم العميق أحد أساليب الذكاء الاصطناعي التي حظيت باهتمام كبير نظًرا لتفوقها في حل المشكلات، وخاصة المهام المتعلقة برؤية الكمبيوتر. استخدمت أحدث الأبحاث حول اكتشاف الأذن أو التعرف عليها أو التحقق منها التعلم العميق لإكمال تلك المهام وأثبتت أنها أسفرت عن أداء أفضل مقابل تقنيات التعلم الآلي الكلاسيكية. وبالتالي، استخدمنا التعلم العميق لمعالجة جميع المشكلات التي حددناها خلال أبحاثنا.

اقترحنا عملاً تجريبياً قويًا من خلال إدخال مناهج جديدة لتحسين عملية التعرف على الأذن. كانت المشكلة الأولى التي تناولناها هي فقدان معلومات الألوان من صور الاختبار، والتي قد يكون لها تأثير ضار على أداء النموذج. لذلك تم اقتراح نظام جديد يعتمد على الترجمة من صورة إلى صورة يمكنه استعادة البيانات المفقودة. كانت المشكلة الثانية التي عملنا عليها هي حذف البكسل غير الموجود في الأذن من الصور وإنشاء منطقة اصطناعية ذات أهمية للأذن. أخيراً، اقترحنا طريقة جديدة لتحديد الأذن تستخدم التعلم النشط غير الخاضع للإشراف، مما يعني أن نموذج التصنيف يمكن أن يتعلم معلومات جديدة أثناء الاختبار دون الحاجة إلى التوجيه اليدوي أو التصحيح أو اتخاذ القرار. أثناء مرحلة الاختبار، يمكن استخدام معلومات جديدة لتحسين أداء النموذج. وفقًا للنتائج التي تم الحصول عليها، كانت مناهجنا المقترحة متفوقة على العديد من الأعمال الحالية ذات الصلة.

كلمات مفتاحية:

القياسات الحيوية، التعرف على الأذن، الترجمة من صورة إلى صورة، التعلم النشط، منطقة الاهتمام.

Resumé

L'oreille humaine est l'une des modalités biométriques importantes pour l'identification des individus. Elle offre un avantage unique par rapport aux autres modèles biométriques, tels que le visage ou les yeux, au-delà de seulement ; seule l'oreille peut être employée dans certaines circonstances. Contrairement à l'empreinte digitale ou à l'œil, l'oreille peut être enregistrée à l'aide d'une caméra conventionnelle ; cependant, cela présente un inconvénient sérieux qui nécessite l'utilisation d'algorithmes de détection d'oreille avant l'identification d'oreille. Ces dernières années, la biométrie de l'oreille a suscité une attention considérable et a fait l'objet de diverses études. De nombreuses étapes de l'opération biométrique de l'oreille ont été explorées et résolues, depuis la détection de l'oreille, la préparation, l'extraction des caractéristiques jusqu'à la vérification et l'identification.

Les techniques d'apprentissage automatique se sont avérées efficaces pour résoudre différentes tâches de vision par ordinateur telles que la classification d'images, la détection d'objets et la segmentation d'images. Récemment, le Deep Learning est une technique d'intelligence artificielle tendance qui a reçu beaucoup d'attention en raison de sa supériorité dans la résolution de problèmes, en particulier les tâches liées à la vision par ordinateur. Des recherches de pointe sur la détection, l'identification ou la vérification des oreilles ont utilisé l'apprentissage en profondeur pour effectuer ces tâches et ont prouvé qu'il offrait de meilleures performances par rapport aux techniques classiques d'apprentissage automatique. Ainsi, nous avons utilisé l'apprentissage en profondeur pour résoudre tous les problèmes que nous avons identifiés au cours de nos recherches.

Nous avons proposé un solide travail expérimental en introduisant de nouvelles approches pour améliorer le processus d'identification de l'oreille. Le premier problème que nous avons résolu était la perte d'informations sur les couleurs des images de test, ce qui pourrait avoir un impact négatif sur les performances du modèle. Un nouveau système basé sur la traduction d'image à image a donc été suggéré qui peut restaurer les données manquantes. Le deuxième problème sur lequel nous avons travaillé était de supprimer les pixels non auditifs des photographies et de créer une région synthétique d'intérêt de l'oreille. Enfin, nous avons proposé une nouvelle méthode d'identification de l'oreille qui utilise un apprentissage actif non supervisé, ce qui signifie que le modèle de classification peut apprendre de nouvelles informations pendant les tests sans avoir besoin de direction manuelle, de correction ou de prise de décision. Pendant la phase de test, de nouvelles informations peuvent être utilisées pour améliorer les performances du modèle. Selon les résultats obtenus, nos approches proposées étaient supérieures à de nombreux travaux connexes existants.

Mots clés:

Biométrie, Reconnaissance de l'oreille, Traduction d'image à image, Apprentissage actif, Région d'intérêt

Table of contents

List of figures

List of tables

List of Acronyms

ABC	Artificial Bee Colony
AE	Auto-Encoder
AFIS	Automated Fingerprint Recognition System
AMI dataset	Mathematical Analysis of Images dataset
ANN	Artificial Neural Network
ATM	Automatic Teller Machine
AUCMC	Area under the CMC curve
AWE dataset	Annotated Web Ears dataset
BS	Biometric System
CLBP	Completed Local Binary Pattern
CML	Cooperative Machine Learning
ConvNet	Convolutional Neural Network
DART	Dallas Area Rapid Transit
DBN	Deep Belief Network
DCA	Discriminant Correlation Analysis
DCGAN	Deep Convolutional Generative Adversarial Network
DL	Deep Learning
DNN	Deep artificial Neural Network
DUAL	Deep Unsupervised Active Learning
DeconvNet	De-Convolutional Neural Network
EER	Equal Error Rate
FAR	False Acceptance Rate
FAR	False Acceptance Rate
FC	Full-Connected
FMR	False Match Rate
FNMR	False-Non-Match
FRR	False Rejection Rate
FTA	Failure-To-Acquire
FTE	Failure-To-Enroll
FV_g	Gallery ear Feature Vector
FV_p	Probe ear Feature Vector
GAN	Generative Adversarial Network
GAP	Global Average Pooling
GPU	Graphics Processing Unit
GT	Ground-Truth
HOG	Histogram Of Oriented Gradients
ICA	Independent Component Analysis
IDLLE	Improved Locally Linear Embedding
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
ISO	International Organization for Standardization
IoT	Internet of Things
IoU	Intersection Over Union
K-NN	K Nearest Neighbors
LBP	Local Binary Pattern
LOOP	Local Optimal-Oriented Pattern
LPQ	Local Phase Quantization
LSBP	Local Similarity Binary Pattern
MAE	Mean Absolute Error

ML	Machine Learning
MNN	Micro-Neural Network
MRI	Magnetic Resonance Imaging
NEC	Nippon Electric Company
NIN	Network-In-Network
NIR	Near-Infrared
PCA	Principal Component Analysis
PIN	Personal Identification Number
POEM	Patterns Of Oriented Edge Magnitudes
RGB	Red-Green-Blue
ROC	Receiver Operator Characteristic
RR	Recognition Rate
ReLu	Rectified Linear Units
ResNet	Residual Network
RoI	Region-of-Interest
SIFT	Scale-Invariant Feature Transform
SVM	Support Vector Machine
UERC	Unconstrained Ear Recognition Challenge
USTB	University of Science and Technology Beijing
VGG	Visual Geometry Group
cDCGAN	conditional Deep Convolutional Generative Adversarial Network
tanh	Hyperbolic Tangent Function

CHAPTER 1:
Introduction

1.1 Context

People are generally mobile and constantly linked, and information technology, particularly mobile devices and social networks, has a significant impact on their everyday life. Remote access to intelligent devices provides the majority of services in these civilizations. Banks, e-commerce, public services, hotel bookings, and social aid are just a few of the many services offered by the government, as well as a variety of other fields linked to labor, travel, defense, education, business, and interpersonal relationships. Biometric identity and recognition are critical in the modern world, utilizing its various modalities. The ear is one modality that this research is interested in for various reasons that we will discuss later.

From ear detection to ear identification and verification, ear recognition has gained much attention recently. We advance the wheel of study in this thesis by addressing many issues and impediments in ear recognition and presenting novel ways for enhancing the use of this technology.

Ear biometric systems faces many challenges and difficulties, some of them are similar to the problems in other modalities. These problems still stands despite the amount of research that has been conducted on ear biometrics. For example: illumination, pose, occlusions, image resolution…etc. Based on these conditions, we can categorize ear datasets into two categories: constrained and unconstrained. We focused on unconstrained ear datasets in our research as they make an important challenge to research community.

1.2 Research Objectives

The goals behind this research are:

- To investigate state-of-the-art ear recognition methods and identify their limits and weaknesses in order to propose new approaches.
- To address the problem of grayscale ear images and suggest a solution by introducing a new framework for ear images colorization.
- To propose a new scheme for region-of-interest (RoI) synthesis. The proposed method allows eliminating all non-ear pixels, patching ear missing parts, and removing occlusions by generating synthesized new RoI image.
- To introduce a new active learning-based training scheme that allows ear recognition models to gain additional knowledge beyond the training phase.

1.3 Thesis Overview

The following is the structure of the dissertation:

Chapter 2 provides an overview of biometrics, details the biometric-based recognition method and examines how the biometric system's performance is determined.

Chapter 3 exploits the state-of-the-art of ear recognition. It reviews related researches to the field and categorizes them into four categories depending on the method used, either holistic, geometric, local, or deep learning-based.

Chapter 4 is dedicated to deep learning and its applications. It gives a detailed introduction to machine learning and its branches, such as artificial neural networks and convolutional neural networks and their different architectures.

Chapter 5 tackles the problem of grayscale ear images and how to colorize them in order to enhance the model's performance, especially when it is trained with color images. The performance of the proposed framework is then evaluated and compared to state-of-the-art methods.

Chapter 6 introduces a new method to synthesize region-of-interest of the ear using image-to-image translation to remove irrelevant and non-ear pixels from the image and generate all missing parts of the ear due to occlusions. The proposed method is evaluated and compared to state-of-the-art methods.

Chapter 7 presents a new training scheme called unsupervised active learning for ear recognition. During the proposed learning process, a model can acquire new knowledge continuously during the testing phase and retrain itself in an unsupervised way.

Finally, we end this thesis with conclusions and some perspectives.

1.4 List of Publications

The contributions related to this thesis are:

1- Khaldi, Yacine, 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), DOI: 10.1007/s12530-020-09346-1

2- Khaldi, Yacine, and Amir Benzaoui. "Region of Interest Synthesis using Image-to-Image Translation for ear recognition." In 2020 International Conference on Advanced Aspects of Software Engineering (ICAASE), pp. 1-6. IEEE, DOI: 10.1109/ICAASE51408.2020.9380127

3- Khaldi, Yacine, Amir Benzaoui, Abdeldjalil Ouahabi, Sébastien Jacques, and Abdelmalik Taleb-Ahmed. "Ear recognition based on deep unsupervised active learning." IEEE Sensors Journal, 21(18), 20704-20713, (2021), DOI: 10.1109/JSEN.2021.3100151.

CHAPTER 2: Biometric fundamentals

2.1 Introduction

The importance of biometrics has been increased dramatically since its foundation in the 1800s. Different modalities have been proposed, such as fingerprint, face, iris, ear, etc. Each modality has its pros and cons based on different criteria. This chapter introduces biometrics in detail and all its related issues.

This chapter is organized as follows: In Section 2, biometrics is defined and contextualized. Section 3 focuses on the biometric identification and verification challenges, as well as the performance evaluation metrics relevant to these concerns. The biometric system is discussed in the fourth section. In section 5, we address the most well-known biometric applications and the newest trends. Section 6 outlines evaluation protocols followed by biometric modalities in the seventh section. In section 8, we discuss how to choose between biometric modalities. Section 9 provides the motivations behind this thesis. Finally, the last section concludes this chapter.

2.2 What is biometrics?

Sir Francis Galton founded the biometric approach as a set of statistical methods to study continuing inheritance characteristics and aspects of heredity at the population level. Pearson K. and Weldon W.F.R. then developed this approach with enthusiasm and created *Biometrika* in 1901, a leading journal in the field of biometrics [1].

Currently, biometrics is the automated process of confirming or identifying a person's identity based on specific physiological attributes called *biometric authenticators* [2], such as a fingerprint or face pattern, or specific behavioral characteristics, such as handwriting or walking patterns. A physiologically-based biometric system is more accurate than one that takes on behavioral characteristics, though the latter may be easier to incorporate in particular applications [3].

One of the first questions to ask while learning more about biometrics is: Which modalities are most appropriate for a particular biometric recognition problem? As common sense dictates, an excellent biometric feature must exhibit several characteristics. They are mainly [4]:

- **Universality:** the chosen attribute should be possessed by every individual.
- **Distinctiveness** (also known as individuality or uniqueness)**:** any two individuals should be sufficiently dissimilar to be distinguished by this feature.

- **Permanence** (sometimes called stability or immutability)**:** the feature should maintain a sufficient level of consistency throughout time (concerning the matching criterion).
- **Collectability:** the feature should be observable and quantifiable.

However, in a real-world Biometric System (BS), several additional considerations need to be made [4], including the following points:

- **Performance:** the accuracy of identification and the time required for successful recognition must be acceptable.
- **Acceptance:** individuals should be willing to accept the BS and should not perceive it as invasive, hazardous, or causing discomfort.
- **Circumvention:** the prospect of attacking and duping the BS should be highly remote.

Given the abundance of BS, there is no such thing as a flawless one. Depending on the parameters listed above, every biometric modality has its pros and cons, and each has a place in one or more applications [5]. Based on the performance and effectiveness of available biometric systems, the final identification system will show a reasonable certainty where someone is previously enrolled in the user's database or not [6].

2.3 Identification vs. Verification

Biometric recognition is divided into two distinct approaches for matching freshly collected biometric features: Identification and Verification [7]. When we do not wish to distinguish between them, we shall use the term "Recognition." Nevertheless, other authors consider recognition and identification as equivalent terms.

1. The process of establishing a person's identity is referred to as identification. It entails taking the measured characteristic and searching for a match in a collection of individuals with that trait. In a more broad response, the system will list the database's most similar individuals. If the database is extensive, this method may need significant computing power and time. Identification is used extensively in law enforcement, forensic science, and information gathering for intelligence. The identification rate is used to evaluate the system's performance.
2. Verification (alternatively known as authentication) is the process of determining whether a person is whom he claims to be. The algorithm evaluates the claim and either approves it or rejects it. As a result, the algorithm can produce a confidence level for the claim’s validity concerning the previously established verification threshold. In

general, the method entails comparing the measured feature to previously recorded data for that individual. As a one-to-one comparison is all that is required, this approach uses far less processing power and time than the prior one (whereas identification requires one to N comparisons). Passwords, secret keys, magnetic chips, and PINs are the most often used authentication methods. Verification is commonly used to control locations (i.e., physical) or data (i.e., logical) accesses [8].

2.4 Biometric system

The underlying biometric recognition mechanism is comparable independent of the biometric challenge and the final chosen modality. Thus, they all share the same general configuration regardless of the system, as illustrated in Figure 2.1.

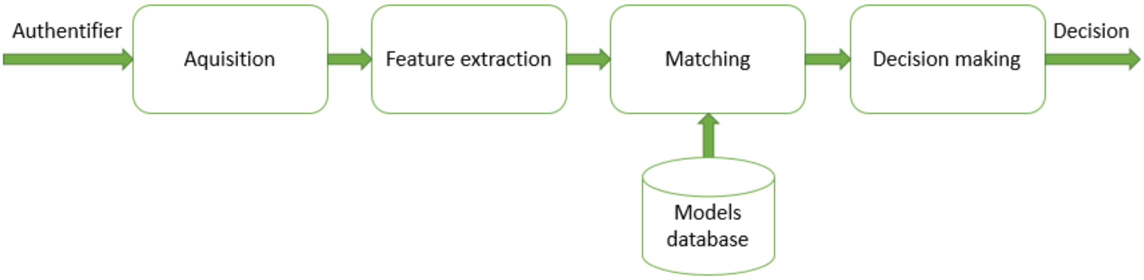


Figure 2.1 Biometric recognition scheme.

The four major steps indicated in the above framework are summarized as follows:

1- Data Collection: A physical or behavioral samples are collected using specialized data collection technology; this is a particularly delicate stage, as most biometric recognition methods are highly dependent on the properties of the gathered data. Thus, if possible, the submitted signal's quality will be verified. A new acquisition will be made if the value is less than a previously determined threshold. Additionally, sufficient samples must be collected to ensure the system's robustness.

2- Feature Extraction: Several digital signal processing techniques extract a set of characteristics from the samples and the user template.

3- Matching: The preceding step's measured parameters are used to create a supplied user model. The entire collection of extracted features is kept and used as a model in enrollment mode. As with people, the system requires a learning phase before recognizing objects. Enrollment is to store a user's characteristics for future usage. After the user data is collected during the enrollment phase, a new sample is taken and compared against all the database's recorded templates for the identification or against the user's template for verification. Considerable distances have been effectively utilized in general ways to accomplish this task.

4- Decision: The system determines if the new sample's collection of extracted features is a match or a miss-match.

2.5 Biometrics applications

The primary advantage of biometric technology is its security. While passwords and Personal Identification Numbers (PIN) are easily obtained, stealing a biometric identification such as a fingerprint or iris scan is incredibly difficult. Because of this mix of security and ease, biometric technology usage will continue to grow in the next years, and biometric security systems will become more prevalent [9, 10].

2.5.1 Law Enforcement

Biometric technology is frequently employed in law enforcement. For over 30 years, Nippon Electric Company (NEC) has worked with law enforcement organizations worldwide, including New Zealand, to provide biometric solutions for identifying criminals.

According to a wired story, government enforcement organizations in the United States have facial recognition data on 117 million Americans. The Home Office in the United Kingdom announced a £26 million investment in police innovation using biometric technologies. They frequently use fingerprints, face, and iris recognition as biometric technology.

2.5.2 Access and Authentication

Smartphone security is likely one of the most prevalent applications for biometric technology in this day and age. In the years following Apple's introduction of Touch ID (i.e., fingerprint recognition system), mobile phone security has expanded to include facial recognition, iris detection, and voice recognition, among other biometric technologies.

2.5.3 Banking

Banking is another industry that incorporates biometrics into a variety of services in order to provide unique experiences for clients. In Japan, seven Banks are experimenting with facial recognition to enhance Automatic Teller Machines (ATM) services. It will be an additional layer of security to verify users. Banks are also employing biometric to strengthen consumer and employee identity management to prevent fraud.

Customers are also concerned about the hassle of needing to authenticate their identities frequently. As a result, an increasing number of clients are looking for banks that use biometric authentication, leading banks to explore and apply the technology.

2.5.4 Home Assistant

The use of voice recognition as a biometric identification is nothing new for anyone familiar with Google Home, Alexa, or Siri. In addition to intelligent lightbulbs, door locks, security cameras, and other IoT devices, the Google Assistant that powers Google Home and the assistant on Android phones and tablets are compatible with a wide range of other Internet of Thing (IoT) devices. Security is a top priority when using your home assistant in conjunction with any of these add-ons. You would not want them in the hands of just anyone; this is a must-have for Google Assistant when it comes to voice recognition.

2.5.5 Public Transport

Security and improving the customer experience are just two of the many possible applications for biometric technology that may be found in public transit, which is still in its infancy in terms of widespread adoption. Dallas Area Rapid Transit (DART), the largest North Texas municipal transit agency, was an early adopter of biometric technology. DART trains now the use of facial recognition cameras.

Smart ID cards and intelligent tickets can also be used with biometric technology to match a person using facial recognition to transport systems, making travel safer and streamlining the ticketing and passenger management processes.

2.6 Performance Evaluation

The performance of a biometric system can also be measured based on other criteria such as-accuracy, efficiency, and the volume of data stored for each person. However, only the accuracy will be assessed while considering the operating mode used, namely: identification and verification. Each of these modes will require different precision measurements.

The evaluation rate is one of the most commonly used measures, but it may be insufficient. Indeed, in the event of an error, it can be useful to know if the correct choice is in the first N. We then plot the cumulative score, which represents the probability that the right choice is among the first N [11].

Precision is the ratio between the number of models correctly found by the system in the database and the total number of models found. The recall is the ratio between the number of models correctly found in the database and the total number of models which should have been found. The type of mistake made by this kind of system is to attribute to the individual presenting an identity other than his own. The performance of this system is measured using the identification rate. This parameter depends on the number of people contained in the database. Indeed, the greater the number of tests, the greater the error rate is likely to be [12].

The International Organization for Standardization (ISO) defined the standard ISO / IEC 19795-1, dividing the error rates into three classes: the fundamental error rates, the error rates of authentication systems, and error rates of identification systems. The fundamental error rates are the following:

1. Failure-to-acquire rate (FTA): proportion of verification or identification attempts for which the biometric system could not acquire the required biometric information;
2. Failure-to-enroll rate (FTE): the proportion of individuals for whom the system could not generate the biometric model during enrollment. Take, for example, the case of fingerprints; some people who do not have fingerprints for genetic reasons, or fingerprints almost nonexistent for medical reasons or very damaged by their profession;
3. False-non-match rate (FNMR): proportion of false rejections by the comparison between the acquired biometric data and the corresponding model;
4. False match rate (FMR): the proportion of false acceptance by comparing the acquired biometric data and the model corresponding to another individual.

2.6.1 Identification Evaluation

The Recognition Rate (RR) can be used to measure the performance of an identification system. RR's information is straightforward: Proportion of previously enrolled participants whose identities were correctly recognized; this is mathematically expressed in Eq. 1.1.

$$RR = \frac{Number\ of\ correctly\ recognized\ images}{Total\ number\ of\ images} \tag{1.1}$$

2.6.2 Verification System Evaluation

To verify the system, we may utilize the False Acceptance Rate (FAR) and False Rejection Rate (FRR) [13]. The True Acceptance Rate (TAR) is also available for positive reasoning. However, it is less frequently used than the FAR index. Both mistakes are detrimental and must be carefully weighed to ensure that the appropriate mix of security measures is obtained. Typically, this required trade-off between them is created by adjusting a threshold. The performance under consideration can be visualized using a Receiver Operator Characteristic (ROC) curve.

2.6.2.1 False Rejection and False Acceptance Error rates

False rejection rate is the proportion of legitimate requester transactions rejected in error. For a single-attempt verification transaction and a fixed threshold at τ (τ depending on the comparison algorithm), the false reject rate is calculated as follows:

$$FRR(\tau) = FTA + FNMR(\tau) * (1 - \mathrm{FTA}) \qquad (1.2)$$

False Acceptance Rate is the proportion of transactions from impostors wrongly accepted. For a single-attempt verification transaction and a fixed threshold at τ, the false acceptance rate is calculated by:

$$FAR(\tau) = FMR(\tau) * (1 - \mathrm{FTA}) \qquad (1.3)$$

The two error rates, FAR and FRR, are related and depend on a decision threshold that is fixed according to the biometric system's security level (high or low). Figure 2.2 shows the theoretical distribution of the likelihood rates of legitimate users and impostors. As we can see, the lower the fixed threshold, the higher the false acceptance rate, which means that the biometric system will accept impostors. Conversely, the higher the threshold, the lower the false acceptance rate; the biometric system will, in this case, be robust to impostors but will reject many more legitimate users.

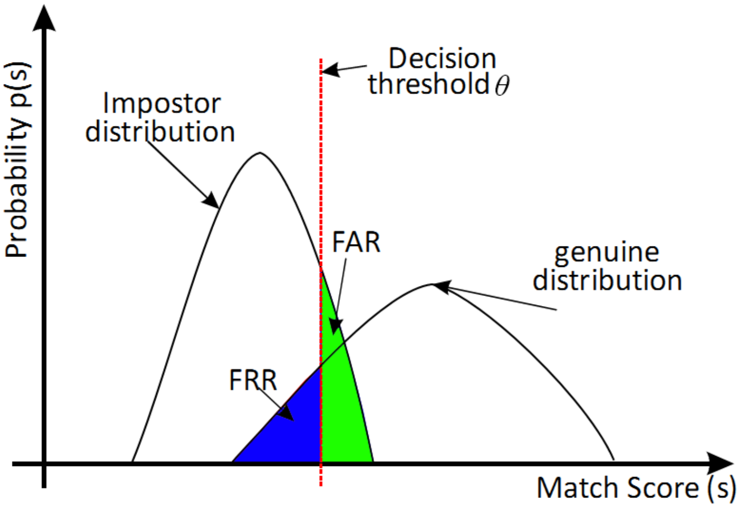


Figure 2.2 Representation of the likelihood rate of legitimate users as well as impostors on a biometric authentication system.

2.6.2.2 ROC curve

This curve is one of the most widely used methods to assess the overall performance of a biometric authentication system. The ROC curve represents the relationship between the false acceptance rate (FAR) and the false rejection rate (FRR) for different decision threshold values. The term DET (Detection Error Tradeoff) is then used; in this case, the term ROC is reserved for the representation of the rate of true rejections (1-FRR) at the FAR. Figure 2.3 illustrates the representation of a ROC curve. The main advantage of this single curve is that one obtains a compact representation of the performance of a biometric system, which allows objectively comparing different biometric systems [14].

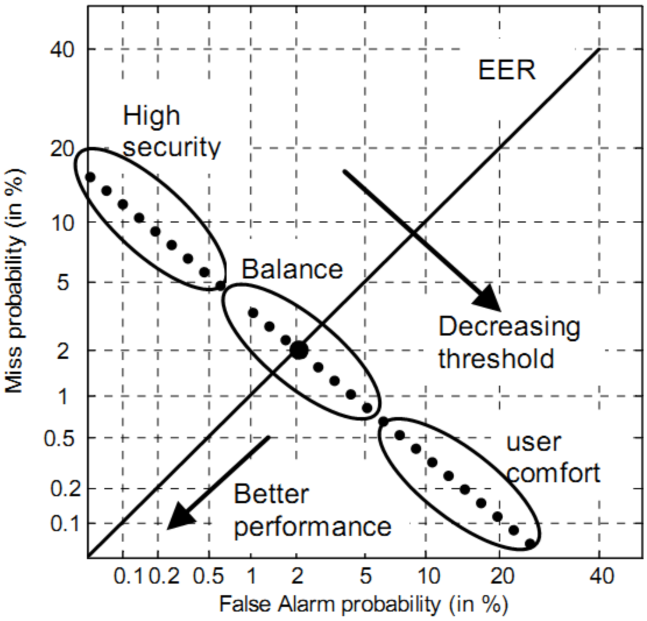


Figure 2.3 False Rejection Rate vs. False Acceptance Rate and ROC curve (dotted line).

Depending on the application for which the biometric verification system is intended, the threshold can be adjusted to provide the desired level of security (low, medium, or high).

2.6.2.3 Equal Error Rate

The Equal Error Rate (EER) is when the FAR and the FRR are equal. The lower the EER threshold, the more precise the system. These characteristics make it easy to compare several systems and thus determine which one best meets the needs [15]. Furthermore, the weaknesses in tampering are not with physical particularity but rather with how they measure it and the margin of error they allow. However, as a single criterion, the EER does not capture all of the system's features.

2.7 Biometric modalities

Biometric systems use a range of biometric modalities, such as the fingerprint, the face, the ear, the iris, the retina, the palm-print, the vein, the voice, the signature, the gait, and the odor, among others. The most widely used biometric modalities are summarized in Figure 2.4.

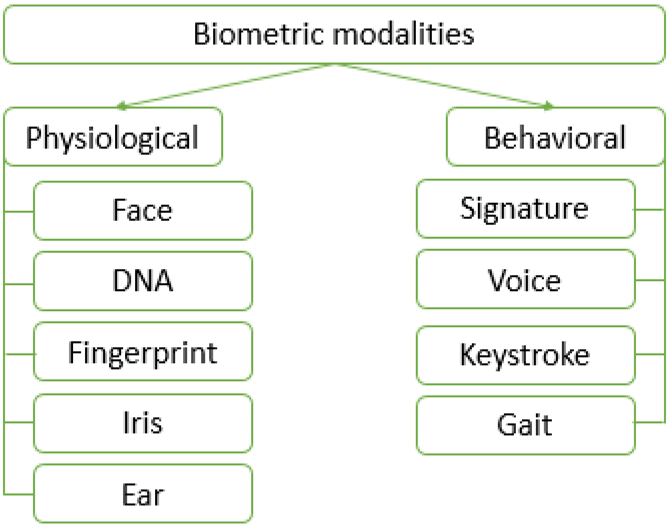


Figure 2.4 Widely used biometric modalities.

2.7.1 Fingerprint

The fingerprint has been the most widely used biometric feature in the world since 1888, when F. Galton discovered the permanence and inalterability of the papillary pattern from birth to death. It has been used for a century for criminal identification. It corresponds to the bulk of the current market, and its use will undoubtedly develop. It has a sufficient reliability rate to allow the identification of individuals in large databases. It has been used for a long time in the police context; it is not always very well accepted by users but presents a good compromise between the constraints of use and the desired reliability [16].

The ridge flow pattern of the fingerprint contains local discontinuities known as terminations and minutiae. The first is described as the point at which a ridge suddenly terminates, while the second is when a ridge splits or deviates into sub ridges; these types of location and orientation are the characteristics that differentiate fingerprints.

Fingerprints are the most common biometric modalities used by police departments for investigating crimes due to their well-known distinctiveness and consistent properties, and their unique ability to leave a copy of it on previously touched surfaces. AFIS (i.e., Automated Fingerprint Recognition System) is the name of these highly mature identifying systems. The AFIS utilized by the Police is depicted in Figure 2.5.

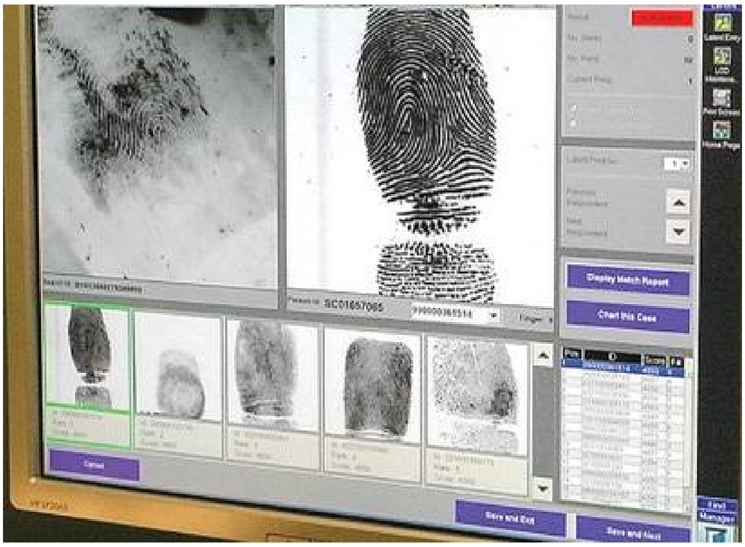
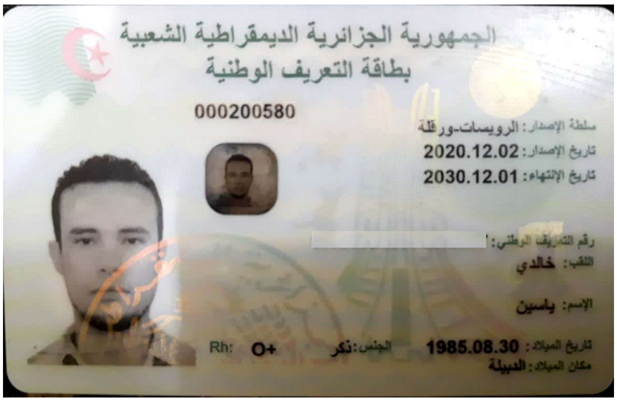


Figure 2.5 AFIS: Police processing of fingerprints in minutes.

2.7.2 Face

Face recognition is among the most promising biometric technologies because it is the most intuitive way for humans to identify each other [17]. In our environment, faces are the most common visual patterns. Thus, it is common for individuals to identify others by their faces. It would not be easy to do so through their card due to the many extant languages, each with its own set of characters. This fact is shown in Figure 2.6.



الجمهورية الجزائرية الديمقراطية الشعبية
بطاقة التعريف الوطنية

000200580

سلطة الإصدار: الرويسات-ورقلة
تاريخ الإصدار: 2020.12.02
تاريخ الإنتهاء: 2030.12.01
رقم التعريف الوطني: [redacted]
اللقب: خالدي
الإسم: ياسين
Rh: O+ الجنس: ذكر تاريخ الميلاد: 1985.08.30
مكان الميلاد: الدبيلة

Figure 2.6 A personal card of an Algerian citizen. Non-Arabic-speaking people can identify the person through the photograph only.

The face is the second most often used biometric modality after fingerprints. One of its benefits is widespread social acceptability among users. Although this technology is most suited to cooperative user applications, interaction with humans is not always necessary throughout the acquisition process.

This last point is especially advantageous for addressing surveillance applications in high-security locations. The significant degree of diversity in faces (expressions, age, lighting circumstances, rotation, changes in appearance, etc.) makes identification difficult. According to comparative research [18, 19], face recognition is not as accurate as fingerprint or iris.

2.7.3 Iris

Several companies develop and sell systems that rely on the human iris to recognize people. The iris is a thin circular diaphragm, which sits between the cornea and the human eye lens. The iris is perforated near its center by a circular opening called a pupil [20]. The iris image is captured by a device that contains an infrared camera when the person stands a short distance from the device. Iris recognition is widely used in identification and verification applications because it is highly distinctive, unique, its shape is stable, and it is protected and very robust. However, acquisition equipment is expensive. Figure 2.7 displays a sample of the iris [21].

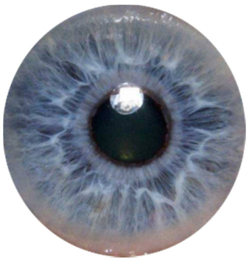


Figure 2.7 Iris pattern.

Iris recognition systems give an aliveness detecting method using the difference in the iris size between two image sequences, similar to how a photography camera's flash works while operating in the dark.

Nonetheless, this is not the only type of light applied directly to the eye. Additionally, a near-infrared (NIR) beam is necessary to highlight and more accurately replicate the texture of the eyeballs. If the naked eye does not detect the light beam and does not respond properly with pupil contraction, ocular damage is a risk [22]. As a result of this and other factors, the populace is uneasy with such arrangements.

As an alternative, retinal scanners that use an infrared light beam to illuminate the retina and scan the pattern of the vascular system are currently being used in contemporary biometric security [23]. Thus, it refers to an invasive procedure and needs a high user involvement. Additionally, this collection procedure may have side impacts on the patients owing to the brief exposure to a low-powered infrared laser beam, which has become the primary impediment for these devices. Due to the high cost of advanced acquisition equipment, only a few firms are developing this technology; the most notable is EyeDentify, the global patent for retinal scanners [24]. Due to the excellent performance of both ocular technologies, military and security applications (research laboratories, nuclear power facilities, and intelligence organizations, for example) prefer them.

2.7.4 Ear

Researchers determined that each individual's ears are unique and distinctive as their fingerprints; no two ears, even on the same person, are identical [25]. Beyond the age of ten years, the ear morphology varies little, and medical research has shown that substantial changes in the ear occur only before the age of eight years and after the age of seventy years. It develops symmetrically and tends to bulge downward with age, but this is a quantifiable impact. According to studies, the ear changes only 1.22 millimeters each year [26]. Additionally, unlike the face, the color distribution of the ear is nearly consistent. The ear is almost precisely in the center of the profile face. From a distance, ear data may be collected without the person being aware. Ear biometrics is an excellent example of passive biometrics since it requires little assistance from the user and satisfies the requirement of the authentication system's confidentiality in the environment. Ear pictures can be collected concurrently with facial images to enhance recognition accuracy [27]. Figure 2.8 illustrates the anatomy of the human ear.

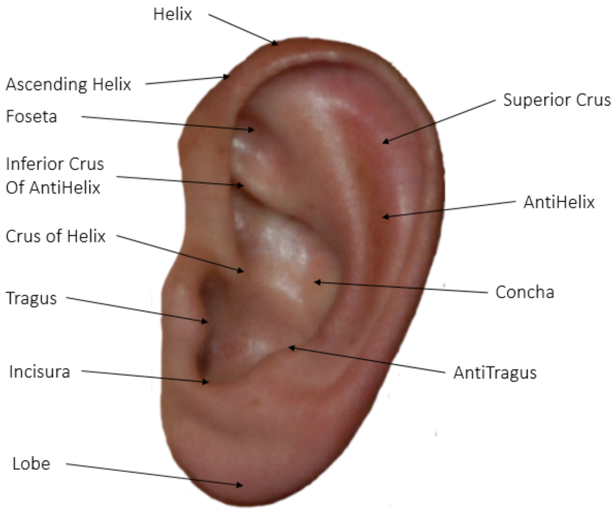


Figure 2.8 Anatomy of the ear (Ear image taken from the AMI dataset[1].

[1] https://ctim.ulpgc.es/research_works/ami_ear_database/

2.7.5 Gait

Gait recognition is a behavioral biometric modality that uses a person's unique walking pattern to identify them. Compared to other first-generation biometric modalities such as fingerprint, gait recognition is less intrusive since it does not need subject interaction [28].

Gait identification is predicated on the premise that each individual has a unique and peculiar manner of walking that is discernible from a biomechanical standpoint. While human movement is composed of coordinated movements of hundreds of muscles and joints, gait varies in time and amplitude between individuals [29]. The complicated biological process of the musculoskeletal system is illustrated in Figure 2.9, which may be split into many sub-events of human gait. The examples depicted in this image are utilized to extract patterns used as an individual's identifying system.

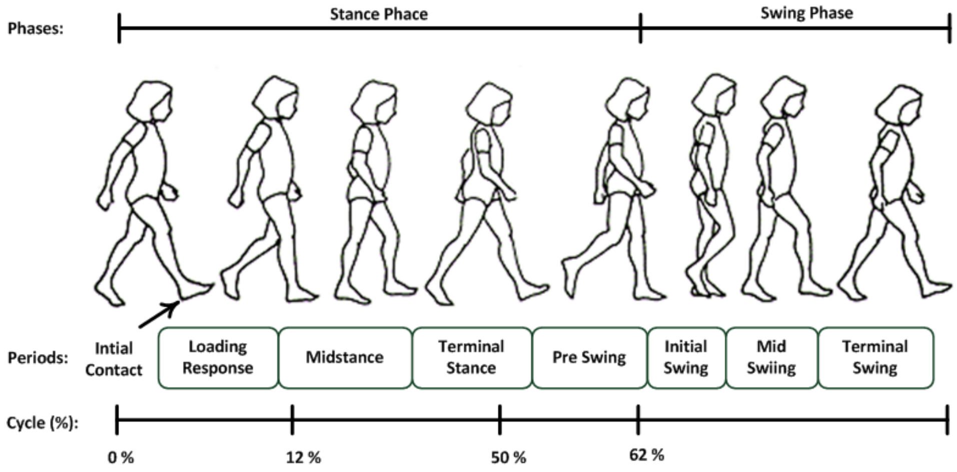


Figure 2.9 The gait cycle divided into two movement phases and five standing phases.

As a result, slight differences in gait style may be utilized to identify individuals using biometrics uniquely. Gait recognition associates spatial-temporal characteristics such as step length, step breadth, walking speed, and cycle time with kinematic variables such as hip and knee rotation, mean hip, knee, ankle joint angles, thigh, trunk, and foot angles. The connection between step length and an individual's height is examined [30].

2.7.6 Hand geometry

Hand geometry is commonly used for physical access control, and timekeeping, especially in some jurisdictions [31]. This form of biometry consists of analyzing 90 characteristics of the hand, including the length and width of the fingers, the palm, the shape of the joints, or the drawing of the lines of the hand. For the capture phase, the person places their hand on a turntable. Then, the positions of the thumb, index, and middle fingers are materialized [32]. An analysis from two different angles is performed to obtain a three-dimensional rendering. Unlike fingerprints, which often still block psychological barriers, the analysis of the hand's shape is much better accepted. Figure 2.10 is an example of a completely functional two-dimensional hand geometry reader device.

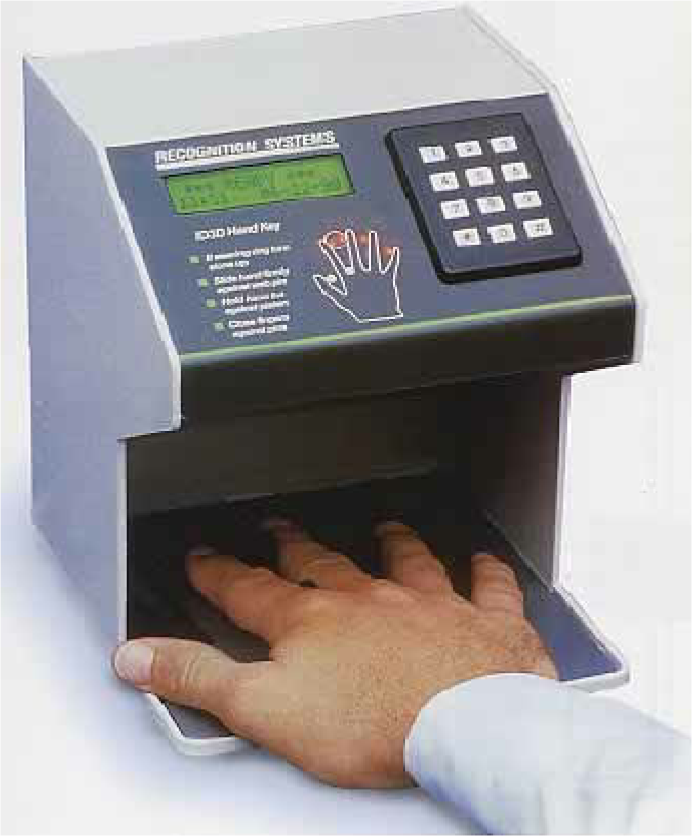


Figure 2.10 Hand geometry reader.

There are verification techniques that need only a few fingers to be measured rather than the whole hand. Although these devices are smaller than hand geometry, they are still significantly more extensive.

2.7.7 Online signature

Signature authentication systems usually include a pen and a digital tablet [33]. The verification is accomplished by analyzing several variables, including the speed with which the signature is made, the accelerations, and the pressure exerted. The difficulty with capturing a signature is that a person never signs the same way twice, even within seconds of each other. Indeed, depending on emotions or fatigue, a signature can change significantly. Hence, the development of very complex algorithms can consider these possible evolutions [34]. Figure 2.11 depicts a tablet in its entirety.

A significant advantage of this technique is that the user may modify this characteristic when needed; this is a unique feature not available with the other biometric authenticators [35].

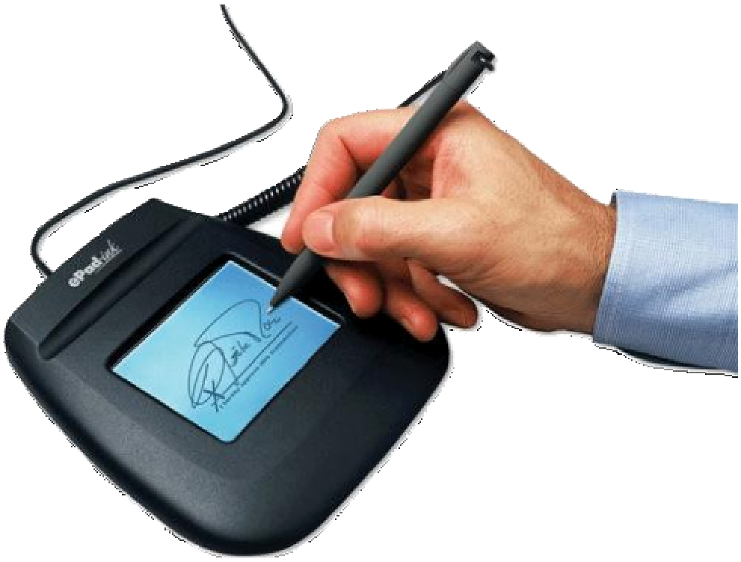


Figure 2.11 Online signature device.

2.8 Biometric modalities… how to choose?

Human characteristics can be used for biometrics in terms of the following parameters:

1. Universality: each person should have the characteristic.
2. Uniqueness: can the biometric separate one individual from another.
3. Permanence: measures how well a biometric resists aging.
4. Collectability: whether a biometric can be measured quantitatively.
5. Performance: accuracy, speed, and robustness of technology used.
6. Acceptability: degree of approval of a technology.
7. Circumvention: ease of use of a substitute.

Table 2.1 shows the most popular biometric modalities' property levels (high, medium, low). It is essential to consider these properties before deciding which modality to use in specific scenarios.

Table 2.1 Properties of biometric modalities [36].
(H: High, M: Medium, L: Low)

Modality	Universality	Uniqueness	Permanence	Collectability	Performance	Acceptability	Circumvention
Face	H	L	M	H	L	H	L
Fingerprint	M	H	H	M	H	M	H
Hand geometry	M	M	M	H	M	M	M
Keystrokes	L	L	L	M	L	M	M
Hand veins	M	M	M	M	M	M	H
Iris	H	H	H	M	H	L	H
Retinal scan	H	H	M	L	H	L	H
Signature	L	L	L	H	L	H	L
Voice	M	L	L	M	L	H	L
Facial thermograph	H	H	L	H	M	H	H
Odor	H	H	H	L	L	M	L
DNA	H	H	H	L	H	L	L
Gait	M	L	L	H	L	H	M
Ear	M	M	H	M	M	H	M

2.9 Why we must have ear biometric?

The human ear is a relatively new modality in biometrics; indeed, there is no commercial software that utilizes this modality at the moment. It is considered one of the most constant anatomical characteristics in humans. According to embryological research, significant changes in the ear shape occur before the age of eight and beyond the age of seventy years [37, 38, 39]. Compared to other modalities, most notably the face, a person's ear does not alter significantly over time, but the face does. The features of the face can be altered via the use of cosmetics, hairstyles, and haircuts. Additionally, human faces vary in response to emotions and expressions such as grief, pleasure, fear, or surprise.

On the other hand, the ear's features are primarily fixed and unaffected by emotions. Unlike facial identification systems, glasses, beards, or mustaches cannot be used to conceal ear pictures during the collection process. However, partial occlusion can be achieved by the presence of hair or curls. It is critical to highlight that the public has a high level of acceptance for ear modality in access control and security applications such as visa and passport programs. Additionally, there is no requirement to contact the sensor, eliminating the hygiene issue; picture capture may be accomplished quietly from a distance and does not require user-sensor interaction.

Additionally, ear pictures are more secure than face photographs, as visually associating an ear image with a particular individual is difficult (most users cannot identify their ear images). Thus, databases holding ear photos should not be safer than databases including face images, as the danger of assault is more significant in the latter scenario. Ear pictures can be collected concurrently with facial photos, increasing recognition accuracy.

Traditionally, ear recognition was performed using classic Machine Learning (ML) algorithms. Although high results were obtained for some constrained databases, those algorithms gave minimal effects on unconstrained databases. Hence comes the need for more advanced ML algorithms such as Deep Learning. Deep learning has been used widely for different image classification tasks, including biometric recognition and identification. Thus, we proposed several promising biometric approaches based on deep learning.

2.10 Conclusion

Biometrics technology attempts to replicate the pattern recognition process of the human brain to identify individuals. It is a secure and dependable authentication system than the traditional secret-based and token-based authentication schemes. Biometrics technologies automate the recognition of persons based on their physiological and behavioral modalities. These qualities must meet specific criteria, including universality and performance.

The recognition process is divided into two stages: Enrolment and Matching. The first phase seeks to teach the system about the person's identification. It begins by extracting some features from the captured data to create a referential database. The template is a representative framework that effectively summarizes the unique biometric features of each individual. The second phase, matching, retrieves the previously saved template to compare it to the newly retrieved characteristics.

We presented the existing methods allowing us to evaluate biometric systems. Due to the vast range between classes and the slight variation within classes in some biometric samples, the system's judgment may be incorrect. Traditionally, the performance of a biometric system has been quantified using two error rates: FRR and FAR. The first triggers when a system rejects a real identity, whereas the second occurs when an impostor identity is erroneously accepted. Equal ERR is a trade-off between these two errors.

It is widely thought that no biometric modality can be perfect; nevertheless, combining numerous biometric traits into a hybrid recognition system can significantly improve the recognition process and therefore enhance performance.

Recent years have seen a surge in interest and concentration on ear biometrics due to several studies. Numerous phases of the ear biometric process have been addressed and studied, ranging from ear detection to verification and identification. The ear, unlike the face, is unaffected by age or expression. Furthermore, ear image capture does not require expensive cameras or scanners, as fingerprint or iris image acquisition does. As a result, we picked the ear as a biometric modality. In the next chapter, we will study in detail the fundamentals of the ear as a biometric modality.

CHAPTER 3: Ear Biometrics

3.1 Introduction

Ear biometrics recently gained an essential quantity of attention due to its importance and advantages over other modalities. In this chapter, we present ear biometrics in detail. Most biometric recognition systems that use two-dimensional ear images begin with feature extraction and then compare the derived vector to the enrolled models; this categorizes ear recognition approaches into four distinct categories: holistic, geometric, local, and deep neural networks (DNNs). This chapter discusses the most recent approaches to ear recognition and categorizes them according to the method utilized.

This chapter is organized as follows: section 2 introduces ear biometrics, followed by ear recognition system presentation in section 3. In section 4, we introduce and categorize ear recognition methods. Section 5 introduces holistic methods, and section 6 introduces geometric approaches. In section 7, we review local methods. In section 8, we review state-of-the-art deep-learning-based methods. Finally, we conclude this chapter in section 9.

3.2 Ear biometrics, why?

Why must we use ear biometrics? This question must be answered before proceeding to the rest of the chapter. Many problems in other modalities, such as face recognition, remain largely unsolved. Some of these problems are:

- A wide variety of imaging problems.
- The face is the most changing part of the body.
- Facial expression, cosmetics, plastic surgery.
- Some modalities require expensive equipment, such as the fingerprint.
- The emission of infrared rays or other kinds of rays can disturb people.

Unlike the face, fingerprint, or iris scanning, the ear biometric system feels a lot more natural. No need to make any physical movements such as moving finger or face. As a result, it is easier to implement continuous authentication. Even if the subject is moving and working, the system works everywhere.

The properties of the ear, on the other hand, are numerous, as indicated in Figure 2.8, and are fixed and unaffected by emotions. Unlike facial recognition systems, glasses, beards, and mustaches cannot be used to hide ear images during the acquisition phase. Hair or curls can, however, partially obstruct vision. It is essential to highlight that the public supports ear modality in access control and government security applications like visas and passports.

Furthermore, because there is no need to touch the sensor, the hygiene issue is avoided; image acquisition can be done discreetly from a distance and does not require user-sensor cooperation. Furthermore, because it is challenging to visually correlate an ear image with a specific individual, ear photos are more secure than facial photographs; most users cannot identify their ear images. As a result, databases with facial photographs should be safer than databases with ear images; in the first situation, the risk of attack is higher.

3.3 Ear Recognition System

We can divide the ear recognition process into four sub-processes, as shown in Figure 3.1. The first step is ear detection which involves finding the ear in images, i.e., localizing the ear box as tight as possible in mixed images that can contain one or several ears. The next step is pre-processing or normalization, an intermediate phase that can enhance and speed up the classification process later by removing unnecessary information from ear images and correcting other conditions, such as illumination and occlusions. The feature extraction phase can be accomplished using either handcrafted or deep features. It is the most crucial phase that can directly affect the system's final performance. Last, a standard classification or identification phase using any distance is sufficient to identify or verify the subject's identity.

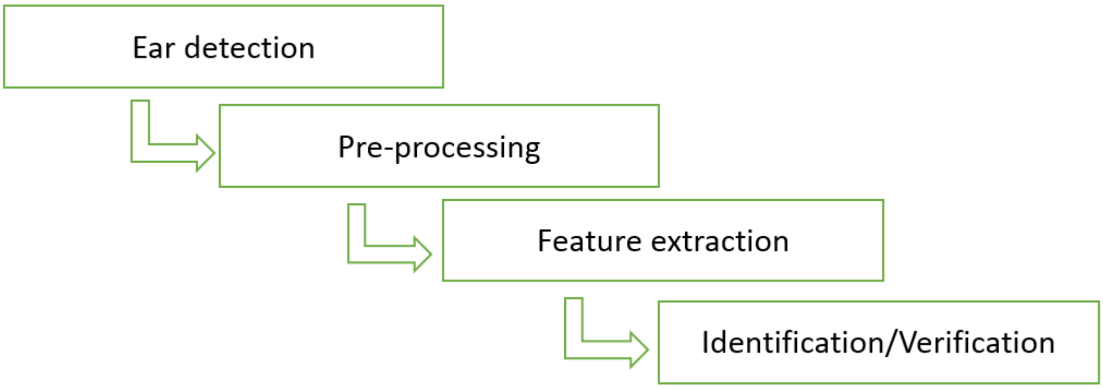


Figure 3.1 Typical ear recognition system.

3.4 Ear recognition methods

Ear recognition techniques can be broadly divided into four categories based on features extraction methodology:

- Local methods that operate using handcrafted local features;
- Geometric methods that focus on the shape of the ear;
- Holistic methods that utilize global information from ear images;
- Deep-learning-based methods that use convolutional neural networks.

Figure 3.2 illustrates the different approaches of ear recognition.

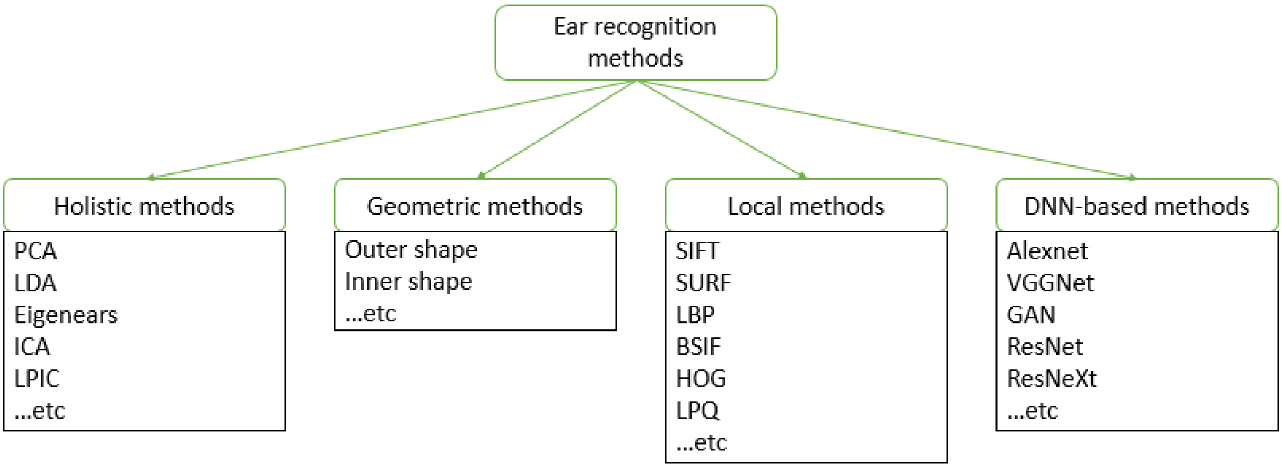


Figure 3.2 Ear recognition methods.

3.4.1 Holistic methods

Holistic techniques presuppose that any collection of ear pictures contains redundancy that may be removed by tensor decomposition. This approach generates basis vectors resembling the original set of images. Each ear picture can be rebuilt in the sub-space using the set of basis vectors. Hurley et al. [40] proposed the force field transform that treats the ear image as a collection of Gaussian attractors acting as the source of a force field. This transformation takes advantage of the directional qualities to determine a small set of potential energies, which is the foundation for ear description.

Gutierrez et al. [41] enhanced ear identification performance using a modular neural network. A 2D wavelet analysis based on global thresholding was used for image compression. The proposed system consists of nine modules; each module has been trained on a subset of the training data to identify a particular part (i.e., Helix, Concha, or Lobule).

Chang et al. [42] applied Principal Component Analysis (PCA) to facial and ear biometrics and found that combining the two modalities resulted in a higher recognition rate. Zhang et al. [43] developed an ear identification system using Independent Component Analysis (ICA) in conjunction with a neural network classifier. They proved that using ICA instead of PCA resulted in a considerable improvement.

Xie and Mu [44] introduced a technique for multi-pose ear identification called Improved Locally Linear Embedding (IDLLE). Hanmandlu and Mamta [45] recently proposed an extension to PCA called local principal independent components (LPIC) that outperformed standard PCA.

In summary, holistic approaches to biometric recognition system building are widespread. They are, nevertheless, subject to background changes and misalignment. As a result, even a minor misalignment of the ears might result in significant categorization errors.

3.4.2 Geometric methods

Geometric techniques attempt to harness the wealth of information included in geometrical ear features such as edge information and ear form when characterizing ear images. Moreno et al. [46] trained multiple neural networks on various geometrical features and then combined the results from each scenario.

Mu et al. [47] offered another geometrical approach in which the feature vector was formed by combining the shape of the outer ear with the inner ear's shape. Choras and Choras [48] presented a work in which they constructed a geometric feature vector using a combination of ear contours and another parametric technique. Rahman et al. [49] developed a geometric feature vector that takes shape, centroid, mean, and Euclidean distance between pixels into account.

Omara et al. [50] recently developed a geometric feature vector built on the minimal ear altitude line that considers the external helix's edge. They used three measure-based criteria in their paper to further improve ear representation.

Geometric approaches appear straightforward to implement and have a low algorithmic complexity level. However, their primary shortcoming is their reliance on the ear’s contours, which can be altered by noise or lightning.

3.4.3 Local methods

Local approaches for ear recognition are based on extracting features from several picture regions, most notably local orientation data. Benzaoui et al. [51] applied and contrasted three different local texture descriptors: Local Binary Pattern (LBP), Local Phase Quantization (LPQ), and binarized statistical image features (BISF). They proved that the BSIF descriptor with the k-NN classifier performed optimally for restricted ear recognition. Benzaoui et al. [49] enhanced their prior work by incorporating embryological and anatomical information about the ear to identify the independent components and their placements at which significant inter-individual variation can be detected. The authors demonstrated that the proposed idea increased the BSIF-k-NN descriptor's efficiency. In another publication, Benzaoui and Boukrouche [52] provided a straightforward and successful method for combining and using color information with local texture descriptors. The authors concluded that the RGB-BSIF descriptor performed more quickly and accurately.

The applicability of a tunable filter bank for feature extraction in ear identification systems was investigated in [53]. The wavelet obtained from a tunable filter based on second-order is applied to each block of the ear image for feature extraction in their suggested approach. The six energy features that resulted from each block of four and six regions produced a final feature vector of 24 and 30 features, respectively. After that, the verification module decides by calculating the distance between Probe ear Feature Vector (FV_p) and Gallery ear Feature Vector (FV_g) within a predetermined threshold using L1, L2, Cosine~~,~~ or Canberra distances.

Guo and Xu [54] presented the Local Similarity Binary Pattern (LSBP) feature extraction approach, which considers similarity and connection features. They combined LSBP and LBP to improve recognition performance. Kumar and Chan [55] efficiently verified the ear's identity by utilizing the sparse representation of surrounding grey-level directions. Al Rahhal et al. [56] recently offered a strategy in which the LPQ descriptor was used for the separated blocs depicting horizontal stripes. The collected characteristics from each stripe were then combined to form a final vector descriptor.

Lakshmanan employed a multi-level fusion of the ear score in [57], taking only the center component into account. He retrieved features from the outer and inner ear in two steps and fused the resulting scores before matching. A recognition rate of 99.2% was attained when the public dataset from the University of Science and Technology Beijing (USTB2) [58] was used. In [59], the authors used a new ear image contrast enhancement approach based on the grey-level mapping technique to solve the acquisition problems of 2D ear images; they used an Artificial Bee Colony (ABC) algorithm as an optimizer. The proposed technique allowed obtaining better-contrasted 2D ear images to be used either in verification or recognition tasks. In another comparative study, Hassaballah et al. [60] conducted a series of experiments on five publically available ear datasets using LBP variants to prove the high discriminative power of LBP-like features.

Local techniques performed well under confined conditions with some ear databases but significantly worse with unconstrained conditions.

3.4.4 DNN-based methods

Deep artificial Neural Networks (DNN), also termed deep learning approaches, have increased interest in computer vision. These methods employ successive convolutional hidden layers of information processing grouped hierarchically for pattern representation, learning, or classification. The literature contains numerous models and architectures for DNNs; the most famous are AlexNet [61], VGGNet [62], Inception [63], and SqueezeNet [64]. These models have demonstrated high performance in various biometric applications, notably facial recognition [65]. Nevertheless, a substantial body of literature on ear biometrics uses deep learning. We can include Tian and Mu's [66] paper among these works; they suggested a three-layer convolutional network for restricted ear recognition. Numerous studies have demonstrated the utility of ensembles of models. Alshazly et al. [67] developed a model ensemble composed of various VGG architectural configurations. Multiple VGG settings are used to extract image features, averaged before being input to a fully connected layer to make a prediction. Experiments on the AMI dataset revealed excellent results, with a 97.50% recognition rate employing a fine-tuned VGG-13-16-19 ensemble. Kacar and Kirci [68] presented the ScoreNet, a novel CNN architecture for unconstrained ear identification, in their landmark paper. They advocated for the establishment of a modality pool in order to boost diversity. They systematically picked the modalities and compared the fusion process to the primary outcome to determine the optimal modality pattern. On the other hand, the recommended training procedure is time-consuming and resource-intensive. Omara et al. [69] retrieved and merged features from different layers of a pre-trained VGG-M utilizing Discriminant Correlation Analysis (DCA). They then used pairwise Support Vector Machine (SVM) to match the resulting feature vector; they regarded the identification problem as a binary classification problem between pair samples. Hansley et al. [70] suggested a two-stage architecture for fusing handcrafted and learned features, beginning with landmark identification using a CNN-based model, followed by feature extraction, both learned and handcrafted, and lastly, score normalization and fusion. Priyadharshini et al. [71] proposed a six-layer deep convolutional neural network architecture for ear identification. Khaldi and Benzaoui [72] tackled the problem of ear grayscale image recognition by employing a Generative Adversarial Network (GAN) to generate color images to enhance the identification process. In another study [73], the authors used generative adversarial networks to synthesize a region-of-interest of the ear free of all non-ear pixels, such as occlusions and hair. The proposed technique also allowed generating missing parts of the ear. In another recent study by Khaldi et al. [39], the authors presented deep unsupervised active learning to enhance the model’s performance during the test phase. Using such a technique, a classification model can acquire new knowledge incremental and unsupervised.

Numerous recent studies have begun to investigate the idea of data augmentation to circumvent the issue of limited training data. It is a well-known difficulty when it comes to training CNNs. Emersic et al. [75] conducted a study in which they attempted to address the latter issue by supplementing the training dataset. They used rotation, scaling, and transformation to generate similar images for each training image. Dodge et al. [75] evaluated and compared the performance of all available deep neural network models, demonstrating that data augmentation can dramatically increase recognition performance. Additionally, they proposed a framework for deep learning to mitigate the effect of over-fitting; this framework obtained superior performance. Transfer learning is another strategy used to address the later issue. This strategy works by transferring knowledge from one domain to another related domain using a new training dataset to fine-tune the deep model and its hyper-parameters [88]. Alshazly et al. [76] recently researched transfer learning utilizing pre-trained deep models, specifically AlexNet, VGGNet, Inception, ResNet [77], and ResNeXt [78]. Additionally, they used ResNeXt101 model ensembles to obtain their best findings. Zhang et al. [79] explored few-shot learning-based methods towards model training with limited training images.

The Unconstrained Ear Recognition Challenge (UERC) [80] kicked off a series of challenges focused on ear recognition utilizing massive unconstrained datasets. The organizers' objective is to evaluate competitors' technique performance and their capacity to adapt to uncontrolled conditions such as generalization to unseen data, sensitivity to rotations, occlusions, and low resolution. Five teams of researchers competed by evaluating and analyzing six ear identification systems using large-scale ear datasets. The competition's next edition was held in 2019 [82]; researchers from 12 institutions submitted a total of 13 distinct techniques, utilizing local descriptor-based methods and CNN-based models, as well as hybrid models. These events made significant contributions to the advancement of biometric recognition, and each edition of the challenge has seen new advances.

Table 3.1 summarizes the studies mentioned in this chapter, categorizing them according to their categories, the datasets used, and the experimental protocols.

Table 3.1 Comprehensive summary of state-of-the-art methods.

Approach	Papers	Feature extraction method	Employed datasets			Evaluation Protocol
			Name	#Sub.	#Img.	
Holistic	Hurley et al. [40] (2000)	Force Field Transform	Private	NA	NA	NA (Not Available)
	Chang et al. [42] (2003)	PCA	Private	NA	NA	NA
	Zhang and Mu [83] (2008)	Geometric Features + ICA	USTB-1	60	180	2 img/sub Training & 1 remaining Test
			USTB-2	77	308	3 img/sub Training & 1 remaining Test
			Private	17	102	5 img/sub Training & 1 remaining Test
	Gutierrez et al. [41] (2010)	Wavelet Transform	USTB-2	77	308	3 img/sub Training & remaining 1 img/sub Test
	Tariq et al. [84] (2011)	Haar Wavelets + Fast Normalized Cross Correlation	USTB-1	60	180	120 Train & 65 Test
			USTB-2	77	308	3 img/sub Training & 1 remaining Test
			IITD-1	125	493	250 Train & 243 Test (Randomly)
Geometric	Moreno et al. [46] (1999)	Geometric Features (Morphological Description)	Private	28	186	3 Train, 1 Validation, & 2 Test
	Burge and Burger [85] (2000)	Adjacency Graphs of Voronoi Diagrams	Private	NA	NA	NA
	Mu et al. [47] (2004)	Geometrical Measures on Edge Images	USTB-2	77	308	3 img/sub Training & 1 remaining Test
	Choras and Choras [48] (2010)	Geometrical Approaches on Longest Ear Contours	Private	NA	NA	NA
	Rahman et al. [49] (2014)	Geometric Features	Private	NA	NA	NA

	Lakshmanan [86] (2013)	Multi-Level Fusion	USTB-2	77	308	3 img/sub Training & 1 remaining Test
	Omara et al. [50] (2016)	Geometric measurements	USTB-1	60	180	60 Train & 120 Test (Randomly, 10 times)
			IITD-1	125	375	125 Train & 250 Test (Randomly, 10 times)
Local	Guo and Xu [54] (2008)	LBP + Cellular NN	USTB-2	77	308	3 img/sub Training & 1 remaining Test
	Benzaoui et al. [51] (2014)	BSIF	IITD-1	125	493	#exp1: 250 Train & 243 Test #exp2: 125 Train & 368 Test
			IITD-2	221	793	#exp1: 442 Train & 351 Test #exp2: 221 Train & 572 Test
			USTB-1	60	185	#exp1: 120 Train & 65 Test #exp2: 60 Train & 125 Test
	Ghoualmi et al. [59] (2016)	SIFT	IITD-1	125	421	3 img/sub Training & remaining Test
			USTB-1	180	60	
			USTB-2	77	308	
	Chowdhury et al. [53] (2018)	Tunable Filter Bank	AMI	100	700	60% Train & 40% Test (Randomly)
			IITD-1	125	493	2 img/sub Train & remaining Test (Randomly)
			UERC-17	3540	11804	2304 Train & 9500 Test
	Al Rahhal et al. [56] (2018)	LPQ	IITD-1	125	465	1 img/sub Train & remaining Test (3 permutations)
			IITD-2	221	793	
	Hassaballah et al. [60] (2019)	Completed LBP	IITD-1	125	493	2 img/sub Train & remaining Test (Randomly)
			IITD-2	221	793	2 img/sub Train & remaining Test (Randomly)
			AMI	100	700	60% Train & 40% Test (Randomly)

			WPUT	474	3348	5-fold cross-validation
			AWE	100	1000	60% Train & 40% Test (Randomly)
			IITD-2	221	793	
Deep Neural Networks	Emeršič et al. [74] (2017)	SqueezeNet	AWE + CVLE + 500 images	166	2304	60% Train & 40% Test (Randomly)
	Hansley et al. [70] (2018)	CNN + HOG	UERC-17	3540	11804	2304 Train & 9500 Test
	Omara et al. [69] (2018)	Pairwise SVM	USTB-1	60	180	#exp1: 1 img/sub Train & remaining Test (3 permutations) #exp2: 2 img/sub Train & remaining Test (Randomly) #exp3: 3 img/sub Training & remaining Test
			USTB-2	77	308	
			IITD-1	125	493	
			IITD-2	221	793	
	Zhang et al. [43] (2018)	VGG-face	AWE	100	1000	60% Train & 40% Test
	Alshazly et al. [77] (2019)	VGG	AMI	100	700	60% Train & 40% Test (Randomly)
			WPUT	474	3348	
	Zhang et al. [80] (2019)	MAML + CNN	AMI	100	700	60% Train & 40% Test (Randomly)
			UERC-17	3540	11804	2304 Train & 9500 Test
	Alshazly et al. [67] (2019)	AlexNet (Fine Tuning)	AMI	100	700	60% Train & 40% Test (Randomly)
			CVLE	16	804	
	Alshazly et al. [77] (2020)	ResNeXt101	EarVN1.0	164	28412	60% Train & 40% Test (Randomly)
	Priyadharshini et al. [71] (2020)	CNN	IITD-2	221	793	490 img Train & 303 Test (Randomly – 10 times)
			AMI	100	700	600 img Train & 100 Test
	Khaldi et al. [72] (2020)	GAN + CNN	AMI	100	700	60% Train & 40% Test
		GAN + CNN	AWE	100	1000	60% Train & 40% Test
	Khaldi et al. [73] (2020)	GAN + CNN	AWE	100	1000	60% Train & 40% Test
	Khaldi et al. [27] (2021)	Active Learning	AMI	100	700	60% Train & 40% Test
			AWE	100	1000	60% Train & 40% Test
			C-USTB2	77	308	60% Train & 40% Test

While advances in ear recognition have been promising, there are many unresolved issues and open challenges such as the variations in pose, occlusion, resolution, and ear synthesis. We continued our research to address some of these challenges by offering different schemes and methods to improve the overall ear recognition process.

3.5 Conclusion

In this chapter, we presented the ear anatomy, ear recognition system, and its components. We also reviewed the most important state-of-the-art ear recognition approaches based on the method used, either: holistic, geometric, local, or DNN-based. We reviewed how each approach can have its pros and cons, but in global point-of-view, deep learning-based methods outperformed other methods, especially in unconstrained scenarios. With the emergence of several unconstrained ear databases (also known as ear in the wild), traditional machine learning algorithms suffered from poor performance. Many of these challenges have been overcome by using deep learning-based approaches.

For this reason, we decided to dive into research using deep learning. Detailed discussion and comparison of our proposed approaches against state-of-the-art methods are provided in the following chapters.

CHAPTER 4: Review of Deep Learning

4.1 Introduction

In Artificial Intelligence (AI), or more specifically Machine Learning (ML), a training model known as "deep learning" is more closely to how the human brain makes decisions. We use the term "brain" to suggest that the algorithms are more complex. Instead of relying on a single layer for processing, numerous complex layers are employed. The layers of a neural network can communicate with one another. Automated learning is a step closer to unsupervised learning.

Before big data and cloud computing, the amount of data and processing power required were not widely accessible. Even if a large volume is required, this does not indicate that the data must be structured. Both organized and unlabeled data can be processed by deep learning. It also generates more complicated statistical models due to this approach of learning. Data adds complexity to the model, but accuracy improves with it as well.

The rest of the chapter is organized as follows: Sections 4.2 and 4.3 represent a background of deep learning and artificial neural networks. Section 4.4 discusses the approaches of deep learning. Section 4.5 outlines convolutional neural networks architectures. Section 4.6 presents some applications of deep learning. Next, we review the standing challenges in deep learning. Last, in section 4.8, we conclude this chapter.

4.2 Machine learning vs. deep learning

ML is a suite of techniques and tools that enables machines to recognize patterns within data and reason about a specific task using this underlying structure [87]. Machines attempt to comprehend these fundamental patterns in a variety of ways. However, what is the relationship between ML and Deep Learning (DL)?

There is a widespread misperception that deep learning is a competitor to machine learning. Figure 4.1 illustrates the relationship between DL and ML; to put things in context, DL is a subdomain of machine learning. With enhanced computer power and massive data sets at their disposal, deep learning algorithms may self-learn hidden patterns within data to make predictions. In summary, consider DL to be a subset of machine learning trained on a vast quantity of data and utilizes many processing units to make predictions [88].

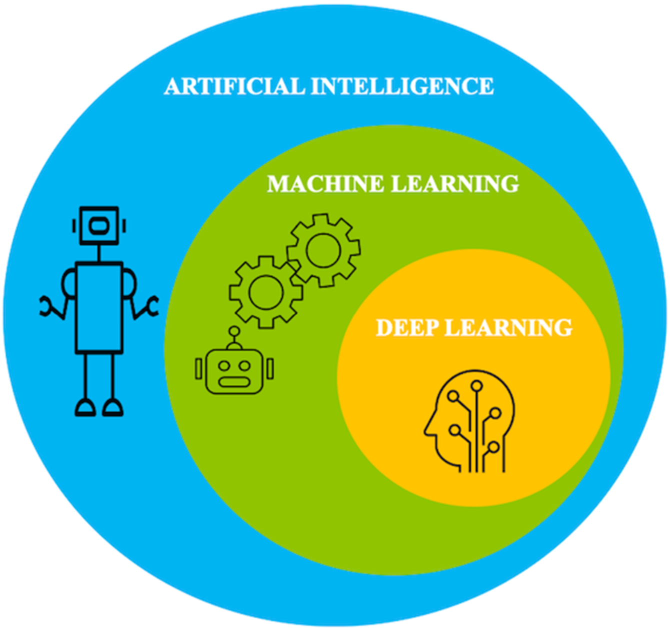


Figure 4.1 The relationship between AI, ML, and DL.

4.3 Artificial Neural Networks

4.3.1 Towards Neural Networks

The perceptron is the essential part of an Artificial Neural Network (ANN); a Perceptron is a learning approach for supervised binary classifiers. Binary classifiers use a series of vectors to determine if an input belongs to a particular class. A single-layer neural network is what a perceptron is, in a nutshell. Weights and bias are included for each of the input value, in addition to a net sum and activation function [89]. The first step is to multiply all the input values by their respective weights. Then, the weighted sum is calculated by multiplying each of the multiple values. Weighted sums are added together and used to a perceptron's activation function to produce its output. Using the activation function, we can ensure that the output is mapped to values like 0 and 1 (or -1 and 1). The strength of a node can be gauged by looking at the weight of an input. Activation function curves can be shifted up or down by varying the bias value of an input in a similar way [90]. The perceptron's workflow is depicted in Figure 4.2.

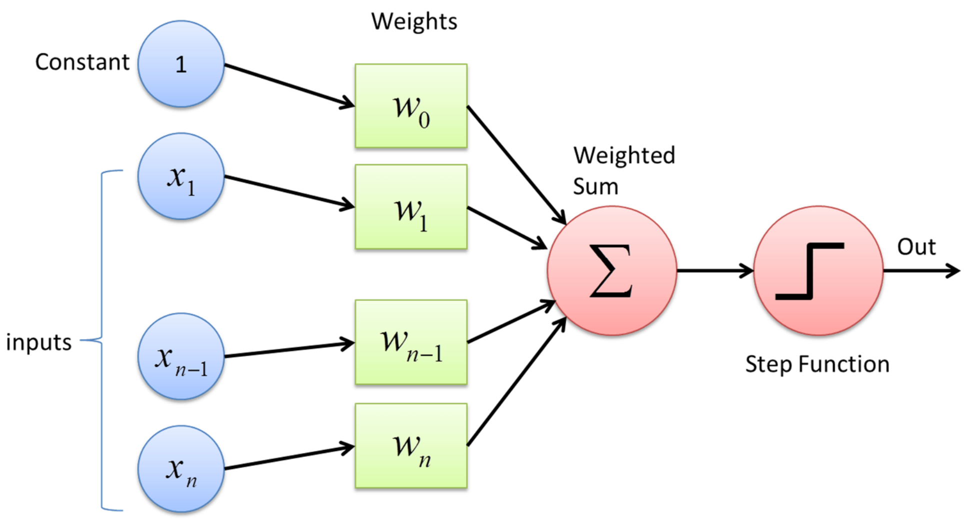


Figure 4.2 Workflow of the perceptron.

ANNs are a logical progression from perceptrons. An example of a feed-forward neural network is a multi-layered perceptron. Input, output, and perceptron neurons (as well as synaptic weights) would all be part of it. Figure 4.3 depicts the ANN's design.

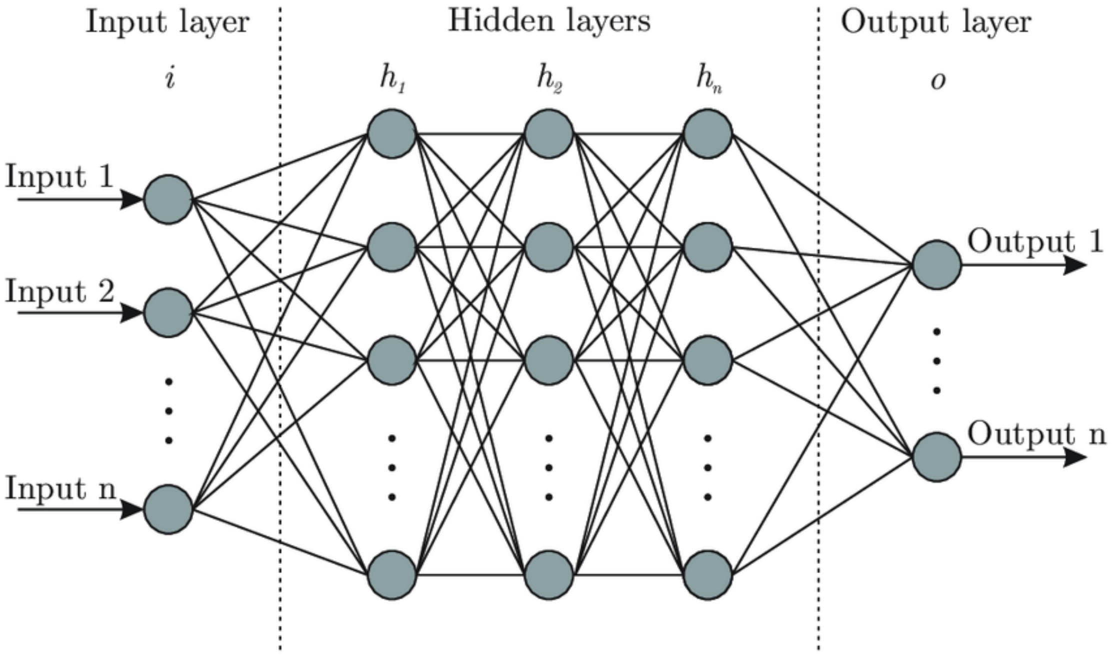


Figure 4.3 The architecture of ANN.

4.3.2 Activation functions

The activation function is critical for an artificial neural network to learn. It converts an input signal to an output signal like any other function. This output signal is the input to the next layer [91]. The activation function determines whether or not to stimulate a neuron by computing the weighted sum and then adding bias. The goal is to introduce non-linearity into a neuron's output. The output signal is just a linear function if no activation function is used. While linear functions are simple to implement, they are insufficient in some cases. Non-linear functions have more than one degree and exhibit curvature. Now, we require a neural network capable of learning and representing practically any complex function [92]. The most used activation functions are:

1) **Threshold Activation Function:** It is called a binary step function based on a threshold. If the input value is greater than a predefined threshold, the neuron is triggered and sends one as a value to the following layer, as shown in Figure 4.4.

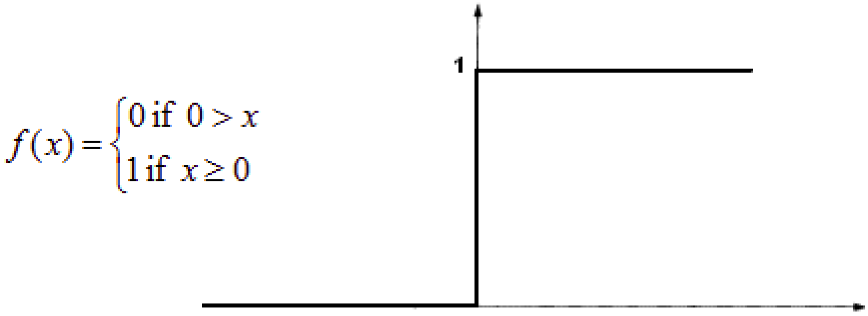


Figure 4.4 A Binary step function.

The limitation of this method is that it can only create binary classifications (1 or 0). However, if we wish to connect numerous neurons to add other classes, in this situation, if all neurons will output 1, the class cannot be determined [93].

2) **Sigmoid Activation Function:** A sigmoid function is a mathematical function with a characteristic S-shaped curve or sigmoid curve that spans the range of 0 to 1, as seen in Figure 4.5. It is utilized in models where the outcome is required to estimate a probability [93].

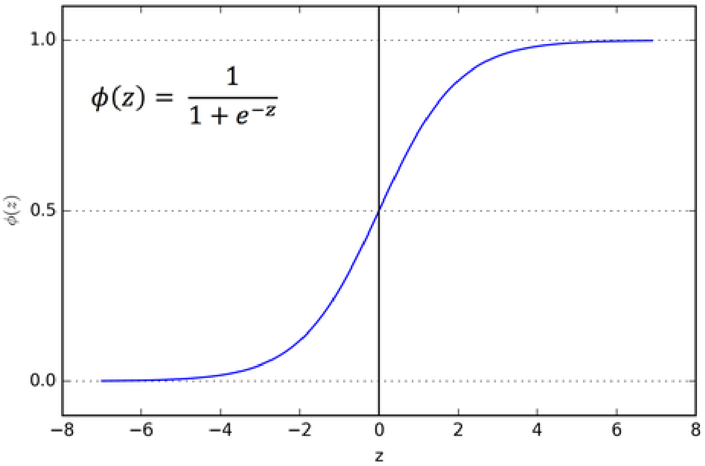


Figure 4.5 Sigmoid curve

The Sigmoid function is differentiable, which implies that it may be used to determine the slope of a curve between any two points. The disadvantage of this activation function is that it might lead the network to become stuck during training if significant negative inputs are provided constantly.

3) **Hyperbolic Tangent Function (tanh):** It is similar to the sigmoid but performs better (Figure 4.6). Nature is nonlinear, which is why we can stack layers. The range of the function is between [-1,1].

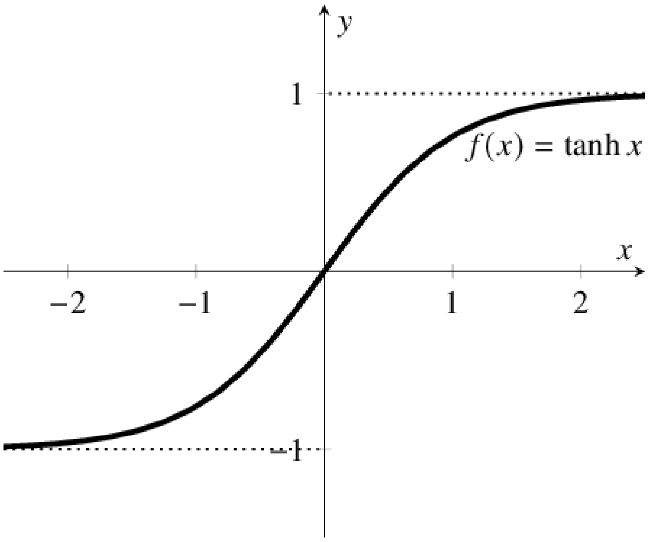


Figure 4.6 The tanh function.

The primary advantage of this function is that it maps substantial negative inputs to negative outputs while mapping zero-valued inputs to near-zero outputs. As a result, the likelihood of becoming stuck during training is reduced.

4) **Rectified Linear Units (ReLu):** ReLu is the most often employed activation function in Convolutional Neural Networks (CNN) and Artificial Neural Networks (ANN), with a range of zero to infinity. It returns a value of *x* if it is positive and 0 otherwise. It appears to have the same linear function difficulty as it does in the positive axis. Relu is a non-linear function by definition, and its combination with other functions is also non-linear. Indeed, it is an excellent approximator, and any function may be approximated with Relu. It is six times more efficient than the hyperbolic tangent function. It should be used solely on the neural network's hidden layers. Thus, for the output layer, we should use the softmax function for classification problems or a linear function for regression problems [94].

One issue is that some gradients are fragile and can die during training. It generates a weight update, preventing it from activating any subsequent data point [95].

Essentially, ReLu may result in the death of neurons. Leaky ReLu was introduced to address the issue of dying neurons. As a result, Leaky ReLu presents a modest slope to maintain the updates. Leaky ReLu has a value between - and +. Figure 4.7 shows the difference between ReLu and Leaky Relu.

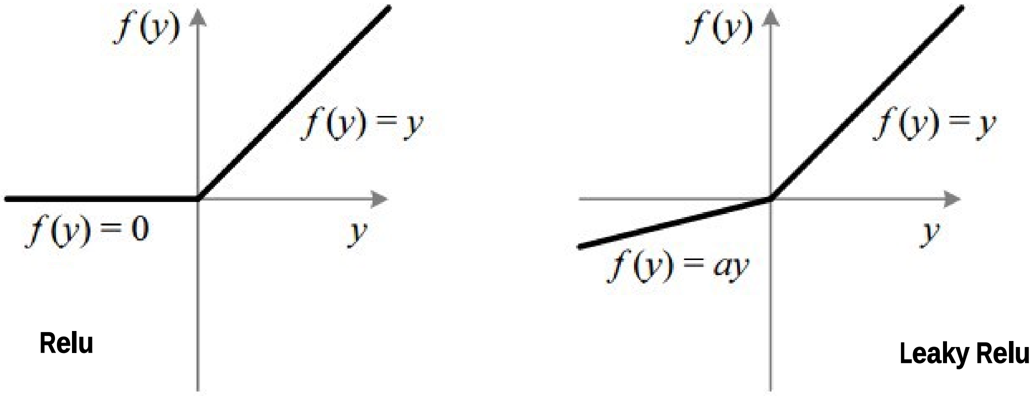


Figure 4.7 ReLu vs. Leaky ReLu.

5) **Softmax activation function:** It is a mathematical function that turns a vector of integers into a vector of probabilities, with the probability of each value proportional to the vector's relative scale as defined in Eq. 4.1.

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{4.1}$$

4.3.3 How do Neural networks to learn?

An analogy may help to understand brain network mechanisms. There are many similarities between neural network learning and human learning, such as learning in our daily lives and activities. For example, when we do an action and receive feedback from a trainer, we improve our performance. A trainer is also needed to explain what the output should have been when it comes to neural networks. Based on this discrepancy between actual and anticipated values, a cost function error value is calculated and sent back to the system.

Cost functions are assessed and utilized to change thresholds and weights for the following input in the network at each layer. As a team, we are working to reduce the cost function. The closer the projected value to the actual value is the lower the cost function. In this approach, the network improves at analyzing values, and the error decreases over time. The results are fed back into the neural network and reprocessed. We can control the weighted synapses that connect input variables to the neuron.

Adjusting the weights is necessary if there is a discrepancy between the actual and projected values. A new cost function will be generated, ideally smaller than the previous one if we adjust them a little and rerun the neural network. To get the cost function down to the smallest possible size, we must go through this process repeatedly. The outlined technique, known as Back-propagation, is continuously applied across a network to keep the error value as low as possible.

4.3.5 Back-propagation

Neural network training relies on back-propagation to do its task. It is a technique for optimizing neural network weights based on the previous epoch's error rate (i.e., iteration). We can obtain lower error rates and increase the model generalization by adjusting the weights [96].

In neural networks, "backward propagation of mistakes" is known as "back-propagation." It is a common practice in artificial neural network training. A loss function's gradient can be calculated using this method for all network weights. This algorithm uses the chain rule to compute a single weight's loss function gradient using back-propagation.

A feed-forward neural network's weight-space gradient concerning a loss function is computed using back-propagation. The chain rule is crucial in back-propagation. Here is a partial differentiation of loss (*L*) in terms of weights/parameters (*w*). An increase or decrease in the weight of an object affects its value. *z(∂z/∂w)* is an essential factor in activation *a*(∂a/∂z), which is affected by even a slight change in the value. The loss function *L(∂L/∂a)* is affected by a tiny modification in the activation a. Equation 4.2 [97] shows how to express it mathematically.

$$\frac{\partial \mathrm{L}}{\partial \mathrm{w}} = \frac{\partial \mathrm{L}}{\partial \mathrm{a}} \times \frac{\partial \mathrm{a}}{\partial \mathrm{z}} \times \frac{\partial \mathrm{z}}{\partial \mathrm{w}} \tag{4.2}$$

where *L* is the loss function, *w* is the weights, *z* is the linear regression, and *a* is the activation function used.

4.4 Convolutional Neural Networks

DL algorithms such as Convolutional Neural Networks (CNNs or ConvNets) can distinguish between distinct aspects and objects in an image and then use that knowledge to create new images based on that information [98]. There is far less pre-processing necessary when using a CNN than other classification methods. CNN can learn certain filters/characteristics if they are given enough training.

A CNN's architecture is similar to the connectivity pattern of neurons in the human brain and was inspired by the visual cortex's arrangement. The receptive field is the area of the visual field in which individual neurons respond to stimuli. A collection of these fields covers the entire visual field. A CNN can capture the spatial and temporal dependencies in an image through relevant filtering techniques. The reduced number of parameters and reusability of weights allow the architecture to fit the picture collection better. In other words, the network may be trained to comprehend the image's complexity better. Figure 4.8 displays a generic CNN architecture [99].

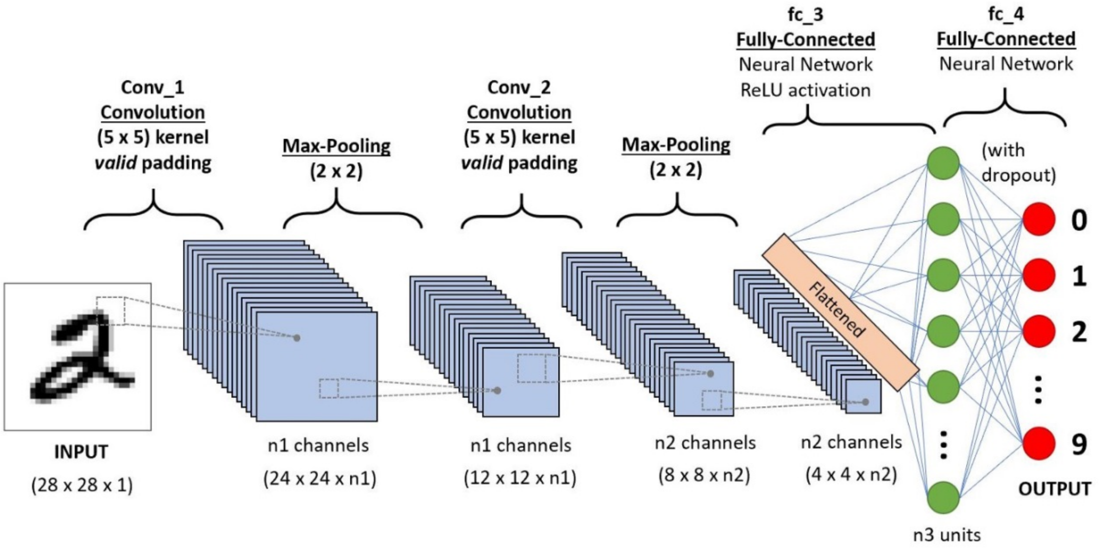


Figure 4.8 A CNN example for handwritten digits classification [100].

4.4.1 Convolutional Layers

The convolutional layer is the fundamental block, as it is responsible for most of the computations. It comprises three components: input data, a filter, and a feature map. The kernel or filter traverses the image's receptive fields, checking for the presence of features; this is referred to as convolution [101]. The kernel is a two-dimensional (2-D) weighted array representing a portion of the image. While filter sizes vary, they are commonly a 3×3 matrix; this also dictates the size of the receptive field. The dot product of the filter and the portion of the input array is calculated and loaded into an output array. Following that, the filter shifts by one stride, and the procedure is repeated until the kernel has swept across the entire image. A feature map, activation map, or convolved feature is the ultimate result of a series of dot products from the input and the filter [102].

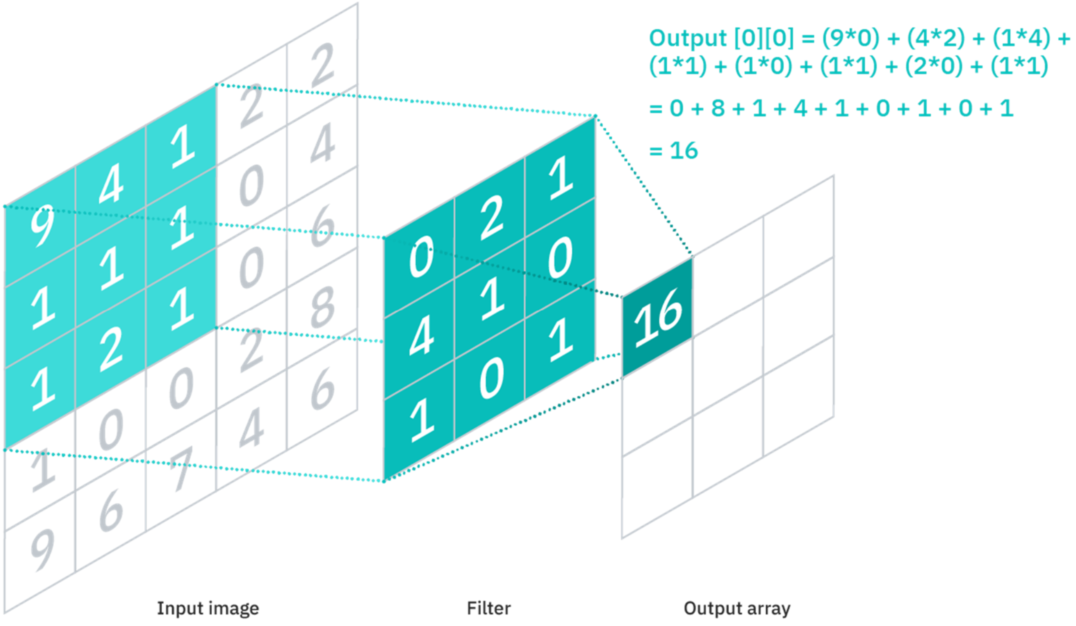


Figure 4.9 The linear convolution operation.

As illustrated in Figure 4.9, the convolution procedure does not require that each output pixel in the feature map be connected to each pixel value in the input array. For that, convolutional layers are frequently referred to as partially-connected layers. The filter weights remain constant as it traverses images, a phenomenon is known as parameter sharing. Specific parameters, such as the weight values, are adjusted during training using back-propagation and gradient descent. However, three hyper-parameters must be set before the neural network training begins. Among them are [101]:

1) The number of filters: It affects the output's depth. *n* distinct filters, for example, would result from *n* different feature maps.
2) The stride parameter: specifies the distance or the number of pixels that the filter moves across the input array.
3) Zero-padding is typically used when the filters do not fit the input image; this reduces the size of all items outside the input matrix to zero, resulting in a larger or equal-sized output. It is classified into three types:
   - Valid padding: Also referred to as no padding.
   - Consistent padding: Ensures that the output layer is identical to the input layer in size.
   - Full padding: Enlarges the output by appending zeros to the input's border.

A CNN adds a Rectified Linear Unit (ReLU) adjustment to the feature map following each convolution operation, bringing nonlinearity into the model.

As previously stated, other convolution layers may be added after the initial convolution layer. When this occurs, the CNN's structure can become hierarchical. For example, suppose we attempt to detect whether an image contains a bicycle. Consider the bicycle as a collection of components. It comprises a frame, handlebars, wheels, and pedals, among other components. Each bicycle component represents a lower-level pattern in the network, while its combination represents a higher-level pattern, resulting in a feature hierarchy within the CNN, as illustrated in Figure 4.10 [102].

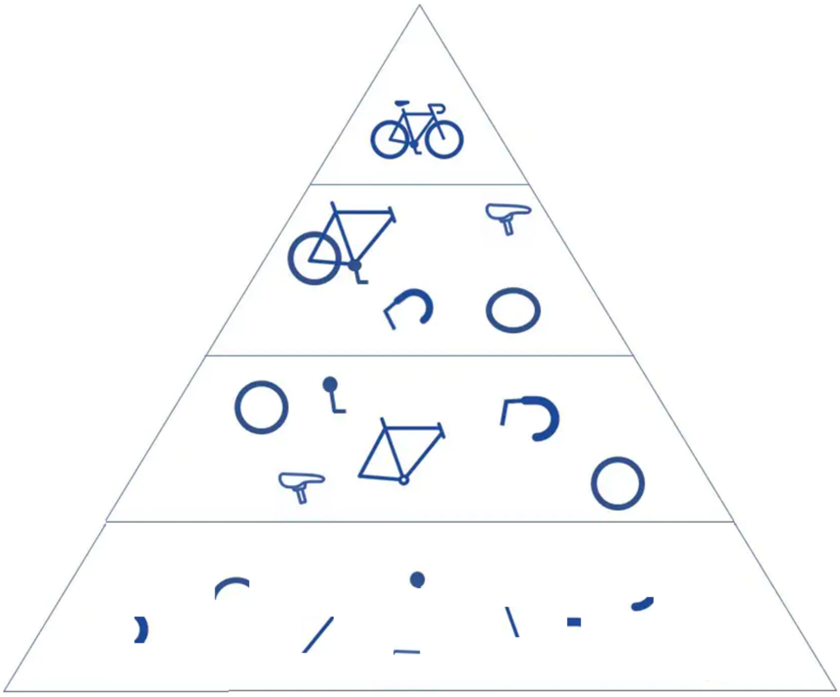


Figure 4.10 Example of feature hierarchy created by CNN.

4.4.1 Pooling Layer

Pooling layers, also called downsampling, is a technique for lowering input dimensionality by reducing the number of factors. The pooling operation sweeps an unweighted filter across the entire input using an aggregation function to populate the output array with the values contained within the receptive field. It can be classified into two broad categories [103]:

- Pooling to a maximum: The filter traverses the input array. It selects the pixel with the highest value for the output array.
- Average pooling: The filter determines the average value contained inside the receptive field.

While the pooling layer loses much information, it also provides several benefits for the CNN. They contribute to the reduction of complexity, the enhancement of efficiency, and the avoidance of over-fitting [104].

4.4.1 Fully-Connected Layer

The Full-Connected (FC) layer is what its name implies. Layers that are only partially connected have no direct connection between the input image and the output layer. Nevertheless, in a completely interconnected layer, every node in the output layer is directly linked to a node in the previous layer [105].

Classification is done here using the information from the previous layers and their various filters. FC layers often employ a softmax activation function to categorize inputs adequately, whereas convolutional and pooling layers typically use ReLu functions, resulting in a probability ranging from zero to one.

4.5 CNN architectures

4.5.1 AlexNet

The origins of deep CNNs date back to the birth of LeNet (Figure 4.11) [97]. At the time, CNN's were limited to jobs involving the recognition of handwritten digits, which could not be scaled to all image classes. AlexNet is highly regarded in deep CNN architecture [64] since it achieved groundbreaking results in image classification. By expanding the depth of the CNN and adopting different parameter optimization procedures, Krizhevesky et al. [61] proposed the AlexNet and subsequently increased the CNN's learning capabilities. The AlexNet architecture is depicted in Figure 4.12.

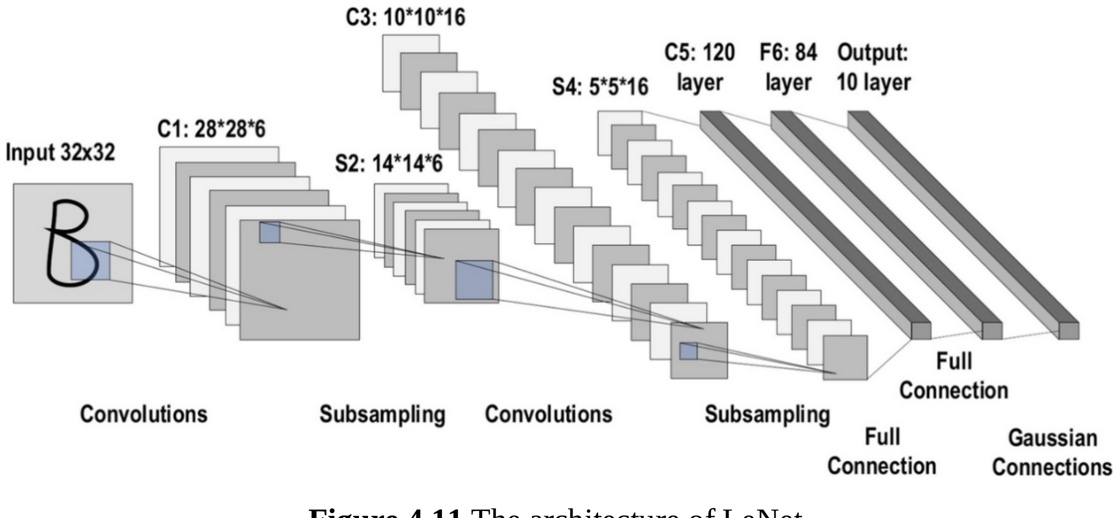


Figure 4.11 The architecture of LeNet.

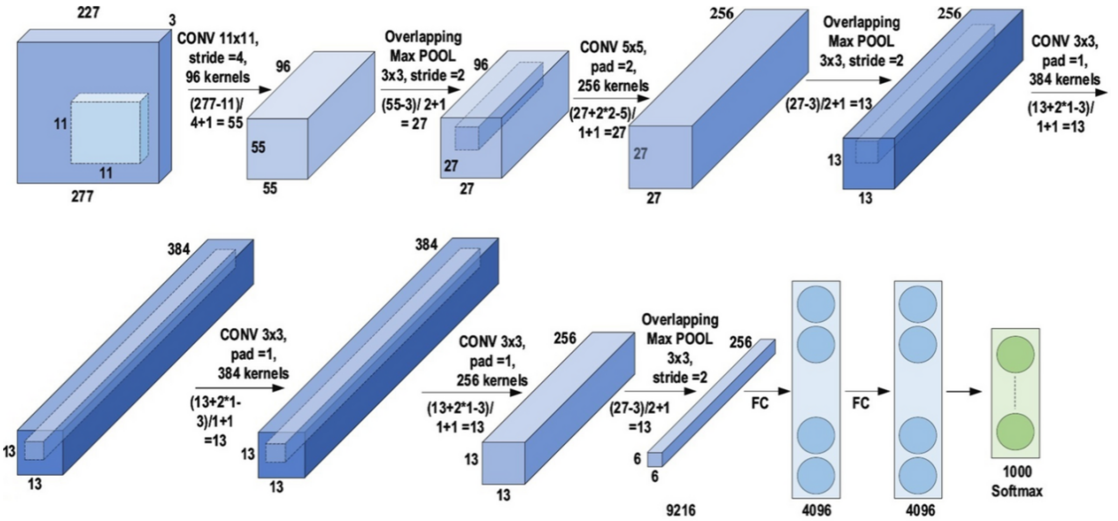


Figure 4.12 The architecture of AlexNet.

Due to hardware limitations, deep CNN's ability to learn was limited. AlexNet was trained on two Graphics Processing Units (GPU) (NVIDIA GTX 580) at the same time to address these hardware restrictions. Additional feature extraction steps have been added in AlexNet so that the CNN can be used for a broader range of image classifications. Even though depth promotes generalization for various image resolutions, overfitting was the main negative associated with depth. Krizhevesky et al. [106] solved this problem. The approach suggested by Krizhevesky et al. randomly passes over numerous transformational units during the training phase to ensure that the features learned by the system are extra resilient. The vanishing gradient problem can be reduced by using ReLU as a non-saturating activation function to speed up convergence [107].

Additionally, overlapping subsampling and normalization of local responses were used to reduce overfitting and improve generalization. The use of large-size filters (5×5 and 11×11) in the first layers of the network was another way to increase the network's performance. In recent CNN versions, AlexNet has had a significant impact, and launched a new era of research in CNN applications.

4.5.2 Network-in-network

The Network-In-Network (NIN) model differs slightly from previous models and introduces two novel notions [108]. The first was accomplished by the use of many layers of perceptual convolution. These convolutions are carried out with the help of one filter, which enables the addition of further nonlinearity to the networks. Additionally, this allows the expansion of the network depth, which can then be regularized via dropout. This concept is widely used in the bottleneck layer of DL models. Global Average Pooling (GAP) layers are also used instead of FC layers, which are the second innovative notion and significantly reduces model parameters.

Additionally, GAP modifies the network architecture significantly. When GAP is used on a prominent feature map, it is possible to generate a final low-dimensional feature vector without reducing the dimension of the feature map [109]. The network's architecture is depicted in Figure 4.13.

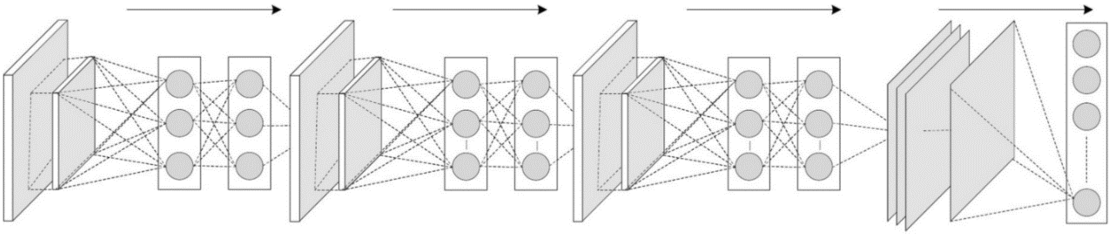


Figure 4.13 The architecture of network-in-network.

4.5.3 ZefNet

Before 2013, the CNN learning process was primarily created through trial and error, making it impossible to determine the precise goal of the upgrade. This issue limited the performance of deep CNN on complex pictures. Zeiler and Fergus responded in 2013 [110] by introducing the multilayer De-Convolutional Neural Network (DeconvNet). This technology became known as ZefNet and was created to examine the network quantitatively. The goal of the network activity visualization was to assess the CNN's performance by analyzing neuron activation. However, Erhan et al. used this exact concept to optimize the performance of Deep Belief Networks (DBN) by displaying the hidden layer characteristics [111]. Additionally, Le et al. evaluated the performance of the deep unsupervised Auto-Encoder (AE) by evaluating the image classes formed using the output neurons [112]. DenconvNet performs similarly to a forward-pass CNN by reversing the operation order of the convolutional and pooling layers. This type of reverse mapping reverses the output of the convolutional layer, resulting in visually discernible image forms that correspond to the interpretation of the internal feature representation learned. ZefNet's fundamental premise was to evaluate the learning scheme throughout the training stage.

Additionally, it used the results to identify a capability issue associated with the model. This concept was proved experimentally on AlexNet using DeconvNet; this revealed that just a subset of neurons was active, while the remainder remained dormant in the network's first two layers. Additionally, it suggested that the second layer's extracted features had aliasing objects. As a result of these outcomes, Zeiler and Fergus altered the CNN topology. Additionally, they optimized the parameters and leveraged CNN learning by reducing the stride and filter sizes. As a result of this reconfiguration of the CNN topology, performance was improved. This reorganization proposes that visualizing the features may be used to uncover design flaws and make relevant parameter changes. The network's architecture is depicted in Figure 4.14.

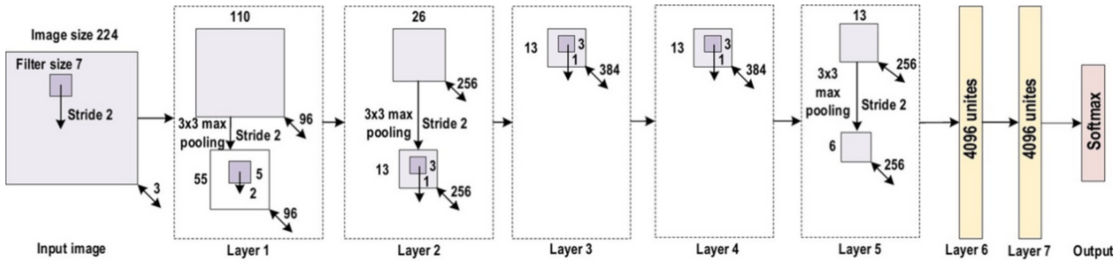


Figure 4.14 The architecture of ZefNet.

4.5.4 Visual Geometry Group (VGG)

Simonyan and Zisserman suggested a straightforward and efficient design principle for CNN after it was proven to be helpful in the field of image recognition. Visual geometry group was the name of this groundbreaking design. Nineteen additional layers than ZefNet [110] and AlexNet [61] were used to simulate the depth of the network representational capacity in-depth in this multilayer model [62]. ZefNet, on the other hand, was the frontier network in the large scale visual recognition challenge 2013 (2013-ILSVRC) competition, proposing that smaller filters could improve the performance of CNN. VGG inserted a layer of 3×3 filters rather than 5×5 and 11×11 filters in ZefNet. These small-size filters were shown experimentally to have the same effect as larger-sized filters when used in parallel. On the other hand, small-size filters produced a receptive field comparable to that of the larger-size filters (77 and 55); using small-size filters allowed for an additional benefit of reducing computing complexity by reducing the number of parameters. These results set the stage for a new paradigm in CNN filter research: working with small-sized filters. In addition, VGG manages network complexity by inserting 1×1 convolutions in the center of the convolutional layers of the network. A linear grouping of the following feature maps is learned by it. Max pooling [113] is placed after the convolutional layer, while the padding is used to keep the spatial resolution in place. For localization and image classification, VGG achieved remarkable results. With its increased depth, homogeneous topology, and simple design, it gained a reputation in the 2014 The ImageNet Large Scale Visual Recognition Challenge (2014-ILSVRC) competition. However, the overuse of 140 million parameters in VGG resulted in an excessively high computational cost. The network's structure is depicted in Figure 4.15.

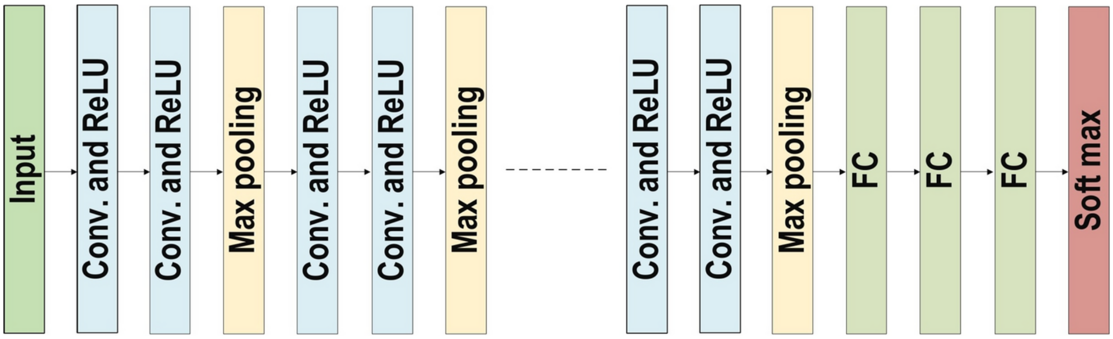


Figure 4.15 The architecture of VGG.

4.5.5 GoogLeNet

The winner of the 2014-ILSVRC competition was GoogleNet (also known as Inception-V1) [63]. The GoogleNet architecture's primary goal is to achieve high-level accuracy while reducing processing costs. In this study, an inception block (module) approach that combines multiple-scale convolutional transformations by applying merge, transform, and split functions for feature extraction was developed for the CNN context. The inception block architecture is shown in Figure 4.16. Channel and spatial information can be captured at various spatial resolutions by using filters of various sizes (55, 33, and 11). Rather than using the standard convolutional layer of GoogLeNet, tiny blocks of Micro-Neural Networks (MNN) were used to replace each layer. The GoogLeNet ideas of merging, transforming, and splitting were used in conjunction with a concern associated with various learning types of variants present in a comparable class of multiple photos. Aims of GoogLeNet were enhancing CNN parameters and increasing learning capabilities. As a bottleneck layer, an 11 convolutional filter is inserted ahead of employing large-size kernels to control the computation. In order to solve the problem of redundant information, GoogleNet used sparse connections. It saved money by ignoring the unnecessary avenues of distribution. As a reminder, only some of the input channels are connected to the output channels. The density of connections was reduced by using a GAP layer instead of an FC layer as the terminal layer. These parameter tunings also reduced the number of parameters from 40 millions to 5 millions. Additionally, RmsProp was employed as an optimizer, and batch normalization was performed [114]. To speed up the process of convergence, GoogleNet recommended using additional learners. GoogleNet, on the other hand, suffers from a topology that is difficult to modify from one module to the next. Representation jam, which reduced feature space in the subsequent layer and caused crucial information loss, is another drawback of GoogleNet.

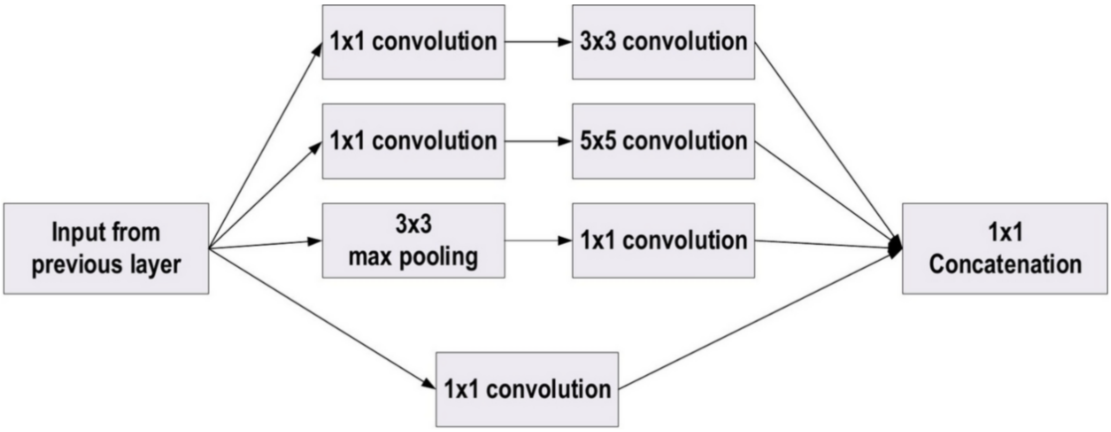


Figure 4.16 The basic structure of GoogLeNet Block.

4.5.6 ResNet

Residual Network (ResNet) was invented by Kaiming He and his colleagues [90]. For this project, the goal was to create an ultra-deep network free of the vanishing gradient problem. ResNet was divided into several distinct categories depending on the number of layers. ResNet50, which has 49 convolutional layers and a single FC layer, was the most frequent form. Network variables totaled 25.5 millions. Figure 4.17 depicts the basic ResNet block diagram-clearly this is a standard feed-forward network with a residual connection included. This layer's (l-1)th outputs can be identified as those given by the previous layer (x_l-1). $F(x_l$-1) results from several operations, such as convolution with variable-size filters, or batch normalization, before applying an activation function like ReLU on (x_l-1). The residual network has a large number of fundamental residual blocks. Operations in the residual block are likewise affected by the residual network design.

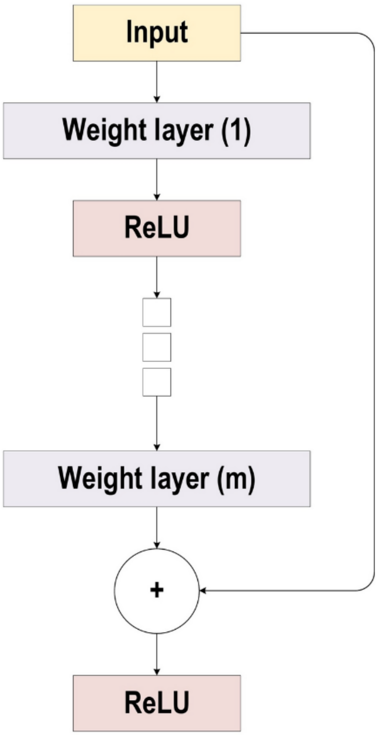


Figure 4.17 The block diagram for ResNet.

For cross-layer connections, ResNet introduced shortcuts within layers that are parameter-free and data-independent, unlike the highway network. Layers describe non-residual functions in the highway network when a gated shortcut is closed. On the other hand, ResNet never closes individuality shortcuts, yet residual information is always transmitted. Another advantage of ResNet is that its shortcut connections (residual links) can help prevent gradient fading and speed up deep network convergence. In the 2015-ILSVRC competition, ResNet has 152 layers of depth, eight times as deep as VGG and twenty times as deep as AlexNet. Even with increased depth, it has a smaller computational burden than VGG.

4.5.7 Inception

An improvement to Inception-V1/2 called Inception-ResNet and Inception-V3/4 was suggested by Szegedy et al. [63, 115]. The idea of Inception-V3 was to reduce the computational cost without affecting the generalization of the network. A bottleneck of convolution before the large-size filters was also used by Szegedy et al. [116]. Consequently, they chose small-size filters (15 and 17) rather than larger filters (77 and 55). Thanks to these developments, traditional convolution is now highly comparable to cross-channel correlation. Previously, Lin et al. used the NIN architecture's 1×1 filtering capabilities. After that, they creatively used the same concept. 3 or 4 isolated spaces are created by applying the 11 convolutional operations in Inception-V3 to transform input data. A 5×5 or 3×3 convolution can then map these relationships in these smaller regions. The residual connection replaces the filter concatenation in Inception-ResNet, which Szegedy et al. use it to combine the inception block and residual learning power [117]. As Szegedy et al. demonstrated, Inception-ResNet (Inception-4 with residual connections) has the same generalization capacity as Inception-V4 with an expanded width and depth but no residual connections. Inception network training can be considerably accelerated by exploiting residual connections in training. Inception Residual unit's fundamental block diagram is shown in Figure 4.18.

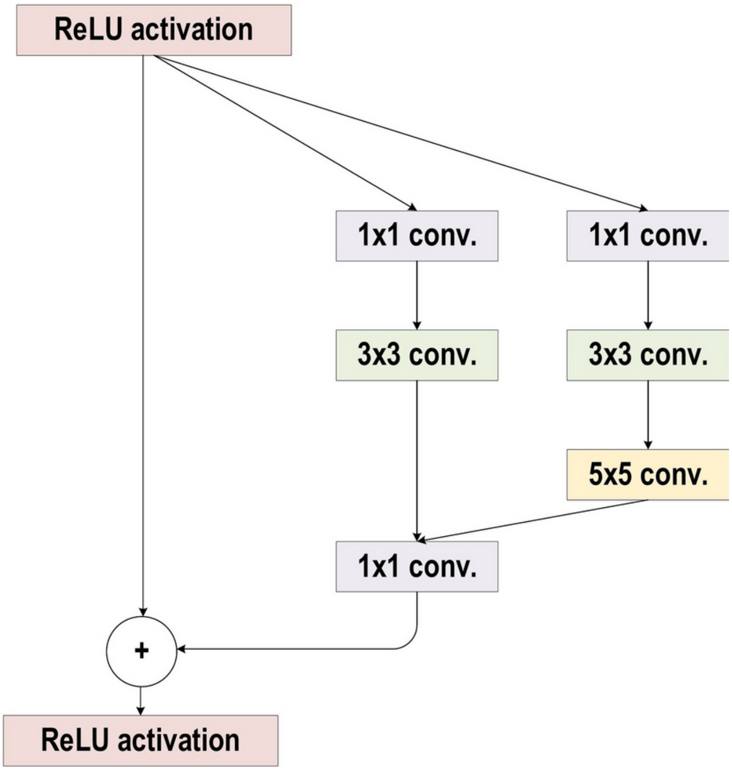


Figure 4.18 The basic block diagram for Inception Residual unit.

4.5.8 DenseNet

DenseNet, following the same path as ResNet and the Highway network, was proposed to address the issue of the vanishing gradient [117, 118, 119]. As numerous layers offer little or no information, ResNet has the drawback of visibly conserving information by preserving individuality modifications. Because each layer has its own set of weights, ResNet has a high number of weights. Cross-layer connectivity was used by DenseNet as a better technique to deal with this problem [120]. It used a feed-forward strategy to connect every layer to every other layer in the network. Therefore, the feature maps from previous layers were used to populate the data in all subsequent ones. Compared to typical CNNs, DenseNets have $\frac{l(l+1)}{2}$ direct connections between the previous and the current layers. As shown by DenseNet, cross-layer depth-wise convolutions have an effect. Because DenseNet concatenates the features of the preceding layers, rather than adding them, the network is better able to distinguish between newly added information and information that has been kept. DenseNet, on the other hand, is more expensive because of its limited layer structure, which makes it more expensive per feature map. Loss-function admittance of all layers to gradients increases network-wide data flow significantly. A regularizing effect is also included, reducing overfitting on tasks and small training batches. DenseNet Network's structure is shown in Figure 4.19.

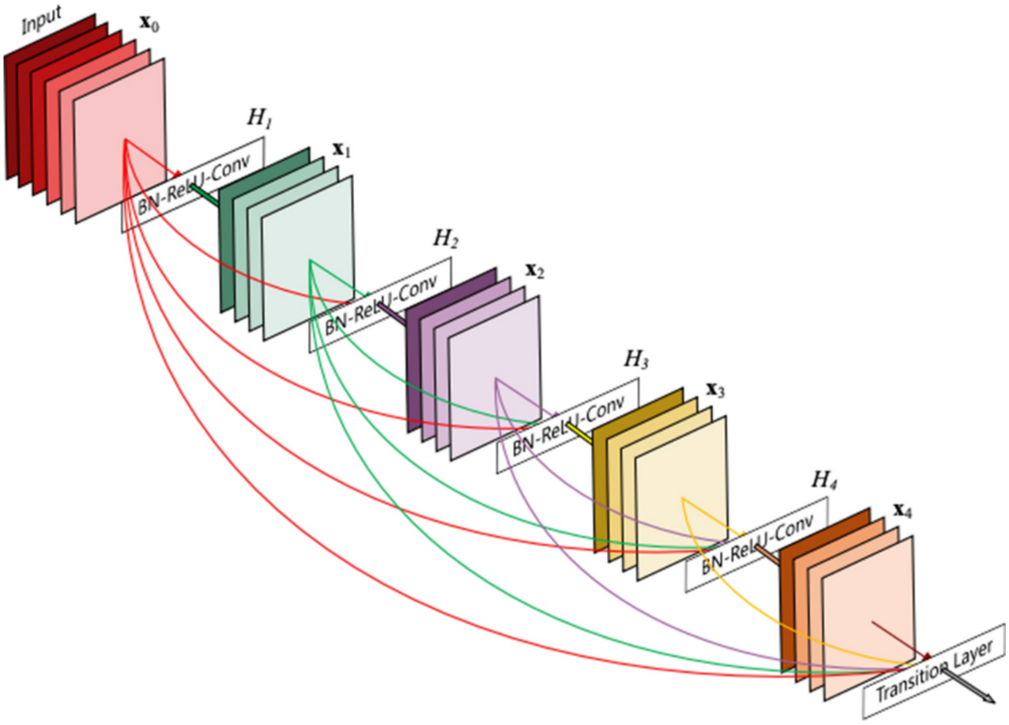


Figure 4.19 The architecture of DenseNet Network.

4.5.9 ResNext

A newer version of the Inception Network [79] is called ResNext. The Aggregated Residual Transform Network is another name for it. The network introduced a new concept, "cardinality," which uses the split, transform, and merge topology. The extra dimension [121] indicates the size of the transformation set. The Inception network, on the other hand, improves the traditional CNN's learning capabilities while also better managing network resources. Different spatial embeddings (e.g., 5×5, 3×3, and 1×1) are employed in the transformation branch. As a result, each layer must be customized individually. On the other hand, ResNet, VGG, and Inception provide the basis of ResNext's distinctive features. The split, transform and merge blocks used the VGG deep homogeneous topology with GoogleNet's fundamental architecture by using 3×3 filters as spatial resolution. ResNext's building blocks are depicted in Figure 4.20. Within the divide, transform, and merge blocks, ResNext implemented multitransformations while also describing these transformations in cardinality terms. It has been shown that raising the cardinality has a considerable impact on performance. In order to keep ResNext's complexity in check, low embedding filters were used before a 3×3 convolution. Dropping connections, on the other hand, is employed to improve training.

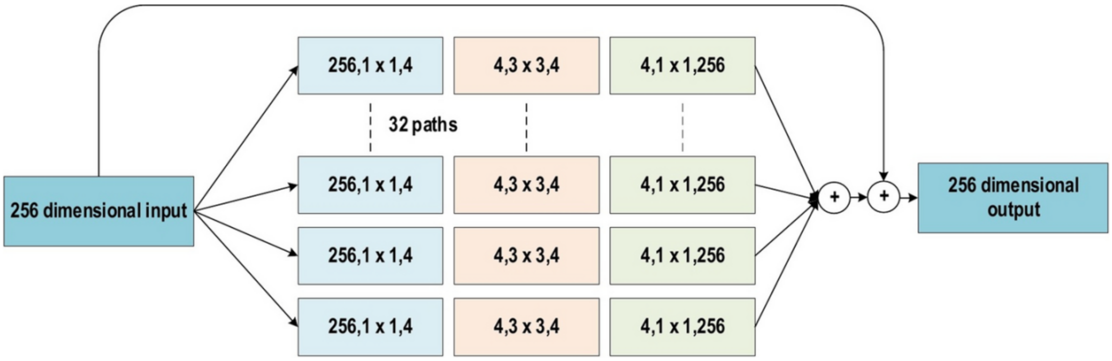


Figure 4.20 The basic block diagram for the ResNext building blocks.

4.5.10 WideResNet

Deep residual networks suffer from a phenomenon known as feature reuse, in which some feature blocks or transformations have a negligible impact on learning. WideResNet was presented as a solution by Zagoruyko and Komodakis [122]. Deep residual networks' core learning abilities are conveyed through residual units that have a supplemental effect on the network's depth. Instead of deepening the ResNet, WideResNet widened it to take advantage of the leftover block power. A new factor, k, was included to deal with network width, which increased the overall width. Layer broadening was a more practical approach to enhancing performance than deeper the residual network. While enhancing representational capacity, deep residual networks have many downsides, like the exploding and vanishing gradient difficulties, feature reuse, and the time-intensive nature of the training. In order to efficiently regularize the network, the authors in [78] included a dropout in each residual block to address the issue of feature reuse. Dropouts were used similarly by Huang and colleagues to tackle the problems of sluggish learning and gradient vanishing. [122] Research in the past has been focused on increasing the depth, so even a slight performance improvement required adding several new layers. WideResNet has twice as many parameters as ResNet, according to an experimental investigation. However, WideResNet offers a better training strategy than deep networks [123]. As a reminder, most previous designs, notably VGG and Inception, were far broader than ResNet. Once this was established, more extensive residual networks were built. Adding a dropout between the convolutional layers (rather than within the residual block) improved learning in WideResNet [124].

4.6 Applications of Deep Learning

Many industries can benefit from deep learning, including healthcare, banking, and image identification [125]. Let us have a look at a few applications in this part:

- **Healthcare:** Deep learning is well-suited to healthcare applications because of the abundance of data and the ease with which it can be used. Image recognition technology has surpassed the accuracy of cancer detection from Magnetic Resonance Imaging (MRI) and x-ray images. Genomics, clinical testing matching, and drug development have also been significant healthcare-based applications.
- **Autonomous vehicles:** Autonomous driving is a dangerous endeavor, but it has just made a turn toward becoming more commonplace in our daily lives. Training and testing deep learning-based models are done in a simulated environment to see how well they do in the real world.
- **E-commerce:** Product suggestions are one of deep learning's most popular and profitable uses. Customers benefit from more personalized and accurate recommendations because they can quickly shop for what they are looking for and see all of their available possibilities. Additionally, this speeds sales, which helps sellers.
- **Personal assistant:** Having a personal assistant is now as simple as purchasing a device like Alexa or Google Assistant. Deep learning is used by these intelligent assistants in various ways, including tailored speech and accent recognition, personalized suggestions, and text production.

Deep learning has many potential applications, and these are just a few examples. Deep learning has also helped forecast the stock market and predict the weather. Figure 4.21 shows some examples of DL applications.

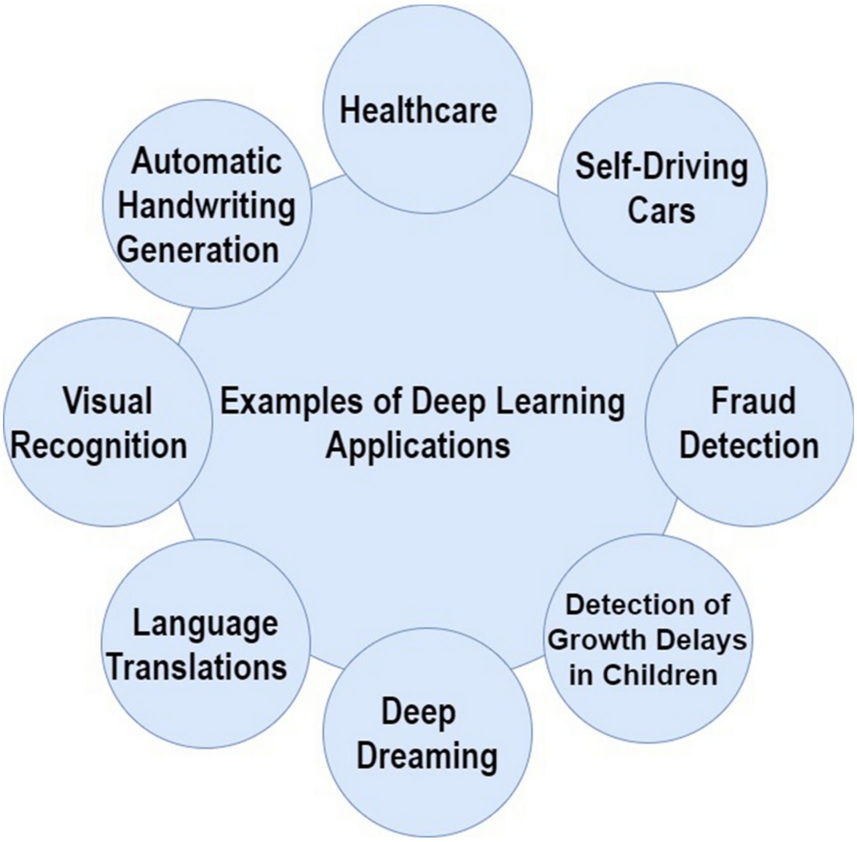

Examples of Deep Learning Applications
Healthcare
Self-Driving Cars
Fraud Detection
Detection of Growth Delays in Children
Deep Dreaming
Language Translations
Visual Recognition
Automatic Handwriting Generation

Figure 4.21 Examples of DL applications.

4.7 Challenges in Deep Learning

Today's deep learning techniques are highly data-intensive and many complicated issues, such as language translation and lack adequate data sets. Deep learning methods for neural machine translation to and from low-resource languages frequently perform poorly. However, in recent years, domain adaptation strategies (applying learnings from high-resource systems to low-resource settings) have shown promise performance. It can be challenging to generate such a large volume of data for tasks such as pose estimation. The synthetic data on which the model is finally trained differs significantly from the "in-the-wild" environment in which the model must perform [126, 127].

Even while deep learning algorithms have been shown to outperform humans in terms of accuracy, there is no obvious way to backtrack and explain each prediction made; this makes it challenging to employ it in applications such as finance, where it is required to justify each loan approval or rejection [128].

Another dimension that frequently causes problems is an inherent bias in the data, resulting in the model performing poorly on critical subsets of the data. Learning agents that employ a reward-based mechanism occasionally cease to act ethically, as all that is required to minimize system error is the maximization of the reward they accrue.

Due to the large number of parameters involved, which are intricately connected, DL models have an extremely high risk of resulting in data overfitting during the training stage. Such circumstances impair the model's ability to perform well on the tested data [129]. This issue is not restricted to a single field but encompasses a variety of tasks. As a result, while proposing DL approaches, this issue should be thoroughly explored and addressed appropriately. As recent research demonstrates [129, 130], the implicit bias of the training process enables the model to overcome critical overfitting difficulties in DL.

Nonetheless, it is vital to create ways to deal with the overfitting problem. Examining the various deep learning techniques that alleviate the overfitting problem reveals three distinct types. The first class affects both the model architecture and parameters of the model and contains the most well-known methods, such as weight decay , batch normalization [131], and dropout [106]. In DL, the default approach is weight decay [132], which is commonly utilized as a universal regularizer in practically all machine learning algorithms. The second class is concerned with model inputs, including data corruption and augmentation [124]. One cause of overfitting is a shortage of training data, which causes the learned distribution to deviate from the accurate distribution. The term "data augmentation" refers to increasing the size of the training data. By contrast, marginalized data corruption benefits the remedy unique to data augmentation. The final class is concerned with the model's output. A recently proposed strategy penalizes overconfident outcomes for model regularization [133]. This technique is capable of regularizing RNNs and CNNs.

4.8 Conclusion

In this chapter, we discussed deep learning, a popular branch of artificial intelligence that is now on the rise. The term "deep learning" refers to a branch of machine learning that is entirely based on artificial neural networks. Because neural networks are intended to imitate the human brain, deep learning is also considered a type of mimic of the human brain. Deep learning eliminates the need to program every aspect of the system explicitly. Deep learning is not a new notion in computer science. It has been around for several years at this point, but it is all the rage these days since we did not have nearly as much processing power or as much data when we first started. Since the processing power has increased tremendously over the past 20 years, deep learning and machine learning have emerged as viable options.

We exploited the power of different deep learning approaches and architectures to achieve outstanding results in all our contributions, as detailed in the following chapters.

CHAPTER 5: A new scheme for gray-level ear images recognition

5.1 Introduction

We address one specific issue in this chapter: the absence of color in test images. It is a challenge of feeding grayscale test images to a color-trained classification model. This issue has a detrimental effect on the overall recognition rate.

Image-to-Image translation [134] contributes significantly toward resolving various image processing challenges, including image generation and colorization via Conditional Deep Convolutional Generative Adversarial Networks (cDCGAN). The colorization of grayscale images is adapted for high-resolution images and optimized in terms of speed and stability [135].

However, no existing study has examined the feasibility of colorizing grayscale ear images to improve recognition rates, indicating a general shortage of research on this subject. As such, this chapter attempts to add to the rapidly emerging field of ear biometrics a new framework for recognizing grayscale ear images with nearly the same efficiency as color image graphs. The proposed system colorizes grayscale test images using cDCGAN before feeding them to a classification model. We conducted an additional experiment to demonstrate that providing grayscale images for the training process is insufficient to identify predicted grayscale test images.

This chapter is organized as follows: In the 2nd section, we introduce generative adversarial networks (GAN). Next, we explain how to perform image colorization using GANs. In the fourth section, we review the used ear datasets. In the fifth section, we introduce the proposed framework, and in the 6th section, we carry out experimental work. Section 7 concludes the chapter.

5.2 Generative Adversarial Networks

To produce new synthetic data instances, Goodfellow et al. [136] suggested GANs. A neural network, referred to as the generating model, is pitted against another adversary network, referred to as the discriminative model, to evaluate if a given sample is genuine or created by the generative model. Both the generator and discriminator are trained concurrently to train the generator to generate samples that the discriminator cannot tell apart from the original. In image processing, both the generator and discriminator are CNNs; thus, we arrive at cDCGAN [137].

The generator learns to transfer a random noise vector *z* to an output image *x*. It is represented by the mapping function $G(z,\theta_g) \rightarrow x$, where *G* is a CNN in the image processing area, with θ_g parameters. On the other hand, the discriminator is represented by a second CNN, *D(x,*θ_d*)*, which determines if x is a *G* creation or legitimate. *G* and *D* engage in a minimax game in which *G* seeks to minimize the probability that *D* correctly labels *x. D,* on the other hand, seeks to maximize that likelihood; this can be stated mathematically using the formula found in Eq. 5.1:

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

where *V(G, D)* is the value function, $\mathbb{E}_x$ is the expected value over all authentic images, and $\mathbb{E}_z$ is the expected value over all generated images G(z).

5.3 Colorization using cDCGAN

Our objective is to colorize grayscale images, so we cannot feed the generator only random noises. These grayscale images are the input to G; in this case, we will employ a variation of GAN called conditional GAN [138].

By taking an additional set of inputs *x* alongside a random noise vector *z*, a conditional GAN learns to map an output *y*. For the colorization problem, the extra information is grayscale image graphs without random noise, which may be described mathematically as $G(0_z | x)$. Likewise, the discriminator must be adjusted to account for the addition of the conditional CNN. It receives as dependent input color images from *G* and the original dataset, as well as grayscale images. Then it attempts to determine which one from the colored images is the true one.

Figure 5.1 depicts how the generator and discriminator compete to colorize and compare images to their ground truth counterparts; this can be described mathematically as the following final cost function:

$$min_G \, max_D \, V(G, D) = \mathbb{E}_x[\log D(y|x)] + \mathbb{E}_z\left[1 - \log D\left(G(0_z|x)\right)\right] \qquad (5.2)$$

The main objective is to train the entire model to reduce the average Euclidean Distance between the colored image and the ground truth at the pixel level:

$$Dist(x, \theta) = \frac{1}{3nm} \sum_{c=1}^{3} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| h(x, \theta)_{i,j}^{c} - y_{i,j}^{c} \right\|_2^2 \qquad (5.3)$$

where *x* is the grayscale image, *y* is the ground truth, θ is the image colored by the generator, *h* is the function that converts from grayscale to color images, *c* is the channel index, *i* and *j* are coordinates of pixels.

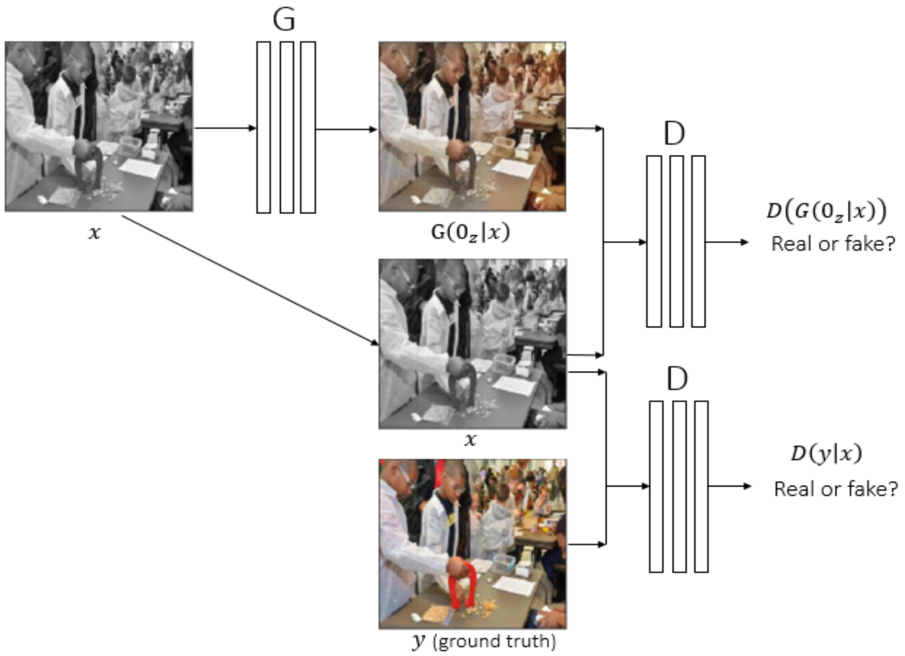


Figure 5.1 Colorization using conditional GAN.

The architectures for the generator/discriminator can be derived from a variety of earlier publications on deep neural networks. For the image-to-image translation problem, [134] used the sequence Conv-BatchNormalization-ReLu for both the generator and discriminator [131]. In image-to-image translation problems, the underlying structure of the input and output are identical; their edges and forms are identical. It is a pixel-wise regression problem [139]. As a result, low-level information must be transferred between symmetric levels via skip connections between layer *i* and layer *n-i* under the U-Net architecture [140].

5.4 Ear datasets

5.4.1 AMI ear dataset

Esther González[2] generated the AMI ear dataset as part of her doctoral research in computer science. It contains uncropped images of the ears taken from 100 people for 700 images taken in an indoor environment. These images have a resolution of 492×702 pixels and were created in *jpeg* format. As described in Table 5.1, each individual has seven images, six of the right ear and one of the left ear.

[2] http://ctim.ulpgc.es/research_works/ami_ear_dataset

Table 5.1 AMI ear images variations.

Profile	Description
ZOOM	Captured with 200mm focal length
FRONT	135mm focal length, subject facing forward
UP	135mm focal length, looking up
DOWN	135mm focal length, looking down
LEFT	135mm focal length, looking left
RIGHT	135mm focal length, looking right
BACK	135mm focal length, left side ear

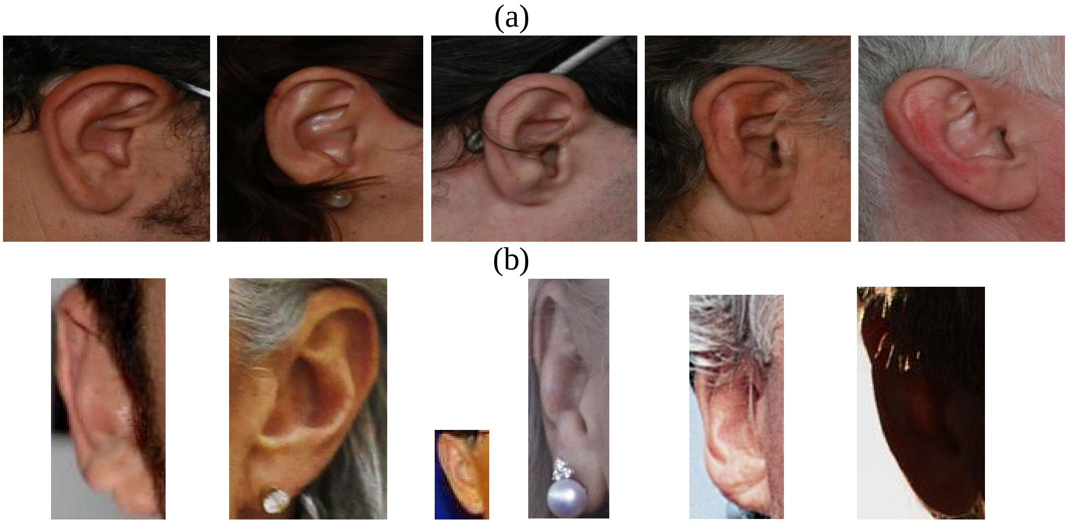


Figure 5.2 Sample ear images from (a) AMI dataset and (b) AWE dataset.

5.4.2 AWE ear dataset

The Annotated Web Ears (AWE) [74, 75] dataset includes 1000 images (left/right) of 100 distinct individuals. These images of public persons were gathered from the internet; they range in size from 473×1022 to 20×32 and are regarded as one of the most difficult ear datasets to deal with in unconstrained conditions. Variations include head rotation, gender, race, occlusion, light, and blurring. Figure 5.2 illustrates several images of various individuals from the AMI and the AWE datasets.

5.5 Proposed framework

In this chapter, we present a new framework that comprises two models: a cDCGAN model for image colorization and a classification CNN model. Both models are trained sequentially on the same dataset. They both participate during the test phase to complete the colorization/classification procedure. The relationship between the two models is illustrated in Figure 5.3.

We used symmetric model architecture with four encoding and decoding blocks to colorize biometric grayscale images. The color space utilized for colorization is CIELab (L*a*b*), where L* denotes lightness and b*, a* denotes color information; this eliminates any visible rapid changes in color or brightness in the RGB color space.

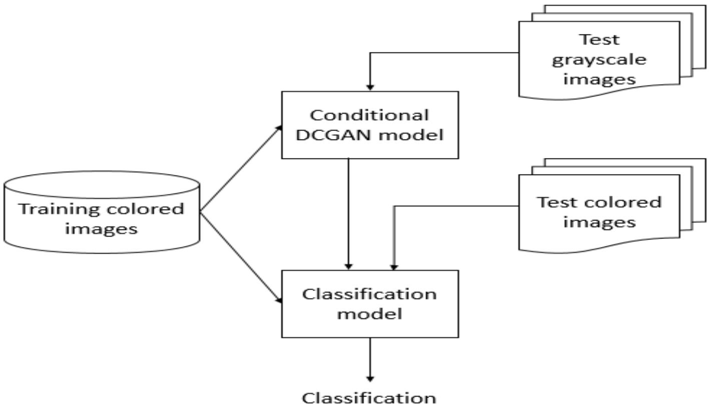


Figure 5.3 The proposed framework.

As depicted in Figure 5.4, the encoding process starts with convolutional layers with filters of size 4×4, then pooling layers applied in 2×2 patches, batch normalization, and the activation function Leaky-ReLu [141]. After upsampling with a 2×2 stride, the decoding process composed of a series of transposed convolutional layers is concatenated with mirror layers from the encoding side. Following batch normalization with the ReLu activation function, the final convolution layer with a 1×1 filter is used as a cross-channel pooling layer. The decoding process concludes with a three-channel output layer, L*a*b*.

The discriminator architecture is built of five convolutional layers is shown in Figure 5.5. The first four layers employ 4×4 (stride 2) filters, followed by batch normalization and an activation filter based on leaky-ReLu. The last layer uses a 4×4 filter striding by one and activating it with the sigmoid function to produce a simple scalar generated by averaging the previous 16×16 patch. This scalar represents the likelihood that the input image is authentic or fraudulent.

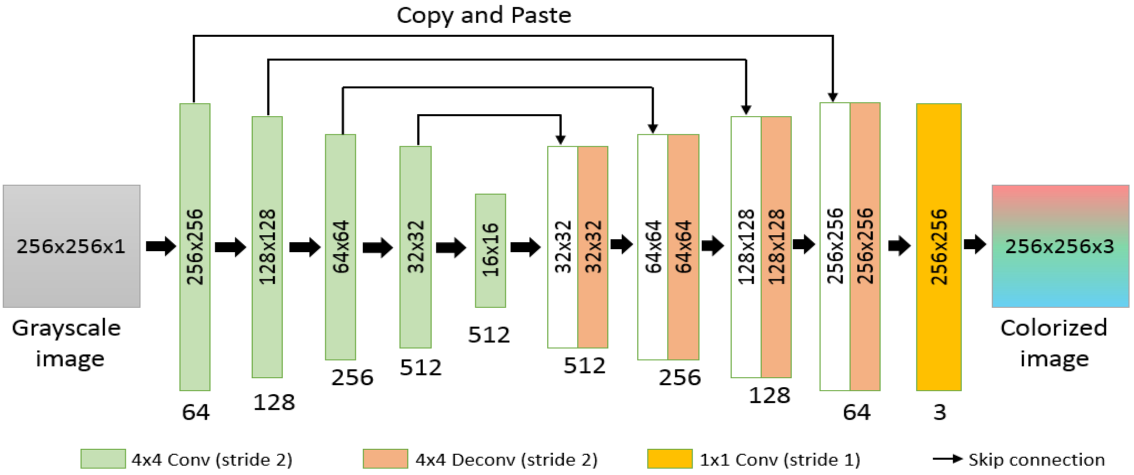


Figure 5.4 The convolutional U-Net architecture of the Generator.

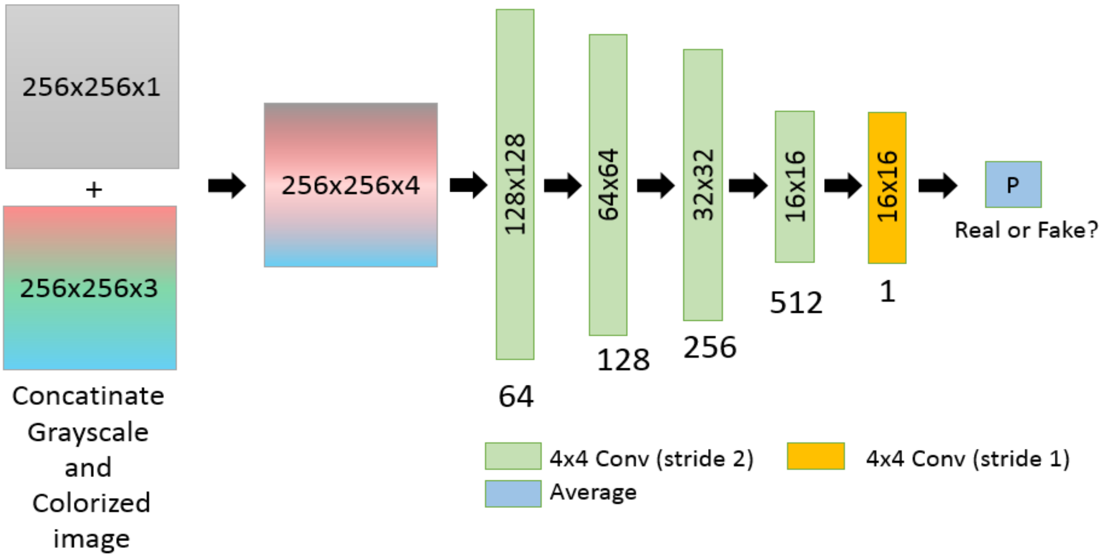


Figure 5.5 The architecture of the Discriminator.

Numerous pre-trained CNN-based classification models exist in the literature. We considered only simple models in our experiments. We did not use model ensembles because optimizing the performance of a single deep model results in optimizing the performance of deep model compositions.

We investigated the AlexNet, VGG16, and VGG19 architectures for the classification model. On top of the pre-trained convolutional layers, we added a final fully connected layer and a softmax output layer. The CNN-based models used in this study were pre-trained using the ImageNet dataset [148]. The global architecture of the resulting classification model is depicted in Table 5.2.

Table 5.2 Classification model's architecture.

Layer	Neurons number	Activation function	Drop
	<pre-trained model Convolutional Layers>		
Dense layer	1024	ReLu	50%
Dense layer	100	Softmax	-

5.6 Experimental results

5.6.1 Datasets

Preparation of the train/test dataset is the initial stage. We have 700 ear images from 100 participants in the AMI dataset. As a result, we selected four images for training and three for testing for each individual. Additionally, all images have been downsized to fit the generative model at 256×256 pixels and the classification model at 224×224 pixels.

On the other hand, we used the provided AWE train/test split. The training set contains 600 image graphs (six images per participant), whereas the testing set has 400 images (four images per subject). We did not augment the training sets in all upcoming experimental scenarios, as opposed to previous works in [74, 137].

5.6.2 Performance metrics

The experiments are conducted according to a predefined experimental protocol. The classification model must identify an input ear image by determining which person it belongs to. We provided the following performance measurements and curves to assess the performance of the suggested approach.

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

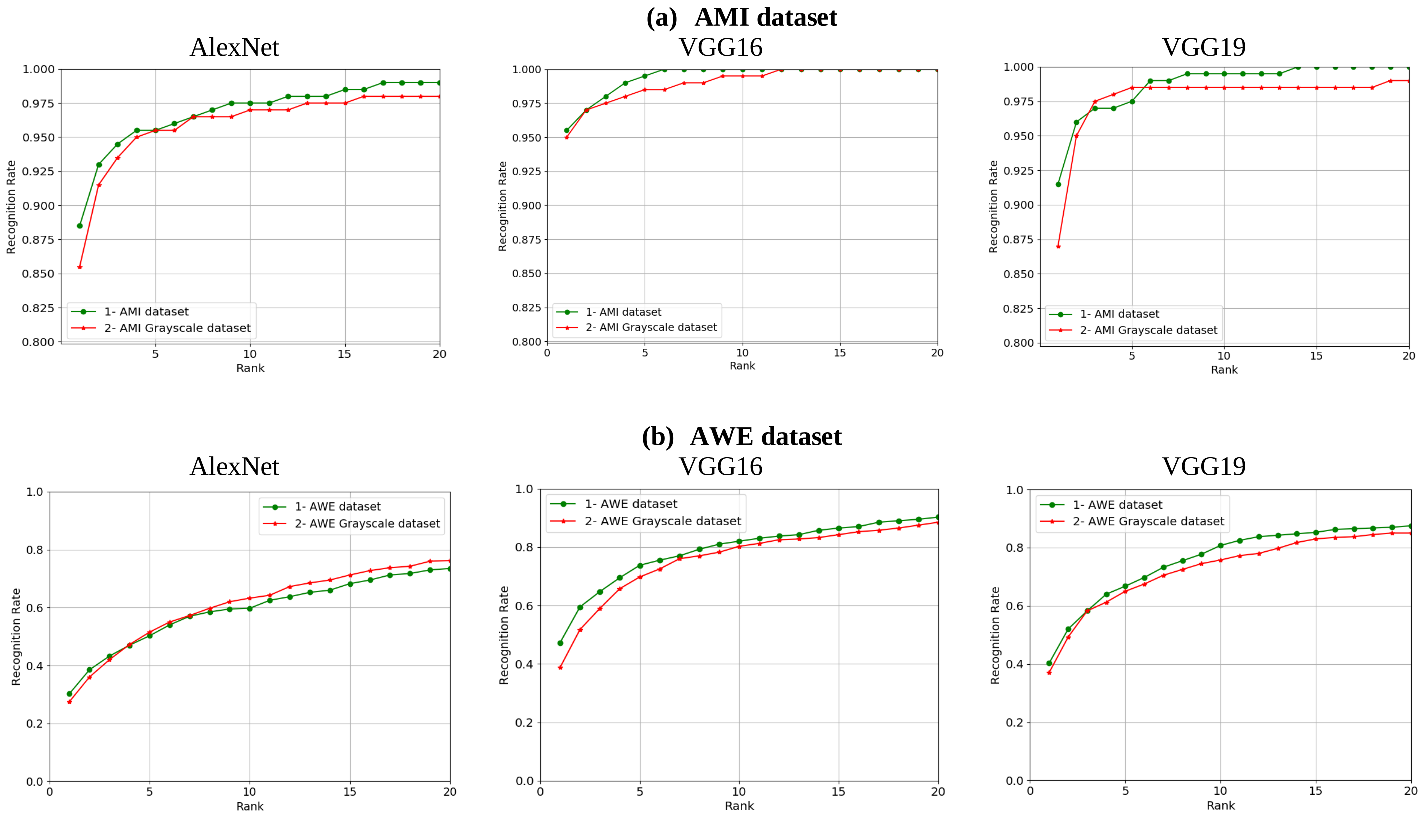
5.6.3 Experimental Scenario #1

The initial trial will provide insight into the detrimental effects of color loss. As a result, we fine-tuned the VGG-based classification models using only grayscale train images and then original color images, observing the results of test phases in both cases. We transformed the AMI and AWE dataset images to grayscale and then carried out the classification process, which included training, testing, and evaluating the model. Table 5.3 summarizes the experiment's findings. The recognition rate is significantly reduced for both the AMI and AWE datasets. Interestingly, training the model exclusively with grayscale images significantly decreased the identification rate, particularly for the AWE dataset. In general, these findings suggest that training the model with a grayscale version of the training dataset may not be sufficient for identifying grayscale test images.

Table 5.3 Classification process results.

Dataset	Rank-1 (%)			Rank-5 (%)			AUCMC (%)		
	AlexNet	VGG16	VGG19	AlexNet	VGG16	VGG19	AlexNet	VGG16	VGG19
AMI	**88.50**	**95.50**	**91.50**	95.50	99.50	97.50	92.11	94.56	93.91
AMI grayscale	85.50	95.00	87.50	95.50	98.50	98.50	91.38	94.17	93.07
AWE	**30.25**	**47.25**	**40.25**	50.25	73.75	66.75	56.54	75.41	72.44
AWE grayscale	27.50	38.75	37.00	51.50	69.75	65.00	58.16	72.64	69.60

According to the CMC data in Figure 5.6, there is a clear trend toward declining recognition rates for both datasets when all recommended models are used. These findings indicate a compelling need for a solution other than merely augmenting the training set with grayscale images. However, if the recognition rate margin obtained is insufficient, the convenience of adding a grayscale version of the training set might be traded off against the computational cost of our suggested system.


(a) AMI dataset

(b) AWE dataset

Figure 5.6 CMC curves of using original color images against grayscale images: (a) AMI dataset, and (b) AWE dataset.

5.6.4 Experimental Scenario #2

We employed three experimental scenarios to demonstrate the proposed framework's reliability and efficiency. We first used genuine colored images for CNN-based classification model fine-tuning, and we used the original test-colored dataset. While in the second situation, we conducted the test phase using grayscale test images converted from the original color image-graphs in the first scenario.

In the third case, we supplemented the training set with images that have been falsely colored. The generative model generated these images, identical in shape and edge to the originals but not in color. The training procedure is depicted in Figure 5.7.

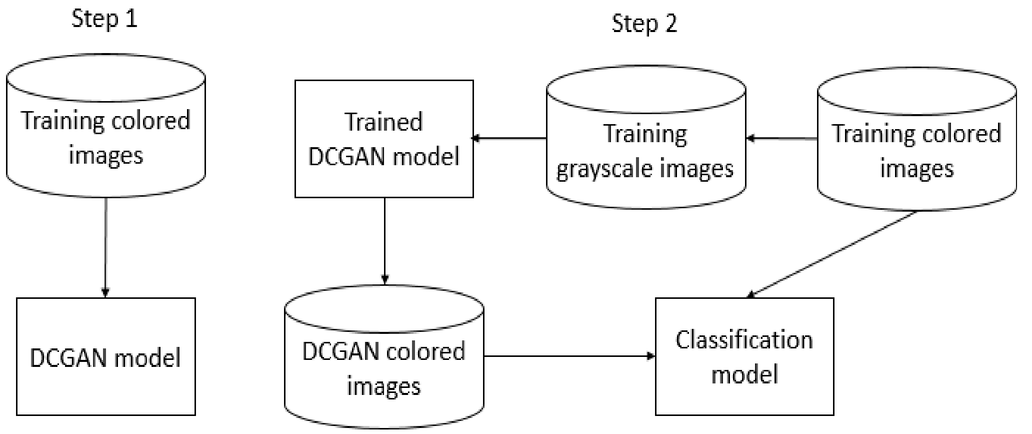


Figure 5.7 The training process in the third scenario.

The cDCGAN was trained using the original color training images to build another dataset of similar grayscale images as a condition input during training. To evaluate its performance, we employed accuracy, which is defined as the ratio of accurately colored pixels to total pixels. A pixel is said to be "accurately colored" if the difference between its colors and the original colors is smaller than a preset threshold ϵ_c for each color channel *c*, as mathematically represented in Eq. 5.4:

$$Accuracy(x, y) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(\prod_{c=1}^{3} 1_{[0,\epsilon_c]}(|x_{i,j}^{c} - y_{i,j}^{c}|)) \qquad (5.4)$$

where *x* is the colorized image, *y* is the associated image of the ground truth image, $1_{[0,\epsilon_c]}$ is the indicator function, *n* and *m* are image dimensions, *i* and *j* are pixel index counters, and *c* is the color channel.

We trained the cDCGAN model for 144 mini-batch iterations in total for the AMI dataset scenarios, obtaining a distance loss of 2.92 for the generator, 1.36 for the discriminator, and a colorization accuracy of 67.86%, as illustrated in Figures 5.8 and 5.9.

In a similar scenario for the AWE dataset, we trained the colorization model using 587 images from 600 train splits; 13 images were eliminated since they were grayscale in the first place and hence could not be used for learning. The model's final accuracy was 24.41%.

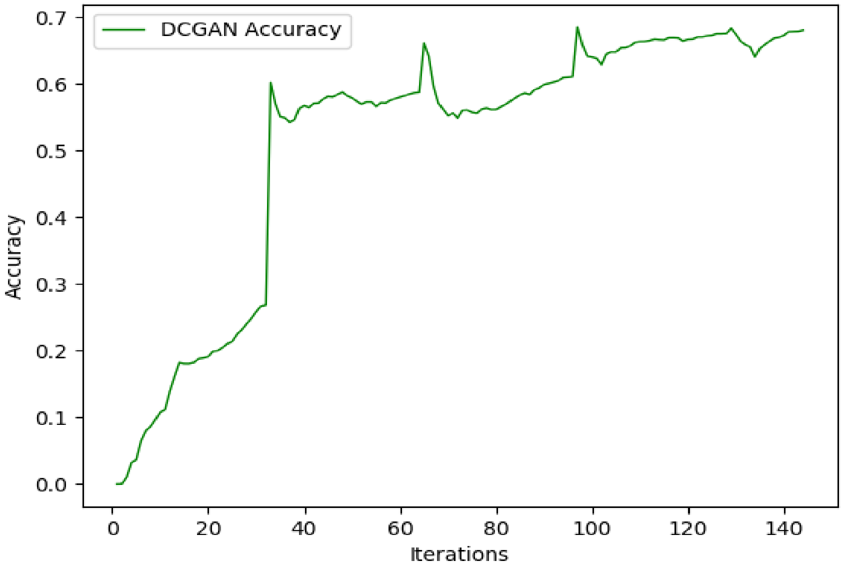


Figure 5.8 cDCGAN training accuracy using the AMI dataset.

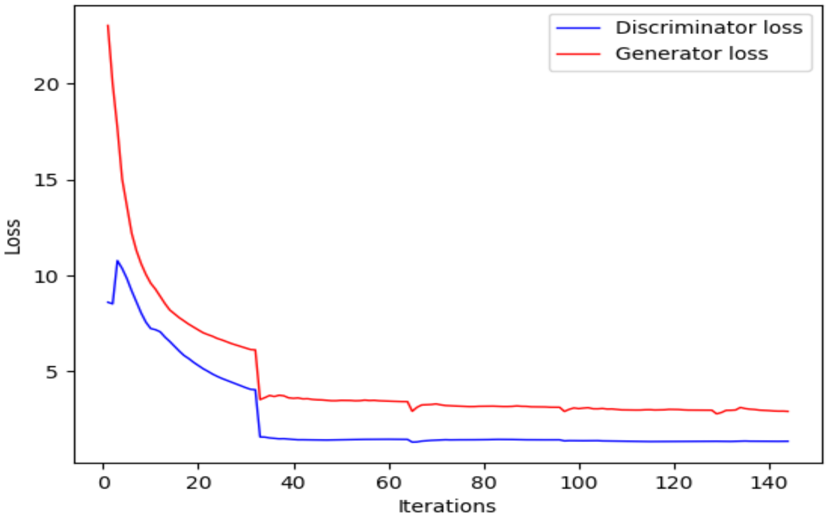


Figure 5.9 Generator/Discriminator training loss.

The colorized images created by the cDCGAN generator model have good quality. The resulting colorized images are shown in Figure 5.10 alongside their original ground truth. We did not restore only missing colors but also equalized color intensity, and brightness using the generative model, which improved the classification identification rate, particularly for the AWE unconstrained images.

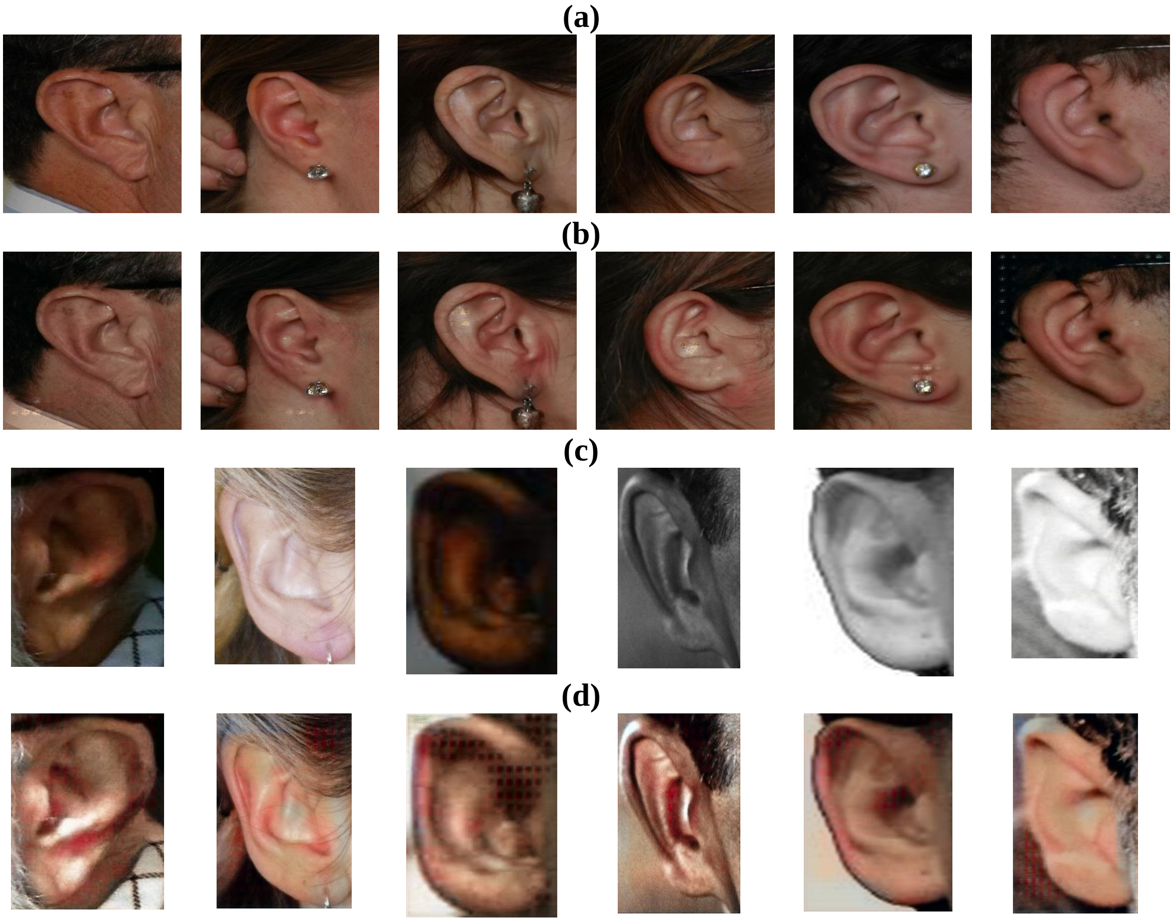


Figure 5.10 Colorization results with the cDCGAN model: (a) Original AMI images, (b) Colorized AMI images, (c) Original AWE images, and (d) Colorized AWE images.

We pre-trained all CNN-based classification models from the ImageNet dataset in our proposed scheme. These models' only purpose is to classify incoming images and determine which individual an image belongs to. We performed the fine-tuning process on the AMI dataset using 400 images from the training set for scenarios 1 and 2. We supplemented the training dataset with 400 new images colorized using the previously developed cDCGAN model. Similarly, as with the AMI dataset, we experimentally fine-tuned the classification models for the AWE dataset using the training set.

Table 5.4 summarizes the recognition rates for rank-1 and rank-5 in the three scenarios stated previously, using the AUCMC. The italicized values represent the results of tests using grayscale images, while the bold values represent the results of a third scenario utilizing colorized images. The collected results demonstrated the severe detrimental influence of color loss on the performance of classification processes.

We evaluated the classification models in the first case using just the original unaltered color images, with no data augmentation. The AlexNet model achieved a Rank-1 recognition rate of 88.50% when trained on the AMI dataset. Whereas VGG16 had the highest rank-1 recognition rate of 95.50%, it was outperformed the VGG19-based model, which had a recognition rate of 91.50%. In the second case, where we gave grayscale images to the classifier, the recognition rate for all models decreased significantly. The VGG19-based model was the most affected, with a recognition rate of 68.50%. These findings demonstrated the critical nature of the color information contained in the original test images, which we duplicated in the following situation.

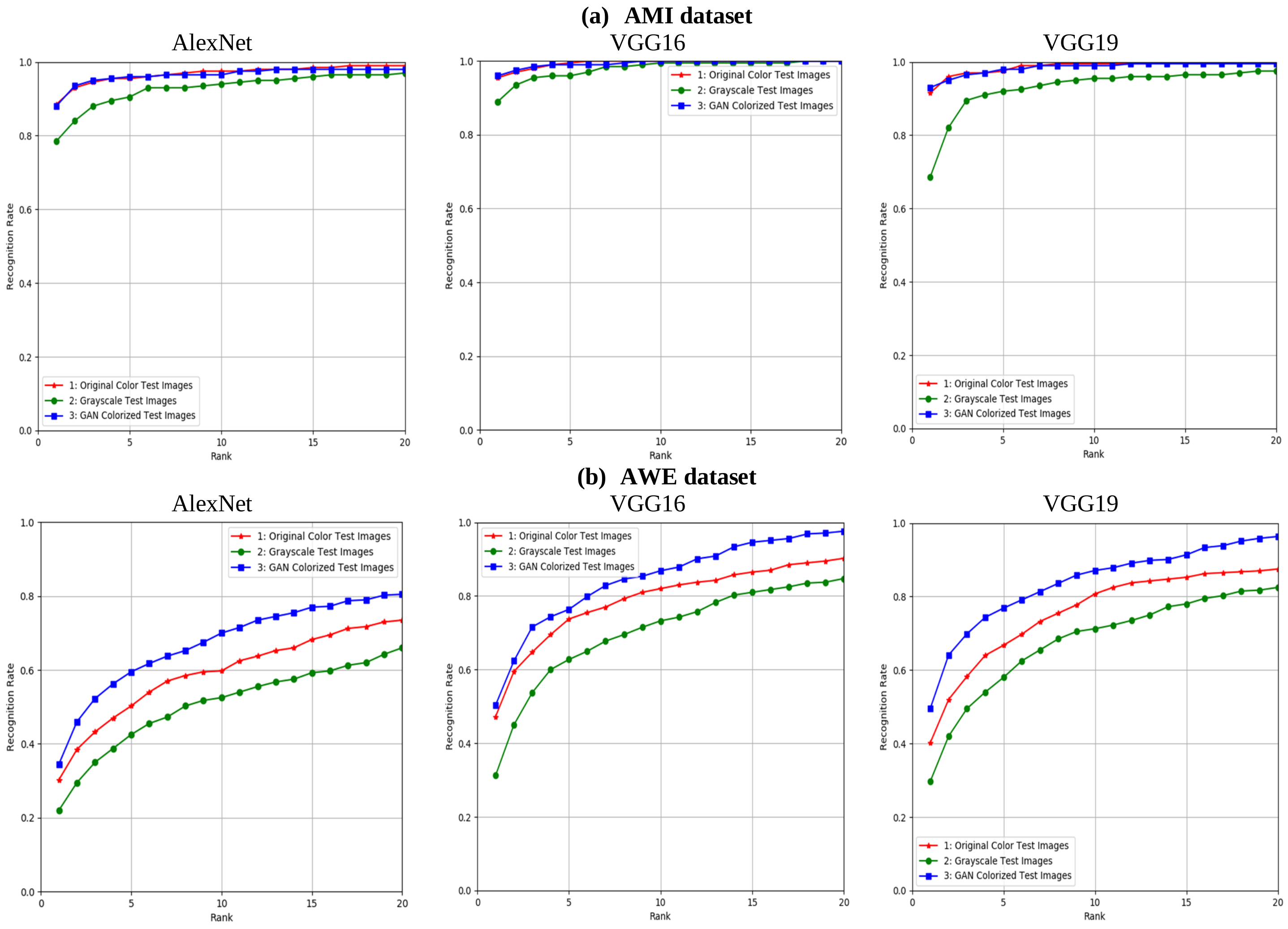
Table 5.4 Resultant performance metrics that show the effect of color information loss.

Dataset	**Scenario**	**Rank-1 (%)**			**Rank-5 (%)**			**AUCMC (%)**		
		AlexNet	VGG16	VGG19	AlexNet	VGG16	VGG19	AlexNet	VGG16	VGG19
AMI	1	88.50	95.50	91.50	95.50	99.50	97.50	92.11	94.56	93.91
AMI	*2*	*78.50*	*89.00*	*68.50*	*90.50*	*96.00*	*92.00*	*88.41*	*93.22*	*88.80*
AMI	**3**	**88.00**	**96.00**	**93.00**	**96.00**	**99.00**	**98.00**	**91.79**	**94.47**	**93.58**
AWE	1	30.25	47.25	40.25	50.25	73.75	66.75	56.54	75.41	72.44
AWE	*2*	*22.00*	*31.25*	*29.75*	*42.50*	*62.75*	*58.00*	*48.36*	*67.37*	*64.84*
AWE	**3**	**34.50**	**50.53**	**49.60**	**59.50**	**76.35**	**76.85**	**64.35**	**80.97**	**80.07**

Not surprisingly, we restored the recognition rate and more in the last experiment in which we colorized all images using the cDCGAN model, much as we did in the AMI dataset experiments using the VGG16-based model; this was possible because of the ability to restore missing colors using a cDCGAN on the one hand and to alter the brightness and equalize the intensity on the other. The AWE dataset experiments produced similar results to the AMI dataset experiments. As shown in Table 3, the recognition rate for all three CNN-based classifiers decreased dramatically when grayscale images were used; however, when artificially colored test images were used, the classification accuracy was restored and even exceeded for all three CNN-based models.

The lack of color information was overcome in the third experiment by utilizing trained conditional cDCGANs, as illustrated in Figure 5.11 of the CMC curves. The recognition rate for the AMI dataset is essentially identical to that in the first scenario when using an AlexNet-based model, but it has improved by 1.50% and 2.50% when using VGG-16 and VGG-19-based models, respectively.

The AWE dataset scenario resulted in a 4.25% increase in the Rank-1 recognition rate over the original test; this increase is attributable to grayscale images in the original train and test datasets, which were colorized using the generative model. Additionally, the cDCGAN model's colorization equalized the intensity and lighting of train and test images. As a result, the classification model distinguished an increased number of ear images. We observed the same improvement of 3.28% and 9.35% for VGG16 and VGG19, respectively, when all VGG-based models were used.


(a) AMI dataset

(b) AWE dataset

Figure 5.11 CMC curves of (a) AMI dataset and (b) AWE dataset.

5.6.5 Comparison

To conduct a comprehensive study, we compared our obtained results to many recent studies on ear recognition. Table 5.5 compares and contrasts the suggested approach's rank-1 identification rate with other well-known and current representative approaches based on 2D ear images. As can be seen, our suggested strategy, which utilizes a cDCGAN model to colorize grayscale and dark images, shows extremely intriguing and competitive performance outperforming current work on ear biometrics under comparable settings.

Table 5.5 A comparison of rank-1of the proposed approach with other representative methods.

Publication	**Dataset**	**Data augmentation**	**Method**	**Rank-1**
Emeršič et al. (2017) [74]	AWE+CVLE+ 500 images	No	SqueezNet	41.26%
		Yes	SqueezNet	62.00%
Alshazly et al. (2019) [67]	AMIC	NO	VGG19	96.78%
	AWE	NO	VGG-face	50.00%
Kacar et al. (2019) [68]	WPUT+AWE +UERC	NO	ScoreNet	47.80%
Emeršič et al. (2017) [139]	AWE	NO	BSIF	48.40%
	AWE	NO	POEM	49.60%
Hassaballah et al. (2019) [60]	AMI	NO	CLBP	73.71%
Alshazly et al. (2018) [143]	AMI	NO	LOOP	72.10%
Chowdhury et al. (2017) [53]	AMI	NO	Tunable Filter	70.58%
This study	AMI	NO	cDCGAN+VGG16	**96.00%**
	AWE	NO	cDCGAN+VGG16	**50.53%**

5.7 Conclusion

One particular issue with ear recognition that we addressed in this chapter is the absence of color information in test images when fed to a model trained on colored images. To address this issue, we suggested an efficient system that utilizes a generative cDCGAN model for colorization of grayscale and dark images and a CNN-based classification model for classification.

We conducted the first experiment to demonstrate that dropping color information from an image harms the model’s accuracy, necessitating our proposed framework’s requirement to re-generate lost color. We used the AMI and AWE datasets to demonstrate that the proposed framework could restore missing color information. Hence, it restored recognition rates to levels comparable to those obtained when using original color images, if not higher, by equalizing intensity and stabilizing illumination using a cDCGAN model.

CHAPTER 6: Region-of-Interest Synthesis using Image-to-Image Translation

6.1 Introduction

We propose in this chapter to incorporate ear segmentation into the recognition process by exposing the input images to a segmentation operation to allow the classification model to handle only ear-related pixels. To obtain a region-of-interest (RoI) segmentation of the ear from the initial image, we employed Image-To-Image translation, i.e., Pix2Pix GAN, to generate it. By deleting as many occlusions as feasible during the image-mapping process, we could manipulate ear pixels to our advantage. Because the RoI synthetic segmentation eliminates all non-ear pixels, the feature extraction techniques and classification phase will concentrate exclusively on ear pixel-by-pixel segmentation. The suggested scheme was assessed using the AWE dataset, a recent and challenging ear dataset comprised of unconstrained ear photos acquired from the web. The segmentation operation was evaluated using pixel-level accuracy and Intersection Over Union (IoU) criteria. We then employed various local feature extraction algorithms to extract texture features for classification from the resulting photos. We repeated the feature extraction and classification studies using the original AWE dataset to emphasize the critical impact of segmentation. The obtained findings validated the efficacy of the suggested approach and demonstrated that feeding precise ear segmentation results in improved classification results.

The rest of the chapter is organized as follows: Section 2 presents the proposed image synthesis approach. Next, we review the RoI segmentation technique in the third section. The fourth section discusses feature extraction and classification using well-known methods. Experimental analysis is carried out in section 5 to conclude the chapter in the 6th section.

6.2 Proposed approach

The suggested ear recognition pipeline is divided into two phases: first, the RoI is synthesized, and then, local features are extracted and classified. The first step is to scale all photos to [-1, 1], as the *tanh* activation function is employed in the generative model's output layer, and the resulting images' pixel values will likewise be in the [-1, 1] range. Then, we use a trained Pix2Pix GAN to synthesize the RoI of each image. Following that, we isolate the RGB-color components of the image, divide them into non-overlapping blocks as necessary, and extract local features. Finally, as indicated in Figure 6.1, we concatenate the histograms before giving them to the classifier.

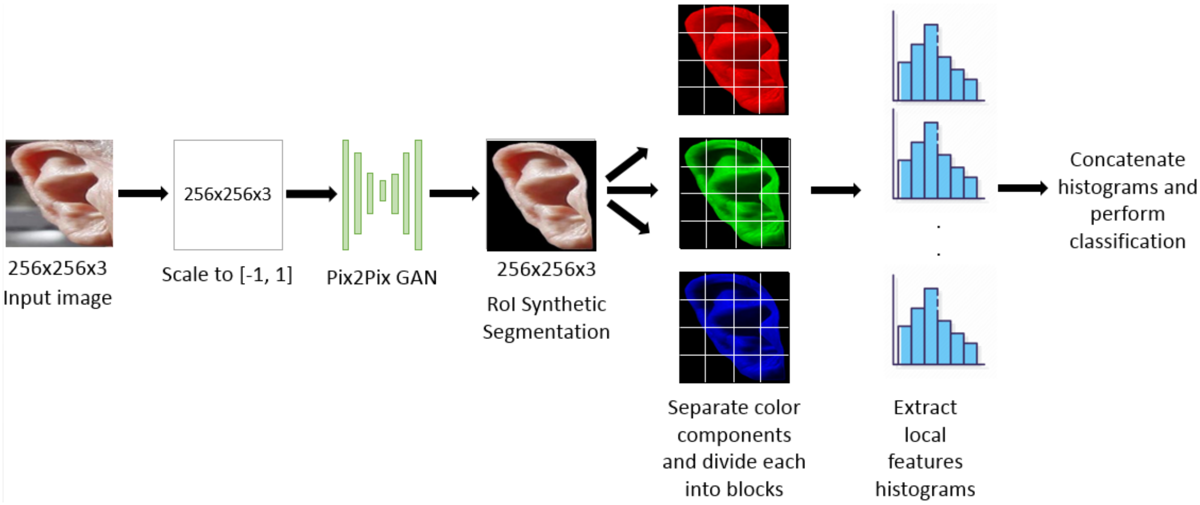

256x256x3
Input image
256x256x3
Scale to [-1, 1]
Pix2Pix GAN
256x256x3
RoI Synthetic
Segmentation
Separate color components and divide each into blocks
Extract local features histograms
Concatenate histograms and perform classification

Figure 6.1 The framework of the proposed method.

6.3 RoI segmentation synthesis

Image-to-Image translation has been widely utilized in recent years to accomplish various tasks, including image super-resolution, image painting, object transformation, and image enhancement. The overall objective is to discover the mapping between an input and an output image. Utilizing a Pix2Pix generative model is one approach to do this task.

We used a cDCGAN model to perform the RoI segmentation synthesis task; the decoding process contains a sequence of transposed convolutional layers concatenated with the mirror layer from the encoding side after upsampling using a 2×2 stride, batch normalization layer, and ReLu activation function. As shown in Figure 6.2, the final convolution layer applies upsampling using three 256×256 filters followed by the *tanh* activation function to generate the target image.

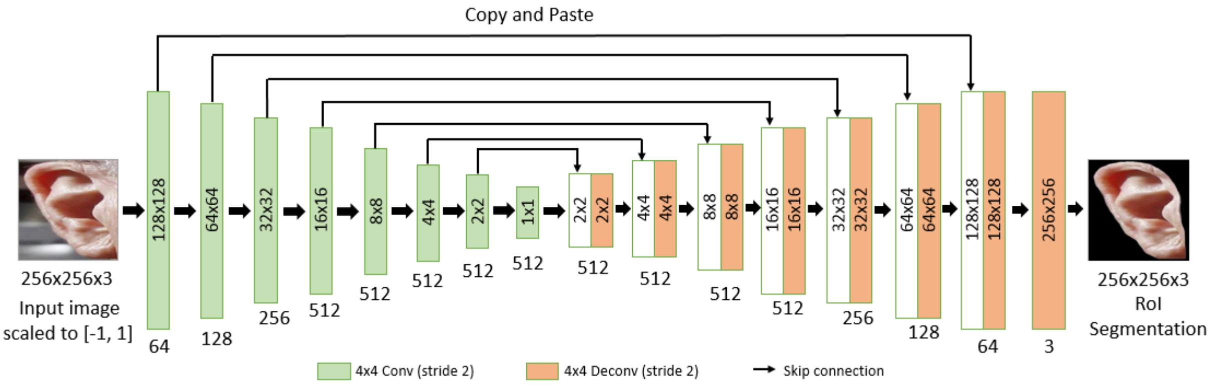


Figure 6.2 The architecture of the generator.

The discriminative model takes an input image and an unknown image and attempts to determine if the generator made the unknown image. Its architecture is more symmetrical than the generator's, as illustrated in Figure 6.3. It is a five-layer convolutional structure. The first four layers convolved with 4×4 filters, sliding by two steps. After each layer, batch normalization and the Leaky-ReLu activation function are applied. The final 30×30 layer represents the credibility of each of the input image's 70×70 patches; thus, it is named PatchGAN. The average of the output layer represents the likelihood that the image is genuine or a forgery.

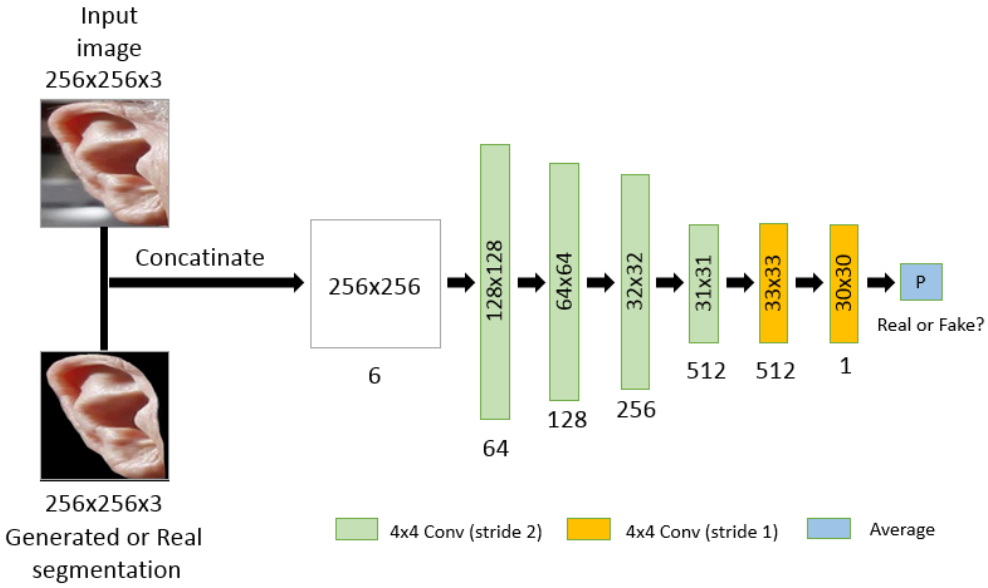


Figure 6.3 The architecture of the discriminator.

6.4 Feature extraction and classification

We extracted features using three well-known local feature descriptors, namely Local Binary Pattern (LBP) [54, 144], Local Phase Quantization (LPQ) [145], and Binarized Statistical Image Features (BSIF) [146]. The similarity of feature vectors is determined by calculating the chi-square distance between them, which is defined as follows:

$$\chi_{(x,y)} = \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - y_i)^2}{(x_i + y_i)} \tag{6.1}$$

where *x* and *y* are feature vectors of size *n*.

LBP is a texture analysis operator that is sometimes referred to as a gray-scale invariant texture measure. It is produced from a general texture description in a local neighborhood. It possesses a solid discriminative ability while using minimal computational resources. Each pixel's local patterns are retrieved by thresholding its neighboring pixels (*P*, *R*) with *P* sampling points on a circle of radius *R*. The equation gives the LBP code of each pixel (x_c, y_c) is expressed in Eq. 6.2:

$$LBP^{P,R}(x_c, y_c) = \sum_{i=1}^{p} S(g_j^{P,R} - g_c)\, 2^{i-1} \tag{6.2}$$

g_c and $g_j^{P,R}$ are the values of the central pixel and its neighbors, respectively, with *S(x)* defined as follows:

$$S(x) = \begin{cases} 1, & if\ x \geq 0 \\ 0, & otherwise \end{cases} \tag{6.3}$$

The LBP operator has been extended to capture large-scale structures in the image by utilizing different sized neighborhoods.

The LPQ technique was introduced to address LBP's relative sensitivity to blur. This technique quantizes the Fourier transformation phase in local neighborhoods. The phase is retrieved using the following formula over a rectangular *M*×*M* neighborhood N_x for each pixel position *x* of the image *f(x)*:

$$F(u, x) = \sum_{y \in N_x} f(x - y) e^{-i2\pi u^T y} \tag{6.4}$$

The transform in Eq. 6.4 can be computed independently for the rows and columns using 1-D convolutions. Then, only four complex coefficients are considered, namely $u_1 = [a, 0]^T$, $u_2 = [0, a]^T$, $u_3 = [a, a]^T$, and $u_4 = [a, -a]^T$, where *a* is a sufficiently small scalar to satisfy *H(u)* > 0, *H* is the point spread function (PSF). Eq. 6.5 defines the F_x vector as the outcome of each pixel position *x*:

$$F_x = [F(u_1, x), F(u_2, x), F(u_3, x), F(u_4, x)] \tag{6.5}$$

$$F_x = [Re\{F_x^c\}, Im\{F_x^c\}]^T \tag{6.6}$$

where Re{.} and Im{.} denote the complex number's real and imaginary parts, respectively, the corresponding 8×M^2 transformation matrix is calculated by observing the signs of the real and imaginary parts of each component in F_x using the quantizer described bellow:

$$q_i(x) = \begin{cases} 1, & if\ f_i(x) \geq 0 \\ 0, & otherwise \end{cases} \tag{6.7}$$

where $f_i(x)$ is the *i*th component of the vector F_x.

Within local regions, histograms of blur-insensitive LPQ features are produced as a texture descriptor. The LPQ descriptor has generated considerable interest in blur-invariant texture recognition. LPQ is indifferent to image blurring, and fuzzy and sharp images are highly effective pattern descriptors.

BSIF was initially proposed for texture categorization. BSIF's goal, influenced by LBP and LPQ, is to automatically learn a predefined set of filters from a limited range of natural images rather than utilizing handcrafted filters like LBP and LPQ. BSIF uses machine learning to generate a statistically meaningful representation of the images, allowing for efficient information encoding. Histograms of pixels characterize the image attributes inside each image block. Every element's value is determined by binarizing the image's response to a linear filter with a zero threshold. Each bit corresponds to a distinct filter, and the length of the bit string dictates the number of filters utilized [49]. Eq. 6.8 states the response s_i of an image patch *X* of size $l \times l$ *t*o a specified filter W_i of the same size:

$$s_i = \sum_{u,v} W_i(u,v)X(u,v) = w_i^T x \tag{6.8}$$

where the index *i* in W_i indicates the i^{th} filter, *w* and *x* are vector notations of W_i and *X*.

The binarized feature b_i is calculated by:

$$b_i = \begin{cases} 1, & if\ s_i \geq 0 \\ 0, & otherwise \end{cases} \tag{6.9}$$

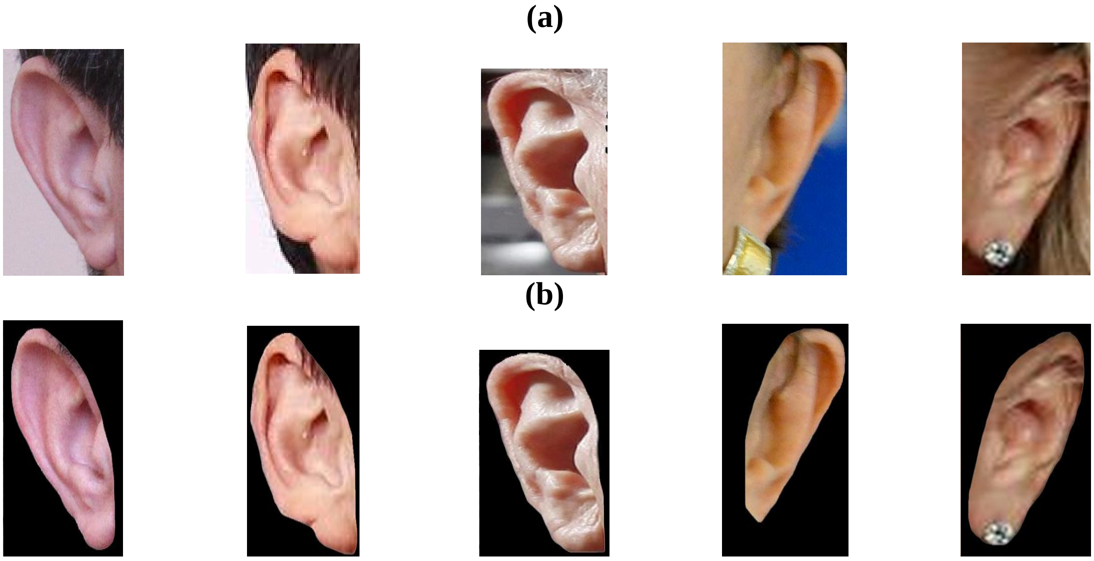
Classification can be accomplished using supervised machine-learning techniques such as K-Nearest Neighbors (K-NN). The K-NN method assumes that similar objects exist near one another. The distance between the feature vectors may quantify the proximity relationship in this case study.

6.5 Experimental analysis

The experimental technique and performance metrics used to evaluate our suggested approach are described in this section before we move on to experiments.

6.5.1 Ear dataset

We used photo-editing software to train the Pix2Pix GAN to generate synthetic segmentation from input ear photos to construct a new version of the AWE ear dataset called RoI-AWE. This dataset comprises only the RoI of each ear image. The RoI is a precise segmentation of the ear pixels without a backdrop and with the fewest possible occlusions; it serves as the ground truth for segmentation. Figure 6.4 illustrates a few photos from the AWE collection with their RoI-AWE counterparts.


(a)

(b)

Figure 6.4 Sample images from the (a) AWE dataset, (b) RoI-AWE dataset.

Table 6.1 Summary of the configurations used by each descriptor.

Descriptor	Configuration
LBP	Uniform LBP, radius: 2 pixels, neighborhood size: 8, block size: 8×8 pixels.
LPQ	Window size: 5×5 pixels, block size: 25×25 pixels.
BSIF	Bit string length: 15, filter size: 12×12, block size: 50×50 pixels.

We conducted a basic ear recognition experiment applying local feature descriptors to compare the original AWE photos to the synthesized RoI-AWE. The configurations utilized by each descriptor are summarized in Table 6.1. The rank-1 recognition rate was utilized to evaluate the categorization procedure at this stage.

6.5.2 Experimental protocol

We choose to evaluate the first performance parameter for the segmentation synthesis process which is the pixel-wise color accuracy. Due to the generative model's function of generating ear segmentations, the colors may deviate somewhat from the original segmentation, even if the difference is imperceptible to the human sight. The image color space was transformed from RGB to CIELab (CIEL*a*b*). Measuring the difference in this color space produces more precise findings than calculating the difference in RGB color space, and it is also more representative of the difference perceived by the human eye.

The pixel-level accuracy is defined as the ratio of correctly colored pixels to the total number of pixels. A pixel is considered "properly colored" if the difference between its colors and the original colors is smaller than a predetermined threshold c for each color channel. The formula theoretically states it in Eq. 6.10:

$$\mathrm{Accuracy}(x, y) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\prod_{c=1}^{3} 1_{[0,\epsilon_c]}\left(\left|x_{i,j}^{c} - y_{i,j}^{c}\right|\right)\right) \tag{6.10}$$

where x is the generated RoI image, y is the corresponding ground truth image, and $1_{[0,\epsilon_c]}$ is the indicator function.

The IoU metric was chosen as the second metric; it is one of the most often used performance metrics for segmentation. It is a relatively simple metric; it is the overlap between the predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth, as defined in Eq. 6.11:

$$\mathrm{IoU} = \frac{\mathrm{SEG} \cap \mathrm{GT}}{\mathrm{SEG} \cup \mathrm{GT}} \tag{6.11}$$

where SEG denotes the generated ear segmentation pixels, whereas GT denotes the ground-truth ear segmentation pixels.

On the other hand, we evaluated the classification model's performance using original AWE dataset photos and RoI-AWE synthetic images via the Rank-1 recognition rate.

6.5.3 Results and discussion

We trained the Pix2Pix GAN for 32000 iterations using the RoI-AWE dataset; we utilized the Adam optimizer with an initial learning rate of 2×10^{-4} and a Mean Absolute Error (MAE) loss function. The convergence of the loss for both the generator and discriminator during training epochs is depicted in Figure 6.5.

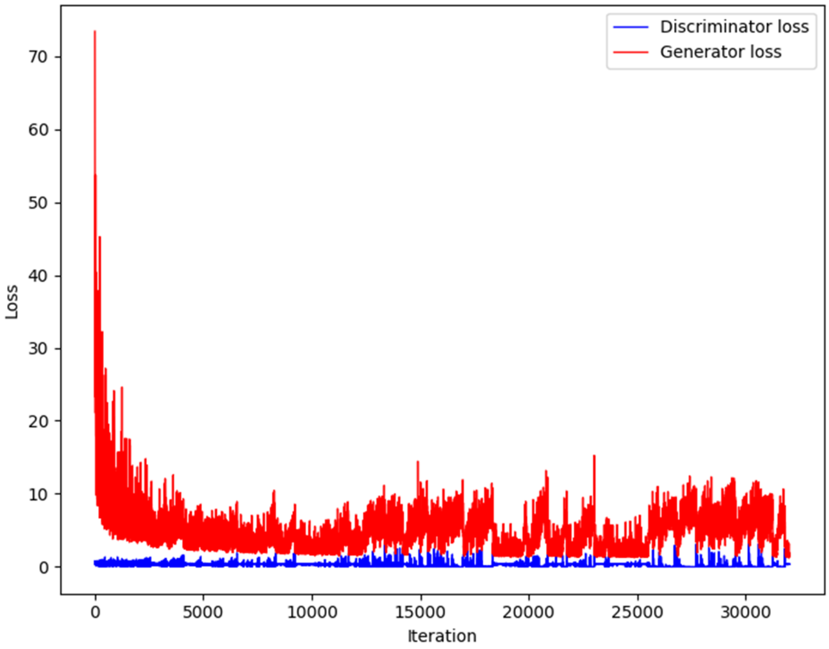
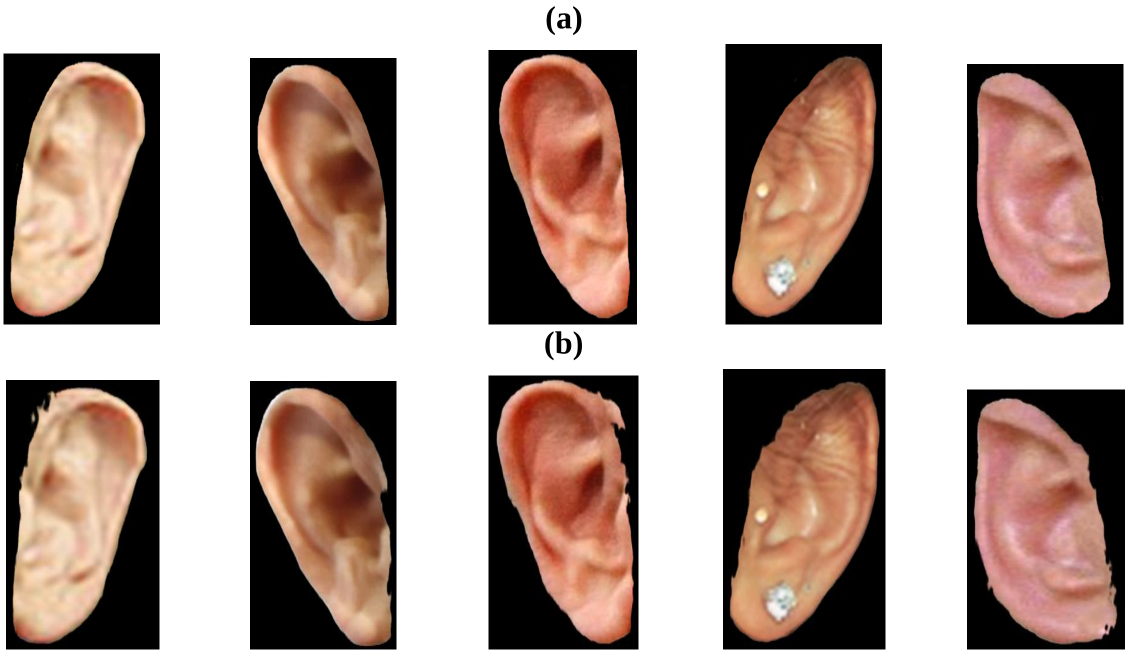


Figure 6.5 Generator/Discriminator loss over training iterations.

Figure 6.6 illustrates examples of RoI segmentation images generated by the GAN. The edges exhibit considerable deformation. We believe that training the generative model on a combination of constrained and unconstrained ear datasets improves performance since the model can recognize a greater variety of ear shapes with varying poses, variations, and scales. However, this needs the creation of a sizable ground-truth RoI dataset.


(a)

(b)

Figure 6.6 Example images of the RoI segmentation synthesis: (a) Ground truth image, (b) Generated RoI.

As shown in Table 6.2, the IoU values of the produced RoI segmentations ranged from 86.24% to 97.21%, with an average of 94.98%.

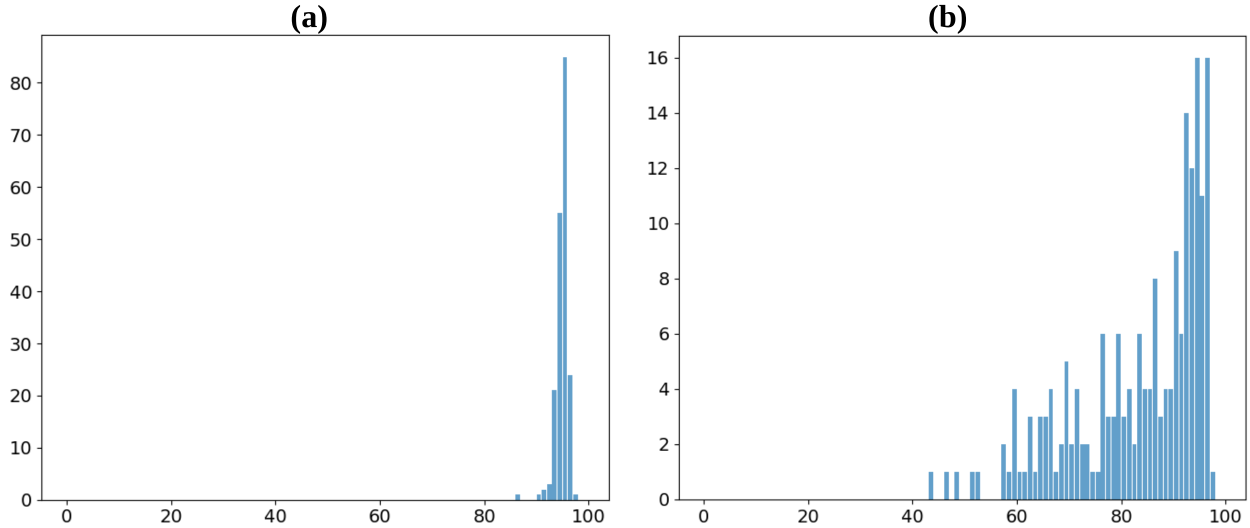


Figure 6.7 Histograms of: (a) IoU, (b) Pixel-wise accuracy.

However, most results are centered about 94%, as illustrated in Figure 6.7's histogram of the IoU distribution. On the other hand, the distribution of pixel-level accuracy was more significant. It ranges between 43.51% and 97.08%, with an average of 83.81%.

Table 6.2 Resultant performance metrics of the RoI segmentation synthesis.

Metric	Minimum	Maximum	Average
IoU	0.8624	0.9721	0.9498
Pixel-wise Accuracy	0.4351	0.9708	0.8331

The second half of the trials is dedicated to featuring extraction and classification, in which we extracted ear features using LBP, LPQ, and BSIF descriptors and fed the histograms to a K-NN classifier. The same experiments are repeated with the same set of train/test images using both the RoI AWE and the original AWE datasets to demonstrate the benefits of RoI segmentation. Synthetic ear RoI pictures achieved a higher rank-1 identification rate in all three cases. As indicated in Table 6.3, the classification of LBP, LPQ, and BSIF characteristics rose by 7,55%, 7,44%, and 3,95%, accordingly.

Table 6.3 Recognition rate of the AWE and the RoI-AWE synthetic datasets.

Method	**AWE Rank-1 (%)**	**RoI AWE Rank-1 (%)**
LBP	19.69	27.24
LPQ	37.25	44.69
BSIF	44.53	48.48

6.6 Conclusion

In this chapter, we attempted to bridge the gap in the literature about ear RoI synthetic segmentation by proposing to use Image-to-Image translation to synthesize ear segmentation and patch absent sections, and eliminate occlusions as much as possible. We investigated the suggested technique on the unconstrained AWE dataset and discovered that it could produce good ear segmentations based on the resulting performance measures.

We used synthetic ear segmentations with local feature extraction and classification approaches such as LBP, LPQ, and BSIF. The recognition rate obtained demonstrated the effectiveness of the method we proposed and its importance in the ear recognition pipeline. Additional research with a larger ear dataset is necessary to train the generative model and even generate ear RoI segmentations with uniform rotations and sizes.

CHAPTER 7: Ear Recognition Based on Deep Unsupervised Active Learning

7.1 Introduction

Traditionally, ear recognition has been accomplished using a machine-learning process, which entails training the model on a subset of labeled data, testing it, and deploying it in the actual world. This strategy has been highly effective. However, one may inquire whether our model can acquire additional features when predicting new data. Cooperative Machine Learning (CML) [147] has been widely used to aid decision-making, data annotation, and other tasks. The basic concept is to train a model on partially labeled data and then predict new labels for new data. Then, using the updated predicted and corrected labels, a human agent or corrector revises the low confidence forecasted data and retrains the model. This concept has been used for various other tasks, including speeding up the annotation of social signals [148], dynamic decision-making [149], and so on.

Along with model prediction, CML is always dependent on human intervention and correction, which means that we must correlate the model with an observing human agent that can monitor and correct model behavior. From here, a critical question arises: What if our model is accurate enough that it can be trusted to gain new knowledge on its own (with a tiny margin of error) throughout the testing phase, without the assistance of a human agent?

This study recommends that active unsupervised learning be used during the test phase of a trained ear recognition model. The classification model aims to forecast and classify test subjects' labels. Meanwhile, the unsupervised active learning stage adds certain test images with their predicted labels to the training dataset and performs additional training epochs (if the predictive confidence is more significant than a predetermined threshold). The proposed training method is called Deep Unsupervised Active Learning (DUAL).

The rest of the chapter is organized as follows: Second section introduces the proposed training workflow. Experimental analysis is carried out in the third section. Finally, Section four concludes the chapter.

7.2 Deep Unsupervised Active Learning workflow

The proposed DUAL training scheme comprises three sequential phases: supervised training, validation and hyper-parameter customization, and unsupervised active learning. We trained the supervised classification model using a labeled training dataset in the first phase. Then, using a small validation set, we conducted a validation experiment to discover the optimal value of the hyper-parameter. Finally, we performed unsupervised active learning on the test images during the test phase. As a result, the unsupervised active learning phase is independent of the original training dataset, i.e., when deploying a biometric model, it should be trained exclusively utilizing the initially labeled dataset. Then, an unsupervised active learning phase is done using real-time test images.

During the typical testing phase, the classification model cannot acquire additional knowledge from the test images (i.e., the model's recognition rate will not improve), even if the model has a high recognition rate. As a result, we present an alternative testing technique in which a model can acquire more knowledge through unsupervised active learning while classifying test images. This testing step is referred to as the unsupervised active learning testing phase.

During the test phase, images categorized with a confidence level greater than the threshold are included in the initial training dataset before completing additional training epochs. As illustrated in Figure 7.1, we proceed through the test subjects one by one.

We used the VGG16 architecture as the basis for our classification model. The purpose of this effort is not to improve the categorization model itself. As a result, we excluded various CNN-based designs such as VGG19, ResNet, and others. It is sufficient to use a classification model architecture with a high recognition rate to validate the suggested method. On top of the convolutional layers of the VGG16 model, which is pre-trained on the ImageNet dataset [61], we added a fully connected layer and a softmax output layer.

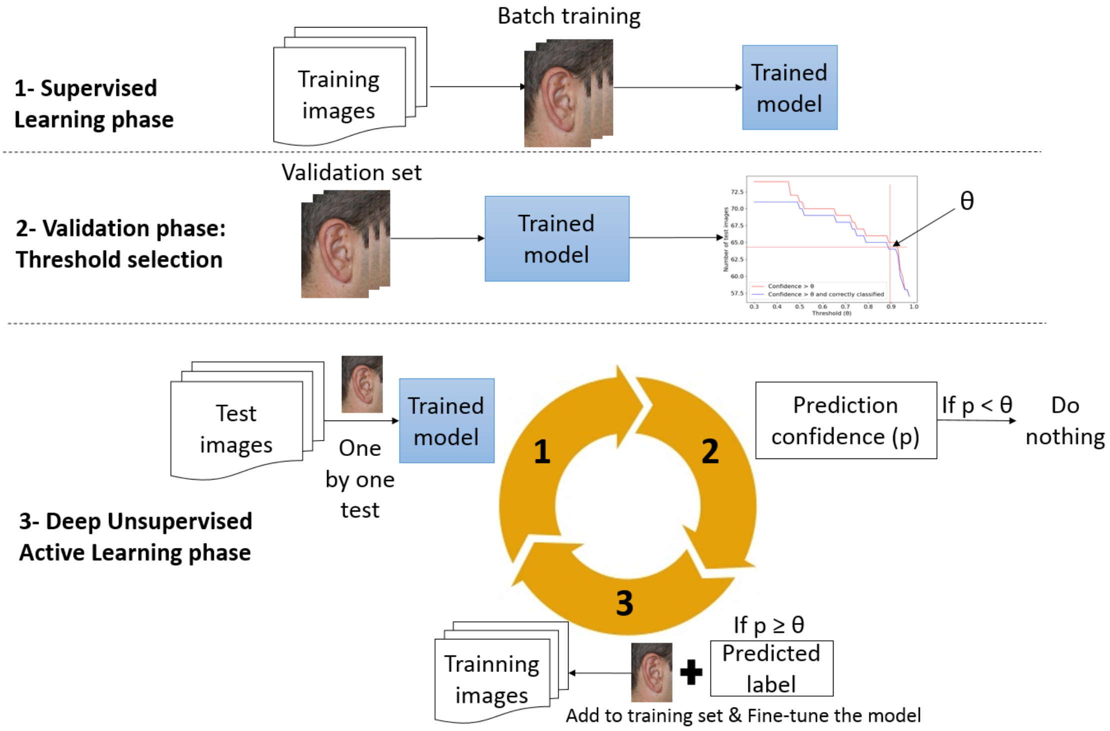


Figure 7.1 Our proposed Deep Unsupervised Active Learning scheme.

The global structure of the classification model is detailed in Table 7.1.

Table 7.1 Details of the parameters of the classification model architecture.

Layer	Neurons number	Activation function	Drop
<VGG16 convolutional layers>			
Dense layer	1024	ReLu	50.00%
Output layer	nbr_of_persons	Softmax	-

We measured our model's performance using categorical cross-entropy as a loss function throughout the training phase. The cross-entropy can be calculated using Eq. (7.1), which is the most straightforward and most frequently used cost function due to its direct relationship to the concept of entropy. On the other hand, we updated the model weights depending on the training data using Adam's well-known optimizer [150].

$$CrossEntropy(p, y) = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \qquad (7.1)$$

where *M* denotes the number of classes in each dataset, *y* is the binary indicator vector if label *c* is the correct classification for image *o*, and *p* is the vector of the estimated probability that image *o* belongs to class *c*.

7.3 Experimental analysis

7.3.1 Experimental Databases

We conducted a series of tests employing ear pictures from the USTB2, AMI, and AWE ear datasets to assess our framework's performance. Figure 7.2 illustrates representative photos from the datasets analyzed.

The university of science and technology of Beijing (USTB) [58] obtained four pictures of 77 subjects' ears under various lighting conditions (students and teachers). The whole collection contains 308 photos that are not cropped. The first image shows the frontal view of the ear under standard illumination; the second and third images show the ear rotated by +30 and -30 degrees, correspondingly; and the fourth image shows the ear under poor illumination.

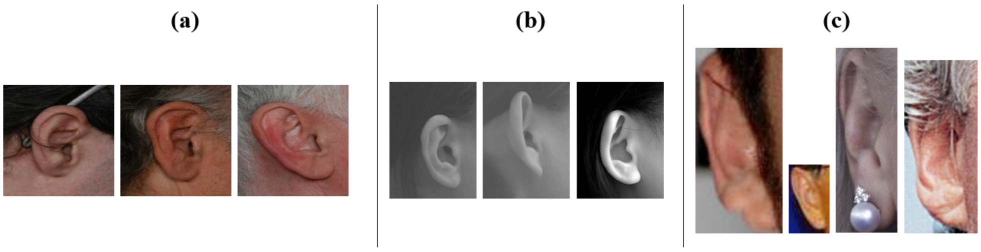


Figure 7.2 Sample ear images from (a) the AMI dataset, (b) the USTB2 dataset, and (c) the AWE dataset.

7.3.2 Setup

Prior to validating the proposed learning strategy, we colored the USTB2 images to increase the recognition rate, defined as the total number of properly-identified images divided by the total number of probe images, as shown in Eq. (7.2).

$$\text{Recognition rate} = \frac{\text{Number of correctly identified probe images}}{\text{Total Number of probe images}} \tag{7.2}$$

Using a cDCGAN model, we created a new colorized dataset named C-USTB2. Because we utilized a VGG16-based model pre-trained on ImageNet using color photos, this colorization step significantly enhances identification rates compared to grayscale images.

To train the cDCGAN colorization model, we used colored photos from the AMI dataset; but, we could use any other ear dataset with colored images; the labels are irrelevant, as we are interested in the colors, not the color labels. For each colored image, the model implicitly generates a corresponding grayscale image and then tries to generate a colorized image. To quantify its performance, we calculated the accuracy, defined as the ratio of correctly colored pixels to total pixels. If the difference between a pixel's RGB of colored image and the original image is less than a particular threshold, the pixel is adequately colored. More precisely, the equation (7.3) [135] defines correctness mathematically as follows:

$$Accuracy(x, y) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\prod_{c=1}^{3} 1_{[0,\epsilon_c]}\left(\left|x_{i,j}^{c} - y_{i,j}^{c}\right|\right)\right) \tag{7.3}$$

where x is the colorized image, y is the corresponding ground truth image, $1_{[0,\epsilon_c]}$ is the indicator function, n and m are the image dimensions, i and j are the pixel indices of the image, c is the color channel, and ϵ_c is the channel threshold.

As shown in Figure 7.3, we trained the model for a total of 62 epochs (more than 1000 mini-batch repetitions) to obtain a distance loss of 2.41 for the generator and 1.38 for the discriminator, as well as a training colorization accuracy of 79.7%. The obtained colorization accuracy is the best to utilize only the AMI color pictures. While the outcome is entirely satisfactory, we believe it might be enhanced by including a broader set of multi-color ear datasets.

As illustrated in Figure 7.4, we obtained a colorized version of USTB2 and equalized the brightness and intensity of the images using cDCGAN.

We extended the training set by creating two more images for each training image, as illustrated in Figure 7.5. One image has been rotated 20° to the left, while the other has been rotated 20° to the right. In the case of AMI, we used 60% of the photos for training and 40%

for testing. We used three images for training and one image for testing in C-USTB2. The AWE datasets' owners have established the train/test distribution, using 600 images for training and 400 images for testing.

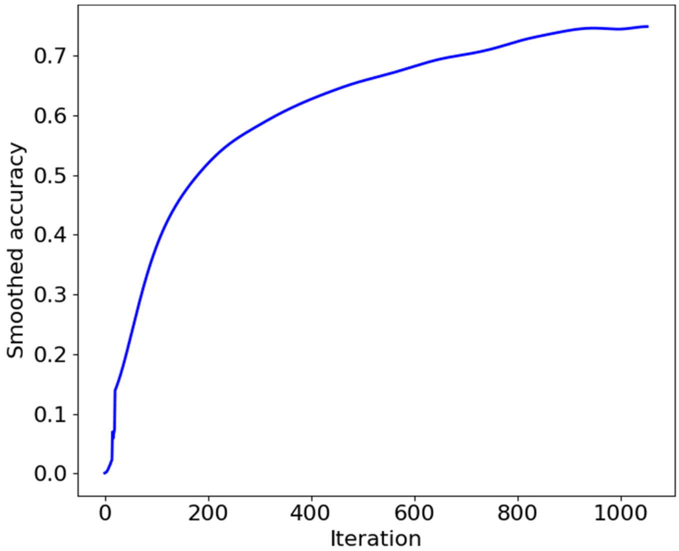


Figure 7.3 Accuracy of the training colorization of the cDCGAN model.

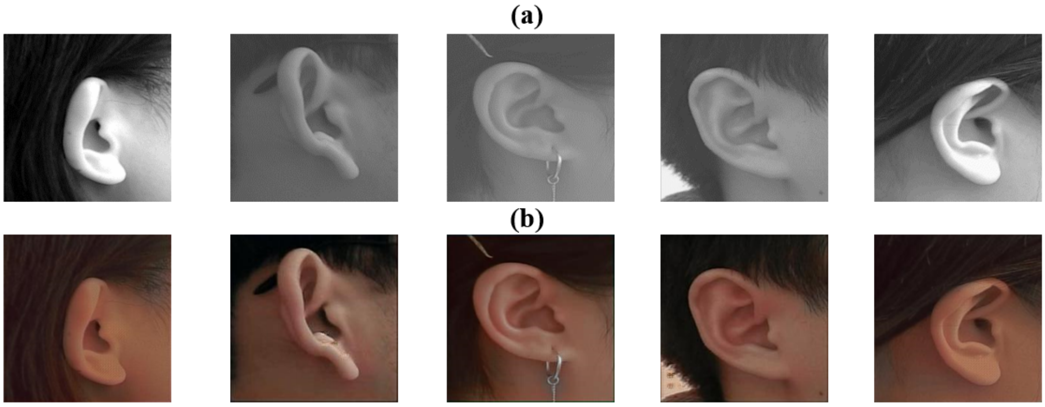
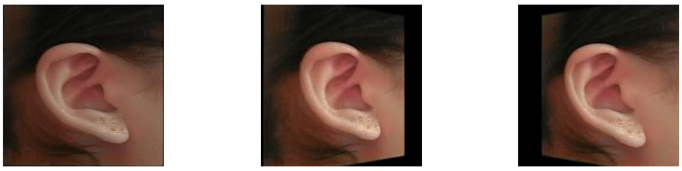


Figure 7.4 Coloring of USTB2 images with cDCGAN (C-USTB2), (a) original images of USTB2, (b) colored images of C-USTB2.



Original image Left rotation by 20° Right rotation by 20°

Figure 7.5 Increase in the training dataset.

7.3.3 Experiment #1

To demonstrate the beneficial effect of coloring the USTB2 dataset on recognition rate, we used the identical configuration of the VGG-based pre-trained classification model with the USTB2 and C-USTB2 datasets. We obtained a recognition rate of 98.70% using the C-USTB2 dataset and 97.40% using the USTB2 grayscale dataset. The CMC curves for both circumstances are depicted in Figure 7.6. As expected, coloring the USTB2 dataset increased recognition rates; this could be attributed to a pre-trained VGG model on color images (ImageNet). Hence it is recommended to employ color images in the following steps.

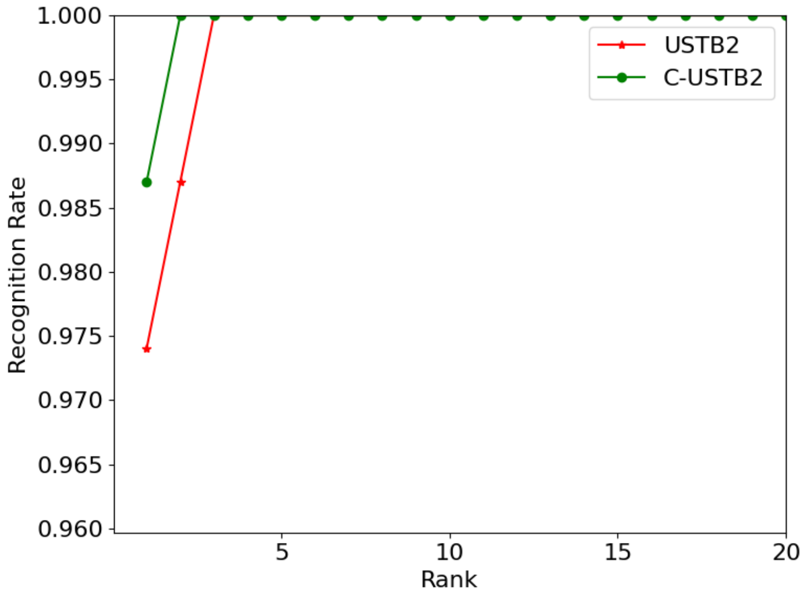
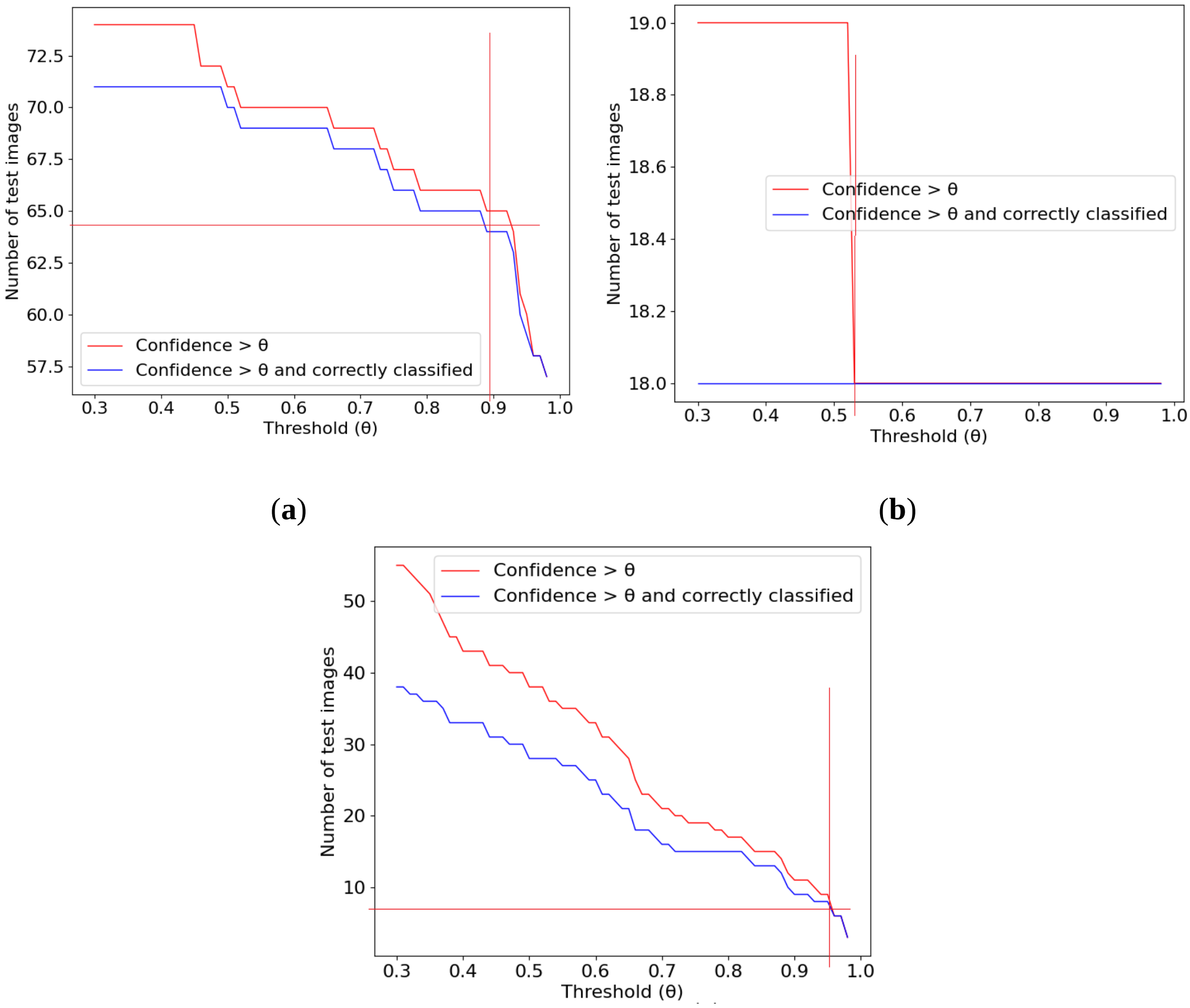


Figure 7.6 Cumulative matching characteristic curves for USTB2 and C-USTB2 datasets.

7.3.4 Experiment #2

We investigated and analyzed the DUAL scheme's effectiveness and effects in this experiment. θ is a confidence level that is established to initiate the fine-tuning process. The optimal value is determined by conducting a single supervised learning test phase using a validation set and observing the number of correctly identified images with a confidence value more significant than a defined θ. We used a quarter of the test set from each dataset as a validation set. Visualizing the facts provides a more distinct perspective from which to decide. The relationship between θ values and the number of correctly recognized test subjects with greater or equal confidence than θ is illustrated in Figure 7.7. We used the best guess to determine the location with a minor vertical difference between the two curves, keeping in mind that our goal is to maximize the number of photos with a confidence level larger than θ.



(a)

(b)

(c)

Figure 7.7 The number of subjects correctly identified with a given θ for (a) the AMI dataset, (b) the C-USTB2 dataset, and (c) the AWE dataset.

The optimum value for the variable is estimated differently depending on several factors, including the type of dataset utilized, whether limited or not, the type of classification model employed, and the variable's introductory recognition rate. In our experiment, we used threshold values of 0.89, 0.52, and 0.95 for the AMI, C-USTB2, and AWE datasets, accordingly.

Table 7.2 Results of the Supervised Learning and the DUAL scheme for AMI, C-USTB2, and AWE datasets.

Method	**AMI**		**C-USTB2**		**AWE**	
	Rank-1	**Rank-5**	**Rank-1**	**Rank-5**	**Rank-1**	**Rank-5**
Supervised Learning	96.00%	99.66%	98.70%	100.00%	49.25%	77.00%
Deep Unsupervised Active Learning	98.33%	100.00%	100.00%	100.00%	51.25%	78.75%

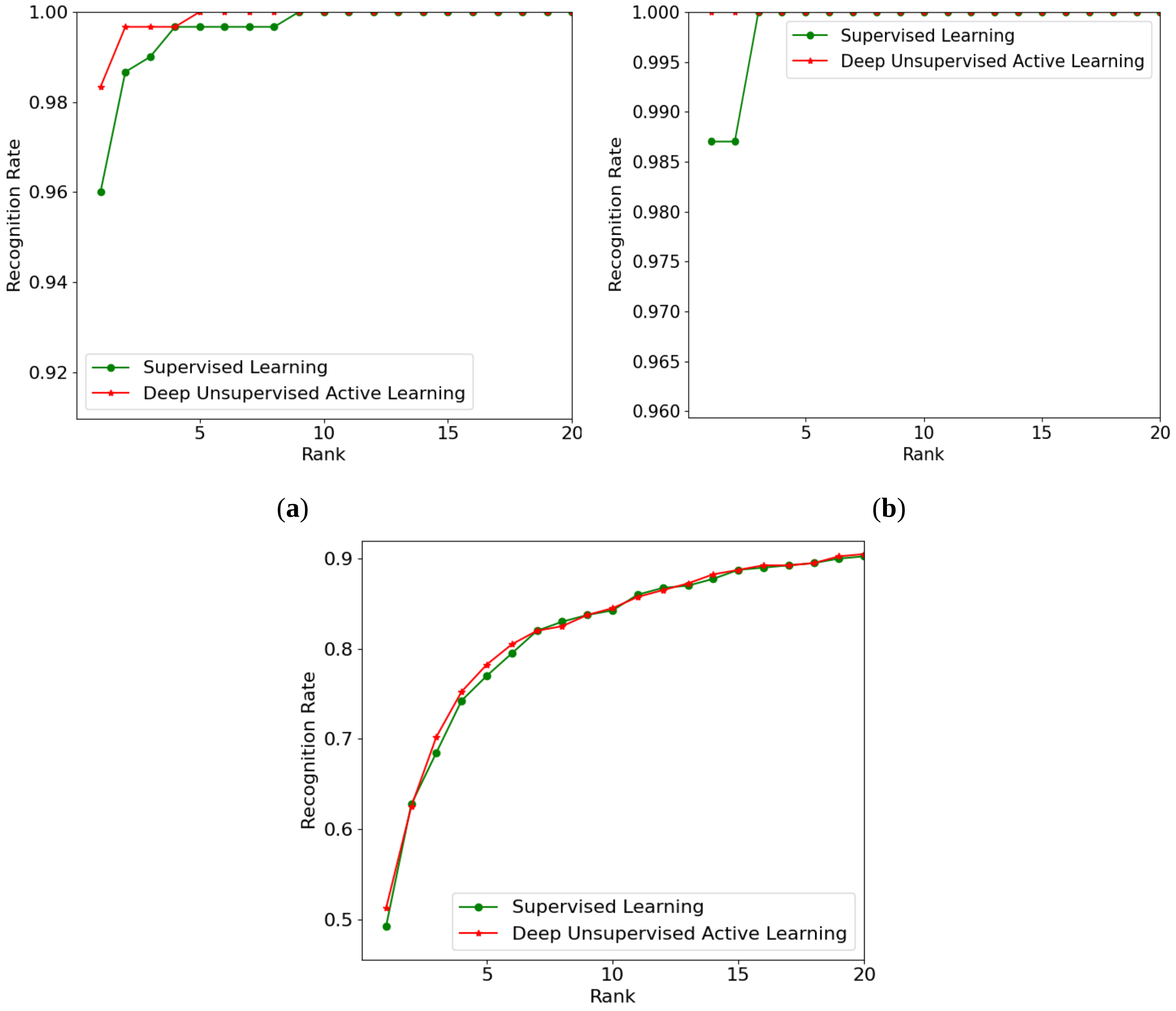
Table 7.2 demonstrates conclusively that the suggested DUAL technique considerably increased the model's recognition rate across all datasets; DUAL reports a greater recognition rate than supervised learning. The recognition rate increased from 96.00% to 98.33% for the AMI dataset. The DUAL method boosted the recognition rate to 100.00% for the C-USTB2 dataset. Additionally, the AWE dataset's recognition rate increased by 2% points.

The number of successfully identified test subjects from all test subjects with confidence θ is shown in Table 7.3. (i.e., test subjects which the model attempts to learn from during the test phase). The recognition rate is improved by executing a new fine-tuning epoch on those test subjects with their projected values. For the AMI dataset, the model was re-trained using new data: 131 newly labeled images. In the second case, the model predicted correct labels for 57 photos with a confidence level greater than θ using the C-USTB2 dataset. As a result, it re-trained itself using accurate data throughout the test phase. Using the preset threshold, the DUAL method actively trained the model on the AWE dataset with 19 new images, 17 of which were identified correctly. Nonetheless, even though 10.52% of new data were misclassified, the recognition rate was increased because of the increased amount of correctly classified data used for active learning.

The CMC curves for the supervised and DUAL learning schemes are depicted in Figure 7.8.

Table 7.3 Supervised learning test statistics.

Test images	**AMI**	**C-USTB2**	**AWE**
Number of test images with confidence $\geq \theta$	132	57	19
Number of images with confidence $\geq \theta$ and Correctly classified	131	57	17



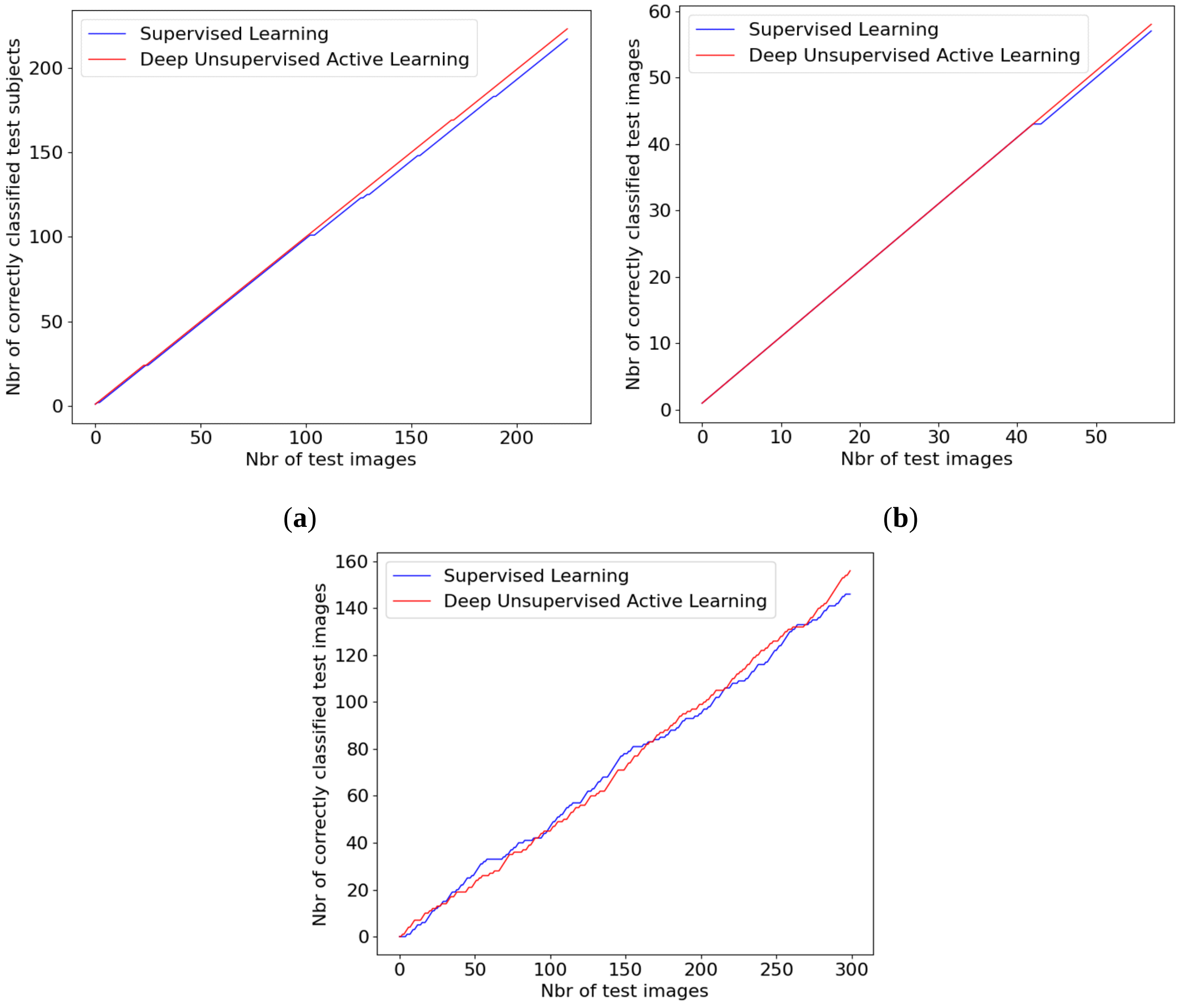
(a)

(b)

(c)

Figure 7.8 Cumulative matching characteristic curves for (a) the AMI dataset, (b) the C-USTB2 dataset, and (c) the AWE dataset.

During the DUAL test phase, we compared the number of correctly recognized photos to the number of correctly recognized images during the supervised learning test phase. As illustrated in Figure 7.9, the model was enhanced to gain additional knowledge by utilizing specific test images, and they are associated with accurately predicted labels anticipated by the DUAL method. This procedure improved the likelihood of recognizing the remaining test images. Retraining the model using test images enhanced the likelihood of adequately predicting the remaining test images in the AMI dataset. Similarly, the DUAL scheme improved the overall identification rate for the remaining test images in the C-USTB2 and AWE situations, respectively, as of test images #42 and #230.



(a) (b)

(c)

Figure 7.9 Number of correctly predicted labels during the test phase for (a) the AMI database, (b) the C-USTB2 database, and (c) the AWE dataset.

7.3.5 Comparison of Rank-1 Recognition Rate

The Rank-1 recognition rate is compared in Table 7.4 between the proposed training strategy and well-known previous approaches that utilized the AMI, USTB2, or AWE datasets. As seen in Table 13, the DUAL approach outperformed state-of-the-art methods by straining the model's performance to its limits. While each approach has several advantages and disadvantages, we concentrated on one significant advantage of the proposed approach over the others in this work, namely the possibility of acquiring new knowledge during the testing phase rather than relying exclusively on what was learned during the learning phase. This, in our opinion, is a critical feature for artificial intelligence systems in general to gain.

While it is essential to keep in mind that the suggested DUAL scheme demands additional processing time and memory space, implementing continuous active learning during the testing phase may be challenging, particularly in real-time.

Table 7.4 A comparison of Rank-1 of the proposed approach with other representative methods.

Publication	Year	Method	USTB2	AMI	AWE
Geometric Methods					
Mu et al. [47]	2004	Geometrical Measures on Edge Images	85.00	-	-
Lakshmanan [57]	2013	Multi-Level Fusion	99.20	-	-
Holistic Methods					
Zhang and Mu [83]	2008	PCA	81.80	-	-
		ICA	92.20	-	-
Gutierrez et al. [41]	2010	Wavelet Transform	97.50	-	-
Tariq et al. [84]	2011	Haar Wavelets + Fast Normalized Cross Correlation	96.10	-	-
Local Methods					
Guo and Xu [54]	2008	LBP + Cellular NN	93.30	-	-
Ghoualmi et al. [59]	2016	SIFT	94.79	-	-
Emeršič et al. [151]	2017	BSIF	-	-	48.40
		POEM	-	-	49.60
Chowdhury et al. [53]	2018	Tunable Filter Bank	-	70.14	-
Hassaballah et al. [60]	2019	Completed LBP	-	73.71	49.60
Deep learning methods					
Omara et al. [69]	2018	Pairwise SVM	99.00	-	-
Alshazly et al. [67]	2019	VGG-13-16-19 ensemble	-	97.50	-
Zhang et al. [152]	2018	VGG-face	-	-	50.00
Alshazly et al. [153]	2019	AlexNet (Fine Tuning)	-	94.50	-
Zhang et al. [80]	2019	MAML + CNN	-	93.96	-
Khaldi and Benzaoui [72]	2020	DCGAN + VGG16	-	96.00	50.53
Priyadharshini et al. [71]	2020	CNN	-	96.99	-
Proposed method	**2021**	**VGG16 + DUAL**	**100.00**	**98.33**	**51.25**

7.4 Conclusion

The purpose of this study is to establish the feasibility of active learning in the field of ear identification and, more broadly, biometrics. We introduced a machine learning technique dubbed *Deep Unsupervised Active Learning (DUAL)* for continually updating a biometric model's knowledge after the training phase. A biometric model then performs additional learning epochs using the test images that have been categorized with a confidence value more significant than a predetermined threshold. We next validated this fact by conducting in-depth tests on the recognition rate under supervised and DUAL learning using the limited AMI and C-USTB2 ear datasets and the unconstrained AWE dataset. The recognition rates for Rank 1 are 100.00% and 98.33%, respectively, for the C-USTB2 and AMI datasets and 51.25% for the challenging AWE dataset.

These preliminary findings lead to two conclusions:

- It is critical to appropriately identify the dataset used when presenting the performance of a method.
- The fact that the AWE dataset is noisy leads to relatively poor performance. As a result, picture pre-processing techniques such as de-noising background noise and texture data using first and second-generation wavelets, as well as multi-resolution analysis, are required.

Although the obtained results are generally adequate, prospective enhancements include pre-processing, better parameter adjustment, and the use of variable thresholds.

Conclusions and future work

Biometrics refers to physical or behavioral features of humans that can be utilized for a variety of purposes, ranging from recognizing human action to identifying and verifying individuals. Biometric accuracy has been a subject of research for decades. Scientists are attempting to enhance the performance of biometric models by integrating as many obstacles and challenges as feasible, such as a pose, occlusions, expressions, and backgrounds. The ear is one physical biometric modality that is concerned with this issue. It has a distinctive structure that can be used to identify individuals. The ear, unlike the face, is unaffected by aging or expression. Ear image acquisition does not require expensive equipment, as fingerprint or iris image acquisition does. As a result, we sought to evaluate our approach using ear datasets.

We were particularly interested in offering novel ideas and schemes for enhancing auditory recognition tasks in this thesis. While the proposed solutions outperformed state-of-the-art methods, several complicated difficulties require additional development.

We have concentrated on the following major issues in this thesis:

1) During the testing phase, the loss of color information, i.e., providing grayscale, monochrome, or dark test images to a model trained on colored images. Because no previous study has examined the prospect of colorizing grayscale ear pictures to improve recognition rates, there is a general shortage of research on this subject. As a result, we want to contribute to the emerging field of ear biometrics by developing a new framework capable of recognizing grayscale ear photos with nearly the same efficiency as color photographs. The proposed system colorizes grayscale test images before feeding them to a trained classification model using Image-to-Image translation. On the other hand, we conducted an additional experiment to demonstrate that providing grayscale photos for the training process is insufficient when attempting to identify predicted grayscale test images.
2) We attempted to bridge the gap in the literature about ear Region-of-Interest synthetic segmentation by proposing to use Image-to-Image translation not only to synthesize ear segmentation but also to patch missing sections and eliminate occlusions as much as feasible. We demonstrated that non-ear pixels, such as backdrops, hair, a portion of the face or neck skin, and even clothing, might adversely affect and degrade the categorization result. Thus, the suggested technique circumvented the issue by synthesizing an ear segmentation composed entirely of ear-related pixels.

3) We investigated and researched the Deep Unsupervised Active Learning technique for machine learning. A classification model incrementally gains new information throughout the test phase using the suggested training strategy without any operator direction or decision-making correction. Thus, a biometric model with a high initial identification rate can be continuously retrained using anticipated labels via test images.

Although this work is extensible to many types of research and future research could, therefore, concentrate on:

- Conduct additional research to generalize the Deep Unsupervised Active Learning scheme utilizing datasets from other areas of image classification, and study the feasibility of applying changeable thresholds in response to changes in the recognition curve.
- Additional research should be conducted on a larger ear dataset to train the generative model and even generate ear RoI segmentations consistent to rotations and sizes.
- For some complex datasets, such as the AWE, it is necessary to investigate additional preprocessing techniques, such as de-noising background noise and texture data using first and second-generation wavelets and multi-resolution analysis.
- Optimization of the DUAL scheme by identifying more efficient and faster techniques for the model to learn information from test images less quickly and without compromising or altering previously gained knowledge.

References

[1] Bulmer, M. (**2003**). Francis Galton: pioneer of heredity and biometry. JHU Press.

[2] Kalyani, C. H. (**2017**). Various biometric authentication techniques: a review. Journal of Biometrics & Biostatistics, 8(05).

[3] Dargan, S., & Kumar, M. (**2020**). A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. Expert Systems with Applications, 143, 113114.

[4] Harakannanavar, S. S., Renukamurthy, P. C., & Raja, K. B. (**2019**). Comprehensive study of biometric authentication systems, challenges and future trends. International Journal of Advanced Networking and Applications, 10(4), 3958-3968.

[5] Saini, R., & Rana, N. (**2014**). Comparison of various biometric methods. International Journal of Advances in Science and Technology, 2(1), 24-30.

[6] Zapata, J. C., Duque, C. M., Rojas-Idarraga, Y., Gonzalez, M. E., Guzmán, J. A., & Botero, M. B. (**2017**, September). Data fusion applied to biometric identification–a review. In Colombian Conference on Computing (pp. 721-733). Springer, Cham.

[7] de Luis-Garcı́a, R., Alberola-Lopez, C., Aghzout, O., & Ruiz-Alzola, J. (**2003**). Biometric identification systems. Signal processing, 83(12), 2539-2557.

[8] Addy, D., & Bala, P. (**2016**, September). Physical access control based on biometrics and GSM. In 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1995-2001). IEEE.

[9] Boulgouris, N. V., Plataniotis, K. N., & Micheli-Tzanakou, E. (**2009**). Biometrics: theory, methods, and applications (Vol. 9). John Wiley & Sons.

[10] Okoh, E., & Awad, A. I. (**2015**, May). Biometrics applications in e-health security: A preliminary survey. In International Conference on Health Information Science (pp. 92-103). Springer, Cham.

[11] Giot, R., El-Abed, M., & Rosenberger, C. (**2013**). Fast computation of the performance evaluation of biometric systems: Application to multibiometrics. Future Generation Computer Systems, 29(3), 788-799.

[12] Ahmed, A. A. E., & Traoré, I. (**2012**). Performance metrics and models for continuous authentication systems. In Continuous Authentication Using Biometrics: Data, Models, and Metrics (pp. 23-39). IGI Global.

[13] Mingote, V., Miguel, A., Ribas, D., Giménez, A. O., & Lleida, E. (**2019**). Optimization of False Acceptance/Rejection Rates and Decision Threshold for End-to-End Text-Dependent Speaker Verification Systems. In INTERSPEECH (pp. 2903-2907).

[14] Hajian-Tilaki, K. (**2013**). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian journal of internal medicine, 4(2), 627.

[15] El-Hameed, H. A. A., Ramadan, N., El-Shafai, W., Khalaf, A. A., Ahmed, H. E. H., Elkhamy, S. E., & El-Samie, F. E. A. (**2021**). Cancelable biometric security system based on advanced chaotic maps. The Visual Computer, 1-17.

[16] Yager, N., & Amin, A. (**2004**). Fingerprint classification: a review. Pattern Analysis and Applications, 7(1), 77-93.

[17] Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (**2003**). Face recognition: A literature survey. ACM computing surveys (CSUR), 35(4), 399-458.

[18] Tripathi, K. P. (**2011**). A comparative study of biometric technologies with reference to human interface. International Journal of Computer Applications, 14(5), 10-15.

[19] Unar, J. A., Seng, W. C., & Abbasi, A. (**2014**). A review of biometric technology along with trends and prospects. Pattern recognition, 47(8), 2673-2688.

[20] Crihalmeanu, S., & Ross, A. (**2012**). Multispectral scleral patterns for ocular biometric recognition. Pattern Recognition Letters, 33(14), 1860-1869.

[21] Hsiao, C. S., & Fan, C. P. (**2021**, June). EfficientNet Based Iris Biometric Recognition Methods with Pupil Positioning by U-Net. In 2021 3rd International Conference on Computer Communication and the Internet (ICCCI) (pp. 1-5). IEEE.

[22] Sliney, D. H. (**1997**). Laser and LED eye hazards: safety standards. Optics and Photonics News, 8(9), 31.

[23] Sadikoglu, F., & Uzelaltinbulat, S. (**2016**). Biometric retina identification based on neural network. Procedia Computer Science, 102, 26-33.

[24] Kemp, R., Palmer, N., Kielmann, T., Seinstra, F., Drost, N., Maassen, J., & Bal, H. (**2009**, December). eyedentify: Multimedia cyber foraging from a smartphone. In 2009 11th IEEE International Symposium on Multimedia (pp. 392-399). IEEE.

[25] Burge, M., & Burger, W. (**1996**). Ear biometrics. In Biometrics (pp. 273-285). Springer, Boston, MA.

[26] Valecha, H., Ahuja, V., Valechha, L., Chawla, T., & Sengupta, S. (**2018**, December). Orisyncrasy-An Ear Biometrics on the Fly Using Machine Learning Techniques. In International conference on Computer Networks, Big data and IoT (pp. 1005-1016). Springer, Cham.

[27] Khaldi, Y., Benzaoui, A., Ouahabi, A., Jacques, S., & Taleb-Ahmed, A. (**2021**). Ear recognition based on deep unsupervised active learning. IEEE Sensors Journal, 21(18), 20704-20713.

[28] Wang, L., Tan, T., Ning, H., & Hu, W. (**2003**). Silhouette analysis-based gait recognition for human identification. IEEE transactions on pattern analysis and machine intelligence, 25(12), 1505-1518.

[29] Wan, C., Wang, L., & Phoha, V. V. (**2018**). A survey on gait recognition. ACM Computing Surveys (CSUR), 51(5), 1-35.

[30] Preis, J., Kessel, M., Werner, M., & Linnhoff-Popien, C. (**2012**, June). Gait recognition with kinect. In 1st international workshop on kinect in pervasive computing (pp. 1-4). New Castle, UK.

[31] Sanchez-Reillo, R., Sanchez-Avila, C., & Gonzalez-Marcos, A. (**2000**). Biometric identification through hand geometry measurements. IEEE Transactions on pattern analysis and machine intelligence, 22(10), 1168-1171.

[32] de-Santos-Sierra, A., Sánchez-Avila, C., Del Pozo, G. B., & Guerra-Casanova, J. (**2011**). Unconstrained and contactless hand geometry biometrics. Sensors, 11(11), 10143-10164.

[33] Guru, D. S., & Prakash, H. N. (**2008**). Online signature verification and recognition: An approach based on symbolic representation. IEEE transactions on pattern analysis and machine intelligence, 31(6), 1059-1073.

[34] Karouni, A., Daya, B., & Bahlak, S. (**2011**). Offline signature recognition using neural networks approach. Procedia Computer Science, 3, 155-161.

[35] Tanwar, S., Obaidat, M. S., Tyagi, S., & Kumar, N. (**2019**). Online signature-based biometric recognition. In Biometric-based physical and cybersecurity systems (pp. 255-285). Springer, Cham.

[36] Jain, A. K., Ross, A. A., & Nandakumar, K. (**2011**). Introduction to biometrics. Springer Science & Business Media.

[37] Benzaoui, A., Adjabi, I., & Boukrouche, A. (**2017**). Experiments and improvements of ear recognition based on local texture descriptors. Optical Engineering, 56(4), 043109.

[38] Arbab-Zavar, B., & Nixon, M. S. (**2011**). On guided model-based analysis for ear biometrics. Computer Vision and Image Understanding, 115(4), 487-502.

[39] Benzaoui, A., Adjabi, I., & Boukrouche, A. (**2016**, October). Person identification based on ear morphology. In 2016 International Conference on Advanced Aspects of Software Engineering (ICAASE) (pp. 1-5). IEEE.

[40] Hurley, D. J., Nixon, M. S., & Carter, J. N. (**2000**, March). Automatic ear recognition by force field transformations. In IEE colloquium on visual biometrics (Ref. No. 2000/018) (pp. 7-1). IET.

[41] Gutierrez, L., Melin, P., & Lopez, M. (**2010**). Modular neural network integrator for human recognition from ear images. In The 2010 international joint conference on neural networks (IJCNN) (pp. 1–5).

[42] Chang, K., Bowyer, K. W., Sarkar, S., & Victor, B. (**2003**). Comparison and combination of ear and face images in appearance-based biometrics. IEEE Transactions on pattern analysis and machine intelligence, 25(9), 1160–1165.

[43] Zhang, H. J., Mu, Z. C., Qu, W., Liu, L. M., & Zhang, C. Y. (**2005**, August). A novel approach for ear recognition based on ICA and RBF network. In IEEE International Conference on Machine Learning and Cybernetics (Vol. 7, pp. 4511-4515).

[44] Xie, Z., & Mu, Z. (**2008**, December). Ear recognition using LLE and IDLLE algorithm. In IEEE 19th International Conference on Pattern Recognition (pp. 1-4).

[45] Hanmandlu, M. (**2013**). Robust ear based authentication using local principal independent components. Expert Systems with Applications, 40(16), 6478-6490.

[46] Moreno, B., Sanchez, A., & Vélez, J. F. (**1999**). On the use of outer ear images for personal identification in security applications. In Proceedings ieee 33rd annual 1999 international carnahan conference on security technology (Cat. No. 99ch36303) (pp. 469–476).

[47] Mu, Z., Yuan, L., Xu, Z., Xi, D., & Qi, S. (**2004**). Shape and structural feature based ear recognition. In Chinese conference on biometric recognition (pp. 663–670).

[48] Choras, M., & Choras, R. S. (**2006**). Geometrical algorithms of ear contour shape representation and feature extraction. In Sixth international conference on intelligent systems design and applications (Vol. 2, pp. 451–456).

[49] Rahman, M., Sadi, M. S., & Islam, M. R. (**2014**). Human ear recognition using geometric features. In 2013 international conference on Electrical Information and Communication Technology (EICT). (pp. 1–4).

[50] Omara, I., Li, F., Zhang, H., & Zuo, W. (**2016**). A novel geometric feature extraction method for ear recognition. Expert Systems with Applications, 65, 127-135.

[51] Benzaoui, A., Hadid, A., & Boukrouche, A. (**2014**). Ear biometric recognition using local texture descriptors. Journal of electronic imaging, 23(5), 053008.

[52] Benzaoui, A., & Boukrouche, A. (**2017**). Ear recognition using local color texture descriptors from one sample image per person. In 2017 4th international conference on control, decision and information technologies (CoDIT). (pp. 0827–0832).

[53] Chowdhury, D. P., Bakshi, S., Guo, G., & Sa, P. K. (**2018**). 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.

[54] Guo, Y., & Xu, Z. (**2008**). Ear recognition using a new local matching approach. In 2008 15th ieee international conference on image processing (pp. 289–292).

[55] Kumar, A., & Chan, T. S. T. (**2013**). Robust ear identification using sparse representation of local texture descriptors. Pattern recognition, 46(1), 73-85.

[56] Al Rahhal, M. M., Mekhalfi, M. L., Guermoui, M., Othman, E., Lei, B., & Mahmood, A. (**2018**). A dense phase descriptor for human ear recognition. IEEE Access, 6, 11883–11887.

[57] Lakshmanan, L. (**2013**). Efficient person authentication based on multi-level fusion of ear scores. IET biometrics, 2(3), 97–106.

[58] The USTB ear dataset. Available online: http://www1.ustb.edu.cn/resb/en/news/news3.htm (accessed on 26/03/2021).

[59] Ghoualmi, L., Draa, A., & Chikhi, S. (**2016**). An ear biometric system based on artificial bees and the scale invariant feature transform. Expert Systems with Applications, 57, 49-61.

[60] Hassaballah, M., Alshazly, H. A., & Ali, A. A. (**2019**). Ear recognition using local binary patterns: A comparative experimental study. Expert Systems with Applications, 118, 182-200.

[61] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (**2012**). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

[62] Simonyan, K., & Zisserman, A. (**2014**). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[63] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Dumitru E., Vincent V., & Rabinovich, A. (**2015**). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[64] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (**2016**). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. arXiv preprint arXiv:1602.07360.

[65] Adjabi, I., Ouahabi, A., Benzaoui, A., & Taleb-Ahmed, A. (**2020**). Past, present, and future of face recognition: a review. Electronics, 9(8), 1188.

[66] Tian, L., & Mu, Z. (**2016**). Ear recognition based on deep convolutional network. In 2016 9th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI) (pp. 437–441).

[67] Alshazly, H., Linse, C., Barth, E., & Martinetz, T. (**2019**). Ensembles of deep learning models and transfer learning for ear recognition. Sensors, 19(19), 4139.

[68] Kacar, U., & Kirci, M. (**2018**). ScoreNet: Deep cascade score level fusion for unconstrained ear recognition. IET Biometrics, 8(2), 109-120.

[69] Omara, I., Wu, X., Zhang, H., Du, Y., & Zuo, W. (**2018**). Learning pairwise SVM on hierarchical deep features for ear recognition. IET Biometrics, 7(6), 557-566.

[70] Hansley, E. E., Segundo, M. P., & Sarkar, S. (**2018**). Employing fusion of learned and handcrafted features for unconstrained ear recognition. IET Biometrics, 7(3), 215-223.

[71] Priyadharshini, R. A., Arivazhagan, S., & Arun, M. (**2021**). A deep learning approach for person identification using ear biometrics. Applied Intelligence, 51(4), 2161-2172.

[72] Khaldi, Y., & Benzaoui, A. (**2020**). A new framework for grayscale ear images recognition using generative adversarial networks under unconstrained conditions. Evolving Systems, 1-12.

[73] Khaldi, Y., & Benzaoui, A. (**2020**, November). Region of Interest Synthesis using Image-to-Image Translation for ear recognition. In 2020 International Conference on Advanced Aspects of Software Engineering (ICAASE) (pp. 1-6). IEEE.

[74] Emeršič, Ž., Štepec, D., Štruc, V., & Peer, P. (**2017**). Training convolutional neural networks with limited training data for ear recognition in the wild. In 12th IEEE International Conference on Automatic Face & Gesture Recognition. Washington, DC. (pp. 987-994).

[75] Dodge, S., Mounsef, J., & Karam, L. (**2018**). Unconstrained ear recognition using deep neural networks. IET Biometrics, 7(3), 207–214.

[76] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (**2018**). A survey on deep transfer learning. In International conference on artificial neural networks. Springer, Cham. (pp. 270-279).

[77] Alshazly, H., Linse, C., Barth, E., & Martinetz, T. (**2020**). Deep Convolutional Neural Networks for Unconstrained Ear Recognition. IEEE Access. 8. (pp.170295-170310).

[78] He, K., Zhang, X., Ren, S., & Sun, J. (**2016**). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778).

[79] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (**2017**). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1492-1500).

[80] Zhang, J., Yu, W., Yang, X., & Deng, F. (**2019**, February). Few-shot learning for ear recognition. In Proceedings of the 2019 International Conference on Image, Video and Signal Processing (pp. 50-54).

[81] Emeršič, Ž., Štepec, D., Štruc, V., Peer, P., George, A., Ahmad, A., Omar, E., Boult, T.E., Safdaii, R., Zhou, Y., & Zafeiriou, S. (**2017**). The unconstrained ear recognition challenge. In IEEE international joint conference on biometrics (IJCB). (pp. 715-724).

[82] Emeršič, Ž., SV, A.K., Harish, B.S., Gutfeter, W., Khiarak, J.N., Pacut, A., Hansley, E., Segundo, M.P., Sarkar, S., Park, H.J., & Nam, G.P. (**2019**). The Unconstrained Ear Recognition Challenge 2019. In IEEE International Conference on Biometrics (ICB). (pp. 1-15).

[83] Zhang, H., & Mu, Z. (**2008**, September). Compound structure classifier system for ear recognition. In 2008 IEEE International Conference on Automation and Logistics (pp. 2306-2309). IEEE.

[84] Tariq, A., Anjum, M. A., & Akram, M. U. (**2011**, December). Personal identification using computerized human ear recognition system. In Proceedings of 2011 International Conference on Computer Science and Network Technology (Vol. 1, pp. 50-54). IEEE.

[85] Burge, M., & Burger, W. (**2000**, September). Ear biometrics in computer vision. In Proceedings 15th International Conference on Pattern Recognition. ICPR-2000 (Vol. 2, pp. 822-826). IEEE.

[86] Lakshmanan, L. (**2013**). Efficient person authentication based on multi-level fusion of ear scores. IET biometrics, 2(3), 97-106.

[87] Jordan, M. I., & Mitchell, T. M. (**2015**). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[88] Zhang, X. D. (**2020**). Machine learning. In A Matrix Algebra Approach to Artificial Intelligence (pp. 223-440). Springer, Singapore.

[89] Zhang, Z. (2018). Artificial neural network. In Multivariate time series analysis in climate and environmental research (pp. 1-35). Springer, Cham.

[90] Maind, S. B., & Wankar, P. (**2014**). Research paper on basic of artificial neural network. International Journal on Recent and Innovation Trends in Computing and Communication, 2(1), 96-100.

[91] Sibi, P., Jones, S. A., & Siddarth, P. (**2013**). Analysis of different activation functions using back propagation neural networks. Journal of theoretical and applied information technology, 47(3), 1264-1268.

[92] Sharma, S., Sharma, S., & Athaiya, A. (**2017**). Activation functions in neural networks. towards data science, 6(12), 310-316.

[93] Karlik, B., & Olgac, A. V. (**2011**). Performance analysis of various activation functions in generalized MLP architectures of neural networks. International Journal of Artificial Intelligence and Expert Systems, 1(4), 111-122.

[94] Agarap, A. F. (**2018**). Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375.

[95] Berner, J., Elbrächter, D., & Grohs, P. (**2019**). How degenerate is the parametrization of neural networks with the ReLU activation function?. arXiv preprint arXiv:1905.09803.

[96] Li, J., Cheng, J. H., Shi, J. Y., & Huang, F. (**2012**). Brief introduction of back propagation (BP) neural network algorithm and its improvement. In Advances in computer science and information engineering (pp. 553-558). Springer, Berlin, Heidelberg.

[97] LeCun, Y., Touresky, D., Hinton, G., & Sejnowski, T. (**1988**, June). A theoretical framework for back-propagation. In Proceedings of the 1988 connectionist models summer school (Vol. 1, pp. 21-28).

[98] Kim, P. (**2017**). Convolutional neural network. In MATLAB deep learning (pp. 121-147). Apress, Berkeley, CA.

[99] Albawi, S., Mohammed, T. A., & Al-Zawi, S. (**2017**, August). Understanding of a convolutional neural network. In 2017 International Conference on Engineering and Technology (ICET) (pp. 1-6). Ieee.

[100] Cooper, H., Holt, B., & Bowden, R. (**2011**). Sign language recognition. In Visual analysis of humans (pp. 539-562). Springer, London.

[101] Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., Adam, M., Gertych, A., & San Tan, R. (**2017**). A deep convolutional neural network model to classify heartbeats. Computers in biology and medicine, 89, 389-396.

[102] Li, Y., Hao, Z. B., & Lei, H. (**2016**). Survey of convolutional neural network. Journal of Computer Applications, 36(9), 2508-2515.

[103] O'Shea, K., & Nash, R. (**2015**). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

[104] Wu, J. (**2017**). Introduction to convolutional neural networks. National Key Lab for Novel Software Technology. Nanjing University. China, 5(23), 495.

[105] Lin, X., Zhao, C., & Pan, W. (**2017**). Towards accurate binary convolutional neural network. arXiv preprint arXiv:1711.11294.

[106] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (**2014**). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[107] Hochreiter, S. (**1998**). The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6(02), 107-116.

[108] Lin, M., Chen, Q., & Yan, S. (**2013**). Network in network. arXiv preprint arXiv:1312.4400.

[109] Hsiao, T. Y., Chang, Y. C., Chou, H. H., & Chiu, C. T. (**2019**). Filter-based deep-compression with global average pooling for convolutional networks. Journal of Systems Architecture, 95, 9-18.

[110] Zeiler, M. D., & Fergus, R. (**2014**, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.

[111] Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (**2009**). Visualizing higher-layer features of a deep network. University of Montreal, 1341(3), 1.

[112] Le, Q. V. (**2013**, May). Building high-level features using large scale unsupervised learning. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 8595-8598). IEEE.

[113] Ranzato, M. A., Huang, F. J., Boureau, Y. L., & LeCun, Y. (**2007**, June). Unsupervised learning of invariant feature hierarchies with applications to object recognition. In 2007 IEEE conference on computer vision and pattern recognition (pp. 1-8). IEEE.

[114] Dauphin, Y. N., De Vries, H., & Bengio, Y. (**2015**). Equilibrated adaptive learning rates for non-convex optimization. arXiv preprint arXiv:1502.04390.

[115] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (**2017**, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence.

[116] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (**2016**). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[117] Wu, S., Zhong, S., & Liu, Y. (**2018**). Deep residual learning for image steganalysis. Multimedia tools and applications, 77(9), 10437-10453.

[118] Srivastava, R. K., Greff, K., & Schmidhuber, J. (**2015**). Highway networks. arXiv preprint arXiv:1505.00387.

[119] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (**2017**). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[120] Kuang, P., Ma, T., Chen, Z., & Li, F. (**2019**). Image super-resolution with densely connected convolutional networks. Applied Intelligence, 49(1), 125-136.

[121] Yadav, D. P., Jalal, A. S., Garlapati, D., Hossain, K., Goyal, A., & Pant, G. (**2020**). Deep learning-based ResNeXt model in phycological studies for future. Algal Research, 50, 102018.

[122] Huang, G., Sun, Y., Liu, Z., Sedra, D., & Weinberger, K. Q. (**2016**, October). Deep networks with stochastic depth. In European conference on computer vision (pp. 646-661). Springer, Cham.

[123] Zagoruyko, S., & Komodakis, N. (**2016**). Wide residual networks. arXiv preprint arXiv:1605.07146.

[124] Takahashi, R., Matsubara, T., & Uehara, K. (**2019**). Data augmentation using random image cropping and patching for deep CNNs. IEEE Transactions on Circuits and Systems for Video Technology, 30(9), 2917-2931.

[125] Deng, L. (**2014**). A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Transactions on Signal and Information Processing, 3.

[126] Angelov, P., & Sperduti, A. (**2016**). Challenges in deep learning.

[127] Arpteg, A., Brinne, B., Crnkovic-Friis, L., & Bosch, J. (**2018**, August). Software engineering challenges of deep learning. In 2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) (pp. 50-59). IEEE.

[128] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., & Farhan, L. (**2021**). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of big Data, 8(1), 1-74.

[129] Xu, Q., Zhang, M., Gu, Z., & Pan, G. (**2019**). Overfitting remedy by sparsifying regularization on fully-connected layers of CNNs. Neurocomputing, 328, 69-74.

[130] 145Sharma, K., Alsadoon, A., Prasad, P. W. C., Al-Dala'in, T., Nguyen, T. Q. V., & Pham, D. T. H. (**2020**). A novel solution of using deep learning for left ventricle detection: Enhanced feature extraction. Computer Methods and Programs in Biomedicine, 197, 105751.

[131] Ioffe, S., & Szegedy, C. (**2015**, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456). PMLR.

[132] Zhang, G., Wang, C., Xu, B., & Grosse, R. (**2018**). Three mechanisms of weight decay regularization. arXiv preprint arXiv:1810.12281.

[133] Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., & Hinton, G. (**2017**). Regularizing neural networks by penalizing confident output distributions. arXiv preprint arXiv:1701.06548.

[134] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (**2017**). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).

[135] Nazeri, K., Ng, E., & Ebrahimi, M. (**2018**, July). Image colorization using generative adversarial networks. In International conference on articulated motion and deformable objects (pp. 85-94). Springer, Cham.

[136] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (**2014**). Generative adversarial nets. Advances in neural information processing systems, 27.

[137] Radford, A., Metz, L., & Chintala, S. (**2015**). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

[138] Mirza, M., & Osindero, S. (**2014**). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

[139] Emeršič, Ž., Gabriel, L. L., Štruc, V., & Peer, P. (**2018**). Convolutional encoder–decoder networks for pixel-wise ear detection and segmentation. IET Biometrics, 7(3), 175-184.

[140] Ronneberger, O., Fischer, P., & Brox, T. (**2015**, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[141] Maas, A. L., Hannun, A. Y., & Ng, A. Y. (**2013**, June). Rectifier nonlinearities improve neural network acoustic models. In Proc. icml (Vol. 30, No. 1, p. 3).

[142] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (**2009**, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[143] Alshazly, H. A., Hassaballah, M., Ahmed, M., & Ali, A. A. (**2018**, September). Ear biometric recognition using gradient-based feature descriptors. In International conference on advanced intelligent systems and informatics (pp. 435-445). Springer, Cham.

[144] Pflug, A., Paul, P. N., & Busch, C. (**2014**, October). A comparative study on texture and surface descriptors for ear biometrics. In 2014 International carnahan conference on security technology (ICCST) (pp. 1-6). IEEE.

[145] Ojansivu, V., & Heikkilä, J. (**2008**, July). Blur insensitive texture classification using local phase quantization. In International conference on image and signal processing (pp. 236-243). Springer, Berlin, Heidelberg.

[146] Kannala, J., & Rahtu, E. (**2012**, November). Bsif: Binarized statistical image features. In Proceedings of the 21st international conference on pattern recognition (ICPR2012) (pp. 1363-1366). IEEE.

[147] Dong, M., & Sun, Z. (**2003**, October). On human machine cooperative learning control. In Proceedings of the 2003 IEEE international symposium on intelligent control (pp. 81-86). IEEE.

[148] Wagner, J., Baur, T., Zhang, Y., Valstar, M. F., Schuller, B., & André, E. (**2018**). Applying cooperative machine learning to speed up the annotation of social signals in large multi-modal corpora. arXiv preprint arXiv:1802.02565.

[149] Vidhate, D., & Kulkarni, P. (**2012**, August). Cooperative machine learning with information fusion for dynamic decision making in diagnostic applications. In 2012 International Conference on Advances in Mobile Network, Communication and Its Applications (pp. 70-74). IEEE.

[150] Kingma, D. P., & Ba, J. (**2014**). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[151] Emeršič, Ž., Štruc, V., & Peer, P. (**2017**). Ear recognition: More than a survey. Neurocomputing, 255, 26-39.

[152] Zhang, Y., Mu, Z., Yuan, L., & Yu, C. (**2018**). Ear verification under uncontrolled conditions with convolutional neural networks. IET Biometrics, 7(3), 185-198.

[153] Alshazly, H., Linse, C., Barth, E., & Martinetz, T. (**2019**). Handcrafted versus CNN features for ear recognition. Symmetry, 11(12), 1493.