REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

AN INTEGRATED SYSTEM FOR GENDER AND AGE DETECTION IN SECURITY SYSTEMS USING DEEP LEARNING

Kanary Alqulub Hatem Habeeb RUBAYE

Master`s Thesis

Supervisor

Asst. Prof. Dr. Muhammad ILYAS

İstanbul, 2024

AN INTEGRATED SYSTEM FOR GENDER AND AGE DETECTION IN SECURITY SYSTEMS USING DEEP LEARNING

Kanary Alqulub Hatem Habeeb RUBAYE

Electrical and Computer Engineering Department

Master of Science

ALTINBAŞ UNIVERSITY

2024

The thesis titled “An Integrated System for Gender and Age Detection in Security Systems Using Deep Learning” prepared by Kanary Alqulub Hatem Habeeb RUBAYE and submitted on (DATE) has been accepted unanimously for the degree of Master of Science in electrical and computer engineering.

Asst. Prof. Dr. Muhammad ILYAS

Thesis Defense Committee Members:

Asst. Prof. Dr. Muhammad Ilyas	Department of Computer Engineering, Altinbas University
Asst. Prof. Dr. Abdullahi Abdu Ibrahim	Department of Computer Engineering, Altinbas University
Assoc. Prof. Dr. Jawad Rasheed	Department of Computer Engineering, Istanbul Zaim University

I hereby declare that this thesis meets all format and submission requirements of an electrical and computer engineering Master’s thesis.

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Kanary Alqulub Hatem Habeeb RUBAYE

Signature

ABSTRACT

AN INTEGRATED SYSTEM FOR GENDER AND AGE DETECTION IN SECURITY SYSTEMS USING DEEP LEARNING

RUBAYE, Kanary Alqulub Hatem Habeeb

M.Sc., Electrical and Computer Engineering Department, Altınbaş University,

Supervisor: Asst. Prof. Dr. Muhammad ILYAS

Date: 06/2024

Pages: 83

The progression of gender and age recognition systems over the course of the past several years paved the way for a plethora of applications in smart real-world scenarios, rigidifying its vital role in the area of computer vision. Even though traditional techniques in particular hand-crafted classifiers, rule-based systems and feature-based methods have indisputably forged the basis for breakthroughs in gender and age recognition. Even so, it manifested certain deficiencies when dealing with subtle intricacies, not to mention wrestling with hurdles posed by varying lighting conditions, diverse facial expressions, and pose variations, thereby restricting the system overall performance. Ergo, the current study proffers to herness the unified might of transfer learning and data augmentations so as to surmount the legacy constraints. There were four prominent deep learning models: Inceptionv3, InceptionResNetv2, Xception, and DenseNet201 were parlayed as a means to develop a robust and unwaveringly reliable gender and age recognition system. As a closing gesture, this study harnessed ensemble model as a means to enhance the accuracy of the system by means of combining the forecasts of the opted models, driving towards more precise predictions which in turn improve the overall performance of the system. The findings of this study set a cutting-edge yardstick for performance, outstripping all prior state-of-the-art models in respect of all the models utilized in this study. Astoundingly the ensemble model shines forth with a striking accuracy of 97.97% when it comes to gender classification task,

together with MAE of 0.77857444 years as regards age estimation. The proposed study irrefutably affirms the pivotal significance of the utilization of transfer learning technique whilst the training phase with an extensive dataset such as IMDB-WIKI and UTKFace. On top of that, it zeroes in on the freezing status alongside data augmentation as potent catalysts for the system performance, eventually bringing about superior accuracy within the scope of gender and age recognition.

Keywords: Convolutional Neural Networks, Ensemble Model, IMDB-WIKI, UTKFace, Gender Classification, Age Estimation.

ÖZET

DERİN ÖĞRENME KULLANILARAK GÜVENLİK SİSTEMLERİNDE CİNSİYET VE YAŞ TESPİTİ İÇİN ENTEGRE BİR SİSTEM

RUBAYE, Kanary Alqulub Hatem Habeeb

Yüksek Lisans, Elektrik ve Bilgisayar Mühendisliği Bölümü, Altınbaş Üniversitesi,

Danışman: Dr. Öğr. Üyesi.Muhammed İLYAS

Tarih: 06/2024

Sayfalar: 83

Geçtiğimiz birkaç yıl boyunca cinsiyet ve yaş tanıma sistemlerinin ilerlemesi, akıllı gerçek dünya senaryolarında çok sayıda uygulamanın yolunu açarak, bilgisayarlı görme alanındaki hayati rolünü sağlamlaştırdı. Geleneksel teknikler, özellikle el yapımı sınıflandırıcılar, kural tabanlı sistemler ve özellik tabanlı yöntemler, cinsiyet ve yaş tanıma konusunda tartışmasız atılımların temelini oluşturmuş olsa da. Buna rağmen, değişen aydınlatma koşulları, farklı yüz ifadeleri ve poz değişimlerinin oluşturduğu engellerle uğraşmanın yanı sıra, incelikli karmaşıklıklarla uğraşırken bazı eksiklikler gösterdi ve bu nedenle sistemin genel performansını kısıtladı. Dolayısıyla mevcut çalışma, eski kısıtlamaların üstesinden gelmek için transfer öğreniminin ve veri artırmanın birleşik gücünü sunuyor. Öne çıkan dört derin öğrenme modeli vardı: Inceptionv3, InceptionResNetv2, Xception ve DenseNet201, sağlam ve değişmez derecede güvenilir bir cinsiyet ve yaş tanıma sistemi geliştirmenin bir yolu olarak kullanıldı. Kapanış olarak bu çalışma, seçilen modellerin tahminlerini birleştirerek sistemin doğruluğunu artırmanın bir yolu olarak topluluk modelini kullandı; daha kesin tahminlere yönelerek sistemin genel performansını artırdı. Bu çalışmanın bulguları, bu çalışmada kullanılan tüm modeller açısından önceki tüm son teknoloji ürünü modelleri geride bırakarak, performans için son teknoloji bir ölçüt oluşturmaktadır. Şaşırtıcı bir şekilde topluluk modeli, yaş tahmini açısından 0,77857444 yıllık MAE ile birlikte cinsiyet sınıflandırma görevi söz konusu olduğunda %97,97'lik çarpıcı bir doğrulukla öne çıkıyor. Önerilen çalışma, IMDB-WIKI ve UTKFace gibi kapsamlı bir veri seti ile eğitim aşamasında transfer öğrenme tekniğinin kullanılmasının hayati önemini inkar edilemez bir şekilde

doğrulamaktadır. Bunun da ötesinde, sistem performansı için güçlü katalizörler olarak veri artırmanın yanı sıra donma durumuna da odaklanıyor ve sonuçta cinsiyet ve yaş tanıma kapsamında üstün doğruluk sağlıyor.

Anahtar Kelimeler: Evrişimsel Sinir Ağları, Topluluk Modeli, IMDB-WIKI, UTKFace, Cinsiyet Sınıflandırması, Yaş Tahmini.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

AI	:	Artificial Intelligence
ML	:	Machine Learning
DL	:	Deep Learning
CNN	:	Convolutional Neural Networks
HOG	:	Histogram of Oriented Gradients
LBP	:	Local Binary Patterns
SIFT	:	Scale-Invariant Feature Transform
SVM	:	Support Vector Machines
CelebA	:	CelebFaces Attributes Dataset
IMDB-WIKI	:	Internet Movie Database-Wikipedia
ANN	:	Artificial Neural Networks
ReLU	:	Rectified Linear Unit
Xception	:	Extreme Inception
ResNet	:	Residual Network
DenseNet	:	Dense Convolutional Network
Adam	:	Adaptive Moment Estimation
MAE	:	Mean Absolute Error
RMSE	:	Root Mean Square Error

1. PRELIMINARY OVERVIEW

1.1 INTRODUCTION

Of late, computer vision has made momentous strides towards the capability of accurately predicting the age and gender of an individual from images [1]. The face has a silent language, it is an astute reservoir of information that reveals a person's true identity [2,3,4]. Nuanced indicators such as the wrinkles around the eyes, shape of the jawline, shape of the nose, the tilt of the chin and others, all of these combined is what offer glimpses to that specific individual's age and gender. This encapsulates the reason why countless sophisticated applications bank on this technique. Extracting High-worth information from a person's singular image possesses enormous prospects, specifically within the sphere of age and gender detection. This technology has become integral in diverse spheres of social interaction and communication [5], extending across a spectrum of domains including surveillance, security, access control, human computer interaction, marketing, advertisements, law application, content personalization, visual observation, among others [6,7].

Through the passage of time, the objective of inferring an individual's gender and/or age experienced a surge of interest [8,9] and a multitude of shifts starting from handcrafted features-based system and finalizing with machine/deep learning algorithms. The preliminary efforts to predict gender/age commenced with handcrafted features-based systems with typical computer vision techniques although those approaches laid the groundwork, it did endure several challenges due to its incapability to deal with the intricacy and inherent variability in facial attributes among multifarious populations not to mention variations in lighting, pose, and image quality [10].

As time progresses, milestones in computer vision and pattern recognition have unsealed the possibility for more intricate and potent techniques. A noteworthy landmark was the development of Haar-Cascade technique, which presented a novel framework for pinpointing faces and basic features founded on cascading classifiers.

Notwithstanding the initial dependence on handcrafted features and shallow learning models, deep learning surfaced as a trailblazer within the realm of computer vision. Deep

learning architectures, notably convolutional neural networks (CNNs), manifested supreme performance in diverse tasks in the field of computer vision, encompassing image classification, object detection, and facial analysis. The developing domain of deep learning went hand in hand with the increasing accessibility of large-scale datasets, like ImageNet and the IMDB-WIKI dataset, wherein served as an invaluable resource for researchers, offering a substantial collection of labeled images for training and evaluation. The amalgamation of data avalanche, enhanced computational capabilities, and algorithmic advancements has considerably enhanced the accuracy and robustness of gender and age detection models.

Notwithstanding the advancements, obstacles remain evident in gender and age detection, encompassing variations in facial appearance on account of many factors such as age, ethnicity, illumination, and expressions. Besides, snags pertaining to dataset bias, privacy concerns, and computational complexity necessitate resolution to empower the practical deployment of these systems.

In a nutshell, the unfolding of gender and age detection through deep learning embodies a journey distinguished by persistent innovation, impelled by breakthroughs in neural network models, data accessibility, and computational capabilities. This voyage has refashioned gender and age detection from an intellectual pursuit into practical technology, profusely impacting contemporary world.

1.2 PROBLEM STATEMENT

During a period marked by heightened security anxieties, there's a more pressing exigency for reliable, efficient and robust systems proficient at accurately identifying an individual's gender and age. Preliminary approaches to age and gender recognition hinged on antiquated machine learning methods, consistently facing noteworthy deficiencies in terms of both accuracy and efficiency. For the purpose of training robust predictive models, early methods exploited handcrafted features for extracting metadata from images. Nevertheless, the intricacies and diversity inherent in human facial features, impacted by many aspects including ethnicity, lighting conditions, and occlusions, pose major barriers for these classical methods. On top of that, the procedure of feature extraction and prediction are computationally burdensome, since it demands substantial computational resources,

particularly in instances where the algorithms aren't tailored for swiftness or are functioning on hardware deficient in dedicated computational specifications.

1.3 THESIS ASPIRATION

Acknowledging the surging demand for high-level security measures and surveillance systems, this study strives to develop a reliable and high-performing gender and age recognition system leveraging deep learning techniques, consequently improving the efficiency and accuracy of security protocols. To narrow it down, the study leveraging transfer learning optimize pre-trained deep learning models to address the task in hand of gender and age recognition. This course of action empowers harnessing prior knowledge imbedded in models trained on extensive datasets, accordingly improving learning efficacy and model performance, particularly when faced with insufficient training data intended for gender and age recognition.

In pursuit of boosting the accuracy and reliability of the models, ensemble learning has been utilized by aggregating the predictions of the opted models. The utilization of ensemble learning was deliberate, intending to alleviate biases or vulnerabilities existing within each model, therefore improve system-wide performance. An essential aspect of the system's design rests in its capability to process data in real-time. With a view to providing expedited analysis of inputted data for real-time applications, it's imperative to pick and optimize models that exhibit computational efficiency, this arises from the fact that real-time applications require immediate feedback, among which are security systems, interactive user interfaces, and live audience analysis.

In a nutshell, the study seeks to unveil a leading-edge solution in the sphere of gender and age recognition by leveraging the power of transfer learning and ensemble models, meanwhile guaranteeing the system possesses the capability to perform accurately and efficiently in real-world scenarios.

1.4 CONTRIBUTIONS

Within the domain of security systems, the integration of cutting-edge technologies has become imperative in improving accuracy, efficiency, and reliability. This study introduces a groundbreaking contribution to this domain through the development of a gender and age

recognition system, harnessing the power of deep learning techniques. The contributions are delineated herein:

a. Harnessing ensemble learning as a view to boost accuracy by combining opted models predictions for both gender classification and age estimation.
b. Leveraging transfer learning technique so as to detect the gender and age of an individual.

1.5 BLUEPRINT OF THE THESIS

This section is dedicated to spotlight the content addressed within the pages of each chapter, elaborated further below:

First Chapter: Within this chapter, a synopsis is depicted embracing the present study, problem statement, and the aim of the thesis.

Second Chapter: This chapter delves into the theoretical underpinnings inclusive of gender and age recognition background, CNN architecture, deep learning, transfer learning, miscellaneous deep learning techniques, models and effective evaluation methods.

Third Chapter: The pursuing chapter illuminates the methodology proposed for this study, encompassing system architecture, dataset acquisition, preprocessing steps, models selection, training procedures, and evaluation methods.

Fourth Chapter: The experimental findings of each model are exhibited and deliberated upon hereon, not to mention specifications of the hardware and software, employed datasets.

Fifth Chapter: This chapter offers a breakdown of the conclusions drawn in addition to recommendations delineated for future undertakings.

2. THEORETICAL UNDERPINNINGS

2.1 INTRODUCTION

The Automated prediction of age and gender from facial images has borne witness to substantial scrutiny in the field of computer vision, which is motivated by its wide-ranging applications in multiple domains namely, security, healthcare, marketing, and human-computer interaction. All in all, leveraging from the vigorous capabilities of deep learning models, remarkably convolutional neural networks (CNNs) has surfaced as a momentous promise approach for gender and age recognition which is stemming from its potential to deliver the highest levels of accuracy and efficiency.

This chapter serve as a stepping stone for the aspire of fostering the development of robust age and gender estimation system, these pages lay the groundwork by elucidating the core concepts for the targeted system. It delves into the CNN architecture, deep learning, transfer learning, miscellaneous deep learning techniques, models and effective evaluation methods.

2.2 GENDER CLASSIFICATION

Gender classification is the process of identifying an individual's gender from visual cues and/or data by utilizing deep neural network architectures. An examination of the historical trajectory of gender classification reveals its presence back to the early days of computer vision research, along with substantial advancements propelled by technological progress and the availability of large-scale datasets. Here's an exploration to the evolution of gender classification through the technological lens:

In the initial phases of computer vision, gender classification principally relied on handcrafted features, rule-based systems and statistical methods. Various algorithms were developed to extract facial features involved jawline shape, eyebrow thickness, eyes, nose, mouth and so on, these extracted features were being fed to the model as a means to train it for gender prediction. Moving on, some of the feature-based methods are Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP) and Scale-Invariant Feature Transform (SIFT), and these were primarily used to capture distinctive patterns and textures in facial images. Still, all of these approaches encountered difficulties in handling variations in poses, illumination, facial expression, and occlusion which resulted in limiting their

effectiveness in real-world scenarios [11]. Eventually, the researchers begin to employee machine learning algorithms such as Support Vector Machines (SVM), random forests to feed the extracted features into the ML classifiers to predict an individual's gender.

Thereafter, the emergence of deep learning marked a turning point in the field of computer vision, achieving considerable breakthroughs in varied tasks. Deep neural network architectures, precisely convolutional neural networks (CNNs), which revealed remarkable progressions [12,13] in the ability of learning and representing features from raw pixel data which in turn resulted in more accurate classifications for gender detection. Furthermore, the availability of large-scale annotated datasets permitted deep learning models to be trained on diverse and representative data which led to improve the performance and generalization of the models. datasets like CelebA, Adience, IMDB-WIKI, and MORPH supplied researchers with immense volumes of facial images labeled with gender annotations.

Putting it all together, the trajectory of gender classification reflects an evident advancement, transitioning from early handcrafted feature-based methods to sophisticated deep learning techniques. Along this path many advancements in algorithms, architectures, and data availability eventuated which entailed in pronounced improvements in accuracy, robustness, and real-world applicability.

2.2.1 Age Estimation

Age estimation, a burgeoning subfield within computer vision, has witnessed noteworthy advancement in the course of time. The bewitching journey of age estimation, demonstrated numerous developments for countless algorithms to estimate an individual's age out of facial images. Here's a quick glimpse into the chronological development of age estimation:

Before the advent of deep learning, age estimation was steered predominantly by rudimentary learning techniques, knowledge-based systems, statistical methods and handcrafted features. Researchers pursued to estimate age accurately by utilizing various facial features, illustrative of wrinkles, skin texture, and facial landmarks with a view to develop age estimation models. However, initial initiatives underwent significant difficulty to achieve high accuracy and robustness, stemming primarily from its disability inability of handcrafted features to competently represent the intricate nuances of facial aging.

With the arrival of machine learning and in particular supervised learning algorithms, researchers granted a potent tool to develop more advanced, sophisticated, effective, data-driven approaches to age estimation, consequently these algorithms were capable of surpassing the constraints of earlier techniques. The presented algorithms exploited machine learning models and leveraged from large-scale datasets of annotated facial images alongside with corresponding age labels to disclose the intricate connections and subtle variations between facial features and age. Hence, the vast quantity of the labelled data shunned the need for manual feature engineering which in turn led to achieve greater accuracy in age estimation. The groundbreaking efforts in age estimation through machine learning utilized many techniques which encompassed Support Vector Machines (SVMs), decision trees, and various regression-based methods so as to learn age-discriminative features from raw pixel data.

The proliferation of deep learning, exceptionally convolutional neural networks (CNNs), triggered a paradigm shift in age estimation, forged a path for the development of highly accurate and data-driven models. Convolutional neural networks (CNNs) surpassed at extracting hierarchical features from raw pixel data, thereby enabled them to unveil intricate patterns, subtle nuances and variations inward facial images [14], bringing about Top-notch performance within the scope of age estimation. Leveraging the potency of deep learning, age estimation models have fulfilled levels of accuracy and robustness far outpacing those of conventional methods banking on handcrafted features.

Notwithstanding substantial improvements in age estimation, there remain significant hurdles to overcome, among them the diverse incarnations of aging across various demographics, the impact of perplexing factors resembling expressions and occlusions, and the need for age estimation models to be accurate, robust and interpretable. By and large, the progression of age estimation emphasizes the paradigm-shifting potency of evolving computer vision and machine/deep learning techniques, transitioning from initial dependence on handcrafted features to the existing dominance of sophisticated deep learning approaches. The proliferating availability of comprehensive datasets are likely to impel further developments in age estimation systems, generating prospects for applications across various domains such as healthcare, security, and entertainment.

2.3 MACHINE LEARNING (ML)

Machine Learning is a subdivision of artificial intelligence, machine learning harness the permanent development of sundry of algorithms and statistical models with a view to furnish computers with the ability to contend with tasks unaided by explicit programming [1,5]. Machine learning capitalizes on data to obtain possession of knowledge and gradationally improve their performance through an iterative learning process [15]. The machine learning process can be decomposed into:

a. Data acquisition: The initial stage mandates the acquisition and meticulous pre-processing of data pertinent to the prescribed task [4]. Machine learning algorithms then persist to lay bare patterns and correlations enclosed in this extensive data, which can manifest in structured formats as a case in point is databases, unstructured formats such as text and images, or semi-structured formats like XML.
b. Model training: The opted algorithm goes through a training process utilizing the provided data. In the course of this process, its internal parameters will be adjusting continually grounded on the identified patterns and relationships [2].
c. Evaluation: The effectiveness of the trained model is rigorously assessed using unseen data; accordingly, the trained model's performance will be gauged [2] to uncover whether efficient for real-world applicability or not.
d. Prediction (a.k.a Decision-Making): Upon achieving satisfactory performance metrics, the trained model will leverage from its acquired knowledge to generate predictions on new data or automate decisions founded on its learned knowledge [2].

An all-inclusive taxonomy of machine learning identifies four major types, differentiated by their learning mechanisms and problem-solving strategies.

a. Supervised machine learning: Could be interpreted as a method that trains algorithms upon labeled datasets [1], in which every example is tied to a target label or outcome, the aim of this endeavor is to train algorithms how to classify data accurately progressing from input features to output labels.
b. Unsupervised machine learning: The fundamental tenet lies in employing algorithms to autonomously uncover patterns, relationships and structures inward unlabeled datasets,

eventually spurring to reveal hidden associations or groupings without the imperative of predefined labels.

c. Semi-supervised learning: It can be stated as it this type of learning bridges the gap between supervised and unsupervised learning [2], it exploits labeled data for structure so as to guide classification, simultaneously it utilizes unlabeled data to extract features with the aim of scalability, which eventually leads to enhance learning efficiency and performance.
d. Reinforcement learning: It can be expressed as it is the learning method in which the algorithm learns by means of trial and error in interactive surroundings, the model learns over time upon the actions it provides. In other words, it receives rewards for positive actions and penalties for negative actions, and according to the sequence of successful outcomes it will be reinforced to develop the superlative recommendation for a specified problem [1].

Machine learning pervades varied domains [6], enabling a throng of applications with substantial impact comprising image recognition, object detection, and self-driving cars, despite the facts that it serves many other disciplines such as healthcare (e.g. disease diagnosis, drug discovery, personalized medicine), finance (e.g. fraud detection, risk assessment, algorithmic trading) and so forth.

2.3.1 Deep Learning (DL)

Deep Learning is a subcategory of machine learning [6], it utilizes deep neural networks to model and solve intricate tasks [5]. Deep learning utilizes complex algorithms and the computer units (neurons) into artificial neural networks (ANNs) with multitudinous layers as a means to train a designated model by mimicking the human brain [5], these layers learn and extract intricate features from data [16,17], resulting in a superior performance in tasks such as image recognition, natural language processing, and so on [17]. The deep learning process can be decomposed into:

a. Data Preparation: The rudimentary step in deep learning entails the meticulous acquisition and pre-processing of data pertinent to the coveted functionality. The utilization of the hierarchical representation learning of data authorizes deep learning models to efficiently navigate and fathom intricate, high-dimensional data structures.

This data manifest in motley formats, spanning from structured elements like databases to unstructured formats like audio, text and images.

b. Artificial Neural Networks (ANNs): The nucleus of deep learning is the ANN; artificial neural networks are a layered structure comprises interconnected nodes dubbed neurons [18]. Apiece neuron procures input signals, performs a calculation on it, and as the closing step it generates the outcome which will be passes to the next layer neurons. The representation in cascade of layers leverages the model, each layer extract evolvingly higher-level features from the data, resulting in exhaustive representation of the cardinal data [17].
c. Training: The backpropagation algorithm acts as the bedrock of deep neural network training [17]. An iterative procedure in which persistently adjusts the model parameters (weights and biases) in an effort to minimize a pre-defined loss function (error) that gauges the divergence between predicted and actual outputs [16,17].
d. Evaluation and Prediction: Post-training, the model undergoes stringent evaluation on unseen data to appraise its effectiveness and generalizability in real-world scenarios [16], breaking ground for prospect application in varied tasks including but not limited to image content prediction, language translation, or realistic text generation.

The pervasive attain of deep learning across myriad domains has empowered a plethora of cutting-edge applications, eliciting significant impact in various arenas [17], encompassing computer vision (e.g. Image recognition, object detection, facial recognition, medical image analysis), natural language processing (i.e. Machine translation, chatbots, sentiment analysis, text summarization) and speech recognition and generation (e.g. Voice assistants, real-time captioning, language learning tools) and suchlike.

2.3.1.1 Convolutional Neural Networks (CNN)

AKA ConvNet, it is a genre of deep learning neural network, was particularly tailored for analysing visual images [17] by processing data about grid-like topology for instance videos and images. Convolutional neural networks have demonstrated remarkable proficiency in extracting features and patterns from unseen data, the data undergo a succession of convolutional and pooling layers [19], eventually it spurred numerous innovations across several domains such as image classification, object detection, image segmentation. The afterwards figure depicts the convolutional neural network structure:

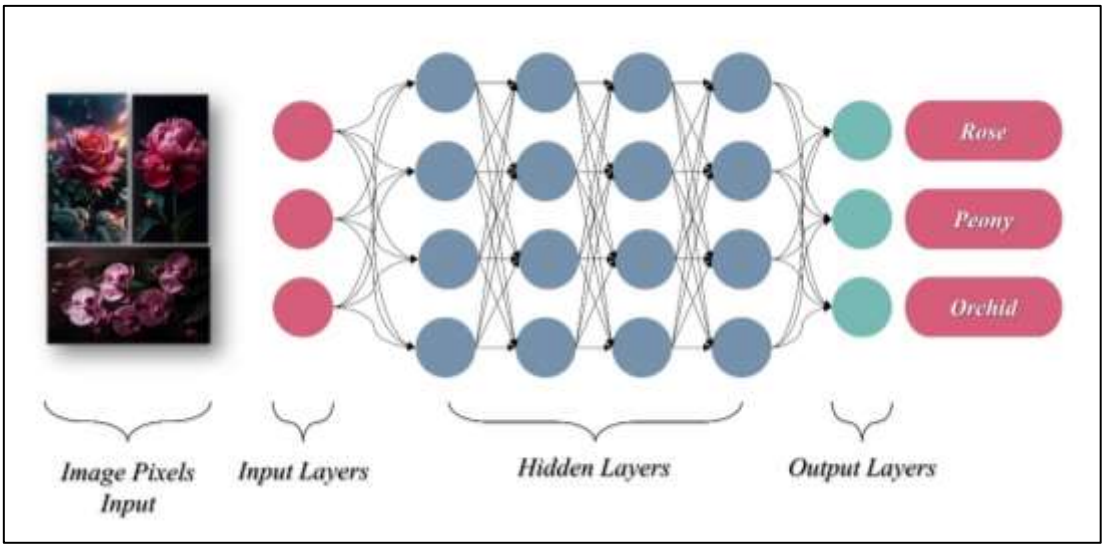

Figure 2.1: CNN Structure.

In the context of CNNs, each and every image is demonstrated in the shape of an array of singular pixel values. In essence, the convolution operation will be performed on the presented arrays, that's basically means that the arrays will be multiplied element-wise and after that the sum of its product will be calculated to form the new array. The subsequent figure manifests the process whereby a CNN recognizes an image:

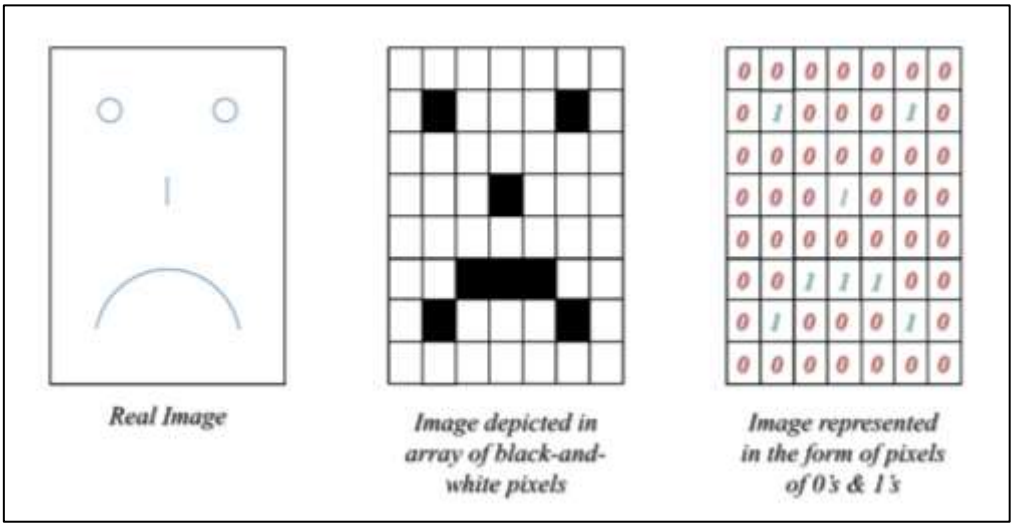

Figure 2.2: How CNN Recognizes an Image.

The structured hierarchy of a convolutional neural network, exemplified by its multitudinous hidden layers, streamlines the gradated extraction of intricate features from an image.

Scrutinizing closely in the foundational layers that constitute the core architecture of a convolutional neural network:

a. Convolution Layers: These layers are the crux of a CNN [16,19], each and every convolutional layer incorporates several filters (alternately termed kernels; usually adopt dimensions of 2x2, 3x3, or 5x5) whereby slide above the input image to execute the convolution operations [16,18]. These filters contribute to capturing valuable features from the input image. Its noteworthy that these filters gradually learn to detect particular patterns within the image (edges, shapes, and textures). Plentiful of these filters result in a profound and nuanced representation of the image content [19].
b. Activation Functions: Upon completion of feature maps extraction, activation function will be utilized to incorporate non-linearity into the network. ReLU symbolizes to Rectified Linear Unit, it is an activation function characterized in tackling the vanishing gradient issue as well as it aids in accelerating the training process. For the aim of posing non-linearity into the network [16,19], ReLU conducts element-wise rectification, exchanging negative pixels with zero whereas conserving the original value for positive pixels, resulting in a rectified feature map being produced as the final output [18]. The succeeding figure illustrates the graphical representation of the ReLU function:

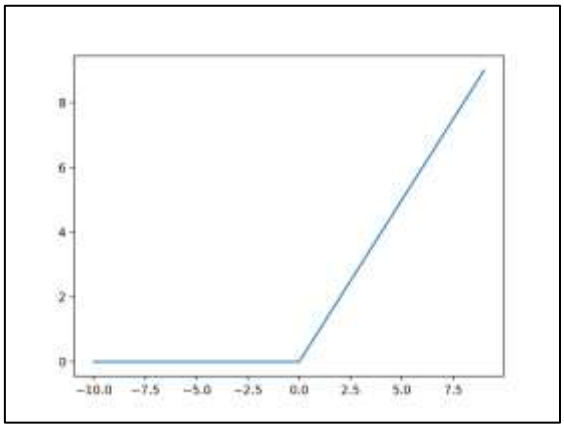

Figure 2.3: ReLU Function.

c. Pooling Layers: With respect to these layers, they fulfil the endeavour of downsampling the feature maps stemming from convolutional layers, leading to effectively diminishing spatial dimensionality [18,19] whereas ensuring that the momentous information will be conserved. The inclusion of pooling layers plays a part in the diminution of the computational cost and overfitting [19]. Pooling layers leverages sundry filters to

identify distinct parts of the image, as an illustration edges, corners, body, eyes, et cetera. Among the miscellaneous pooling strategies utilized in CNNs, max pooling and average pooling stand out as the two most commonly utilized techniques [16]. The pronounced divergence between the max and average pooling techniques is depicted in the figure below:

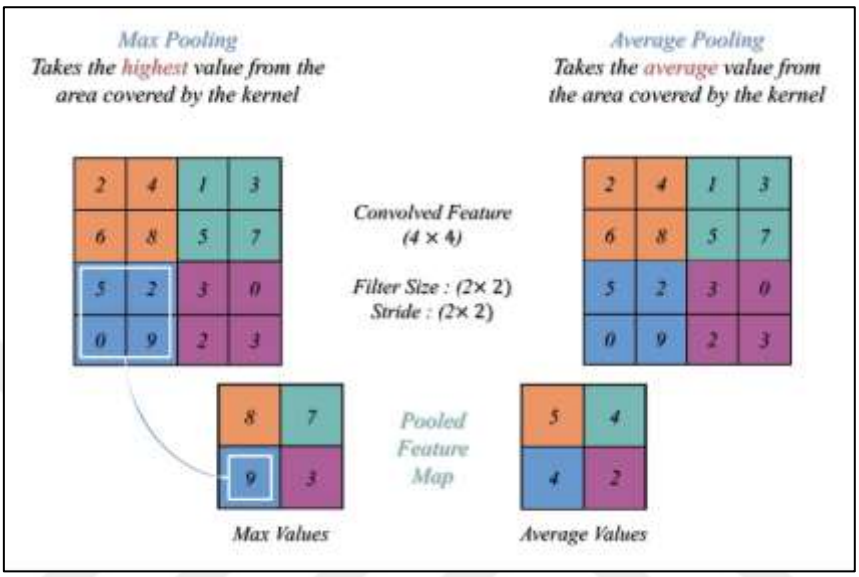

Figure 2.4: Max & Average Pooling Techniques**.**

d. Flattening Layer: The process of flattening entails the conversion of all the 2D and/or 3D arrays from pooled feature maps into a singular, continuous, and elongated one-dimensional vector. To simplify, with the intention of classifying an image, all the matrixes will be reshaped and preprocessed to be properly set up to be fed into the fully connected layers. In the upcoming figures, figure 2.5 showcases the visualization of a flattened matrix, while figure 2.6 illustrates the foundational four layers of the CNN architecture.

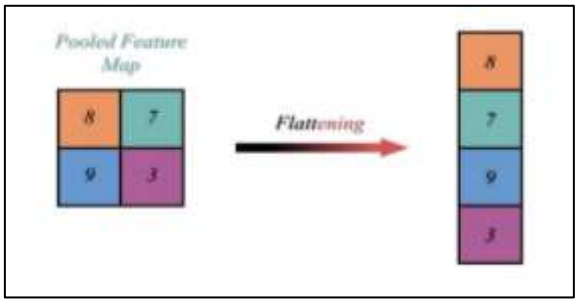

Figure 2.5: Flattened Matrix.

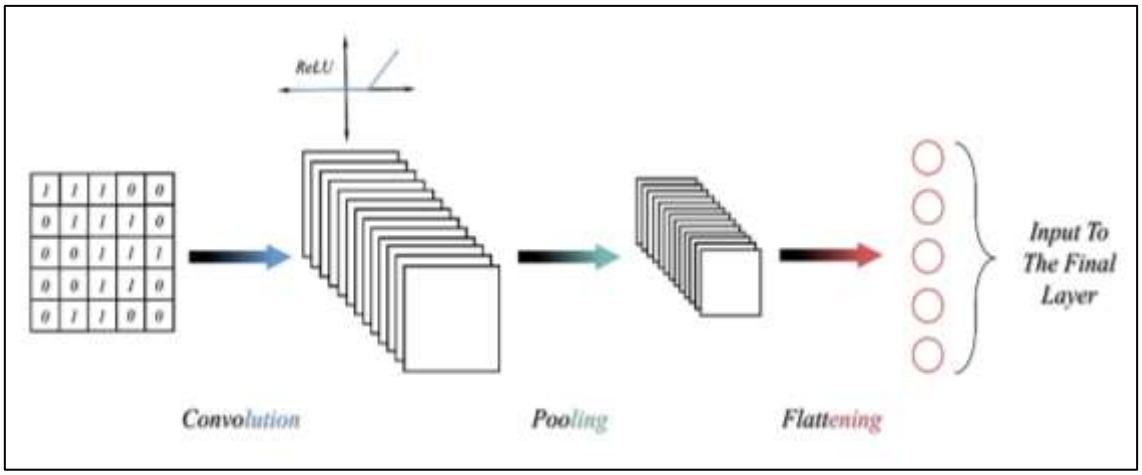

Figure 2.6: CNN Structure Post-Flattening.

e. Fully Connected Layers: Fully connected layers are distinctively stationed at the latter phases of a CNN architecture [19]. These layers seize the high-level features, therein extracted by the convolutional layers, those features will serve as the underpinning for the consequent classification and/or regression tasks to categorize the image [16]. Establishing these fully connected layers ensures the interconnection of all neurons from one layer with those in the succeeding layer [16,19], bringing about the creation of a structure known as a dense network.

f. Dropout Layer: With an eye to alleviate overfitting in ConvNet, a prevalently utilized regularization technique is dropout, which haphazardly drops out a definite percentage of neurons in the course of training [19]. Utilizing this technique, the network's capacity to generalize and effectively handle unseen data is manifestly improves.

g. Batch Normalization Layer: In the pursuit of peak training performance for CNNs, batch normalization is regularly applied. It’s a technique contributes towards escalating both the stability and swiftness of convolutional neural networks. This technique normalizes the activation levels embedded within each layer by way of adjusting and scaling the input data, resulting in accelerating the convergence of the training process [19].

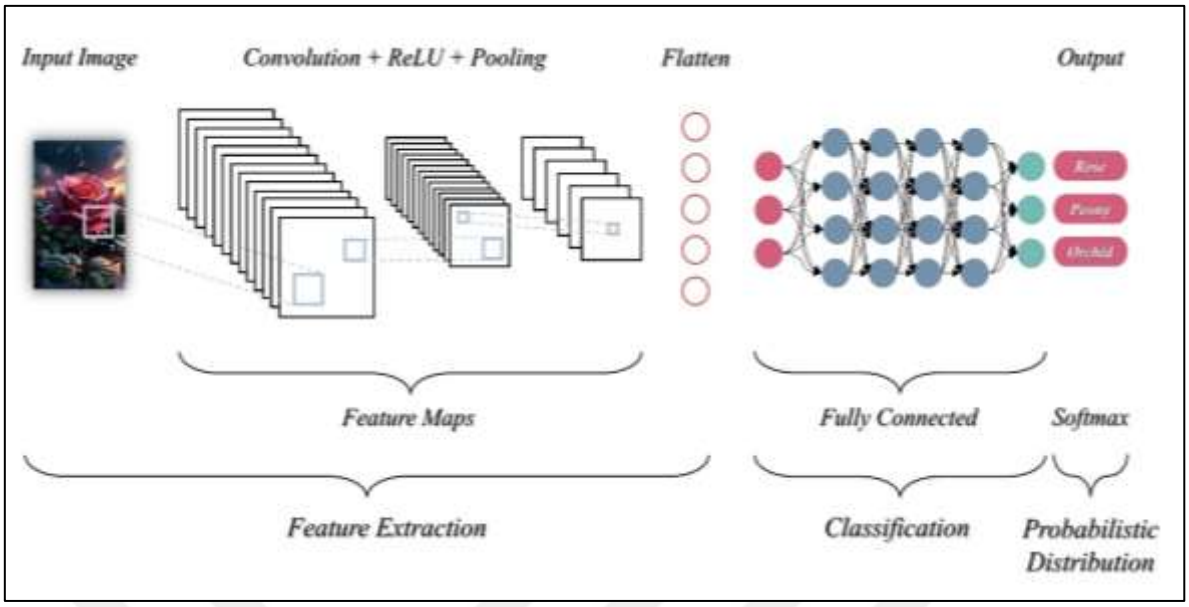

Figure 2.7: CNN Architecture.

2.4 TRANSFER LEARNING

Transfer learning is one of the most potent machine learning techniques, this technique hinges on the maxim of leveraging from the knowledge which was once extracted by resolving a problem so as to train a particular model to expedite the learning process for a distinct but related problem [20]. This technique stands out when the target task exhibits momentous resemblances with the source task, or in cases where data scarcity poses an issue for the target task. Drawing upon the prior acquired knowledge from the source task, the model can accelerate its learning process and achieve more efficient performance on the new task [21]. Here's a breakdown of what the basics of transfer learning:

a. Pre-trained Model: Transfer learning entails the utilization of a pre-trained model that has formerly been trained over a massive dataset like ImageNet (concerning tasks aimed at classifying images) [21]. In view of the fact that these models are trained on a copious amount of data, it learns to extract general features and patterns from the input data.
b. Knowledge Transfer: The pre-trained model will be used as a feature extractor within this step [21]. To simplify, the weights of the pre-trained model will be set as the initial weights and it will be set to be "frozen" which means it will not be updated during the training process. Nevertheless, when addressing the final layers of the model, these are

the task-specific layers which will be eventually trained on the new dataset to adapt to the definite task.

c. Fine-Tuning: This step involves selectively to either partially or completely unfreezing the pre-trained model's layers, while updating the weights in coincidence with the task-specific layers during training process [21]. This process bolsters the model to leverage its pre-acquired knowledge whereas simultaneously adapting to the particulars of the new dataset. The accompanying figure illustrates the process of the transfer learning:

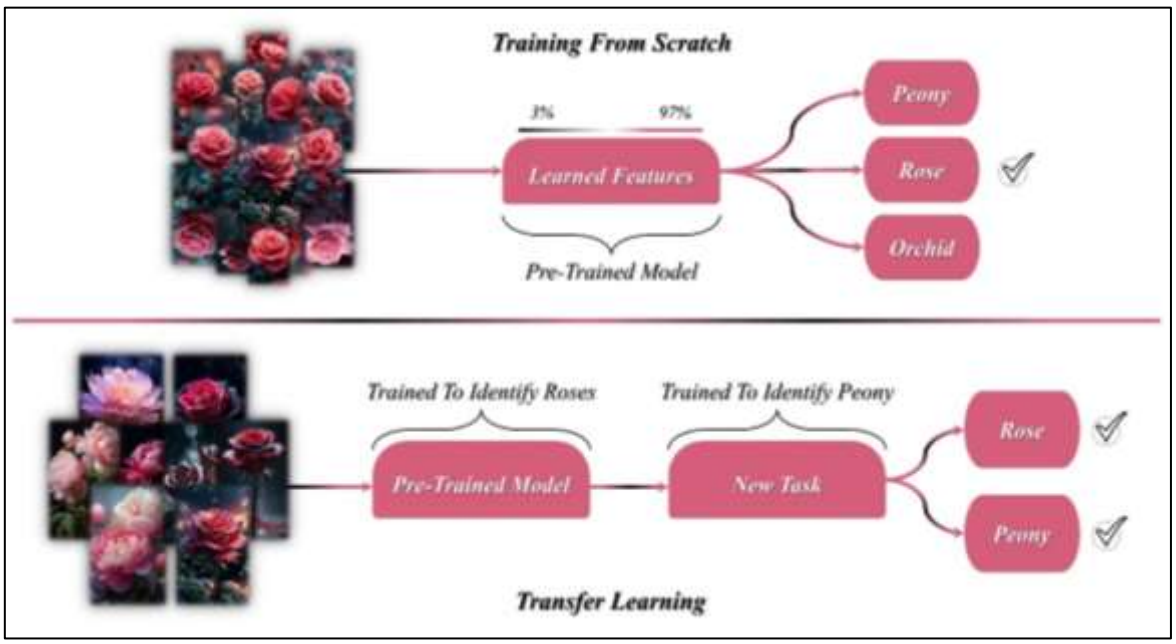

Figure 2.8: Transfer Learning.

Transfer learning possess widespread adoption across a plenitude of domains comprising computer vision (e.g. object detection, image segmentation, sentiment analysis, and image captioning), natural language processing (e.g. sentiment analysis, named entity recognition, and machine translation), and speech recognition (e.g. audio classification and time-series data) [21].

Whereas transfer learning exhibits an effective approach, analysts should be mindful of limitations and obstacles affiliated with its implementation. One of the hurdles is picking the adequate pre-trained model alongside with selecting the appropriate number of layers for freezing or fine-tuning purposes, not mentioning that one of the other dilemmas that could be faced is the heterogeneity between the used datasets because different datasets mean

different characteristics or distributions, which may impact the performance efficacy of the transfer learning technique [20,21].

2.4.1 Inceptionv3 Model

Developed by Google, Inception-v3 is a robust ConvNet architecture notable for its striking performance in the domain of image classifying [24]. Inception-v3 model is affiliated within the firmly established Inception lineage, each and every iteration patently improves image classification performance while preserving an emphasis on computational efficiency [22,24].

Inception-v3 is a multi-layered deep convolutional neural network (CNN), it incorporates numerous convolutional layers which are succeeded by max-pooling layers plus global average pooling layers [23]. This model utilizes a chain of Inception modules which were scrupulously tailored to capture intricate patterns at diverse scales.

As previously stated, the cornerstone elements of the Inception-v3 architecture are inception modules. Each one of these modules carry out manifold convolutions branches in parallel accompanied by various filter sizes (1x1, 3x3, and 5x5) for the aim of extracting features within varying spatial scales [16], on top of that it encompasses a max-pooling branch [22], the joint impact of all of these factors improves the network's performance. Furthermore, and with a view to lessen computational complexity, Inception-v3 employs a dimensionality reduction branch by which converts larger convolutions (i.e., 5x5) into smaller, asymmetric convolutions (e.g., 1x1, 1x3 and 3x1) [22,23]. By doing so, it minimizes the used number of parameters, assuring computational efficiency, whereas minutely maintaining the model's capability of learning and representing intricate patterns. The succeeding figure provides a detailed illustration of the Inceptionv3 model architecture:

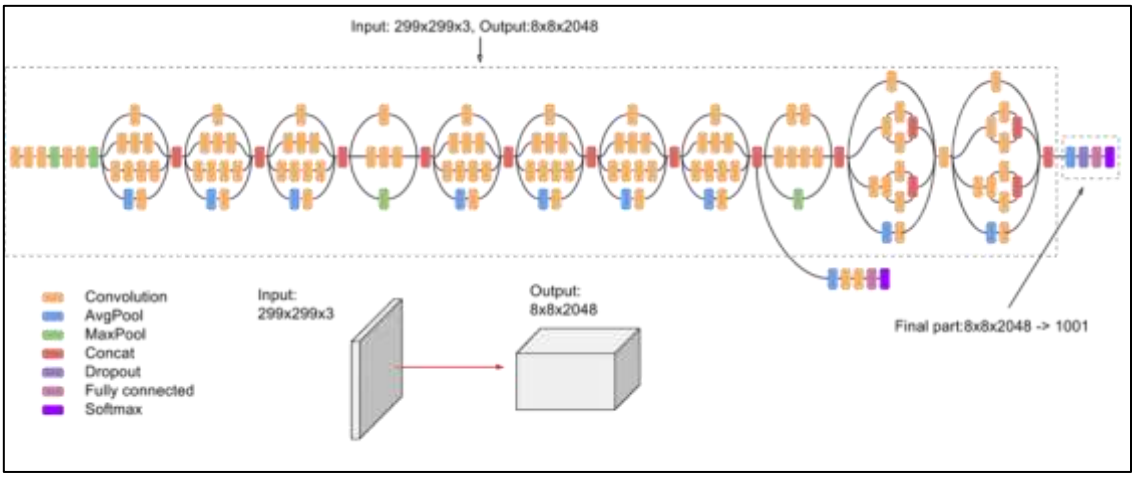

Figure 2.9: Inceptionv3 Architecture [22].

Inception-v3 incorporates auxiliary classifiers in its architecture and precisely in the intermediate layers. These serve as supplementary regularization points [22], facilitating the proliferation of gradients throughout the deeper network in the course of training process, which in turn improves convergence [23]. Additionally, with the aim to alleviates certitude in the model's predictions, Inception-v3 integrate "label smoothing" which is nothing but a regularization technique that slightly adjusts ground truth labels by deterring them from being entirely 0 or 1 [22].

2.4.2 Xception

"Extreme Inception," termed as Xception, is a deep convolutional neural network architecture was specifically designed for image classification and object recognition tasks [25]. It was presented as an enhancement over the Inception architecture, seeking to simplify the network whereas preserving or refining its performance [26]. Here's a breakdown of what the basics of Xception model:

a. Depthwise Separable Convolutions: The depthwise separable convolutions is the central component for the Xception architecture. This process decomposes a standard convolution into two separate operations. Firstly, the depthwise convolution, in which distinct filters are applied to each input channel so as to capture spatial patterns within singular channels [25]. Secondly, pointwise convolution, it combines the output which was acquired from the depthwise convolution by utilizing (1x1) convolutions [27]. This separation considerably diminishes the computational burden and model complexity

whereas maintaining representational capacity, leading to strengthening the model's effectiveness.

b. Aggressive Factorization: The Xception architecture leans heavily on depthwise separable convolutions all over the stages, encompassing the initial and intermediate layers [28]. The aggressive utilization of the factorized convolutions permits further efficient information flow, not to mention the considerable reduction in the number of parameters [27,25].
c. Fully Convolutional Architecture: Xception composed of several stacked depthwise separable convolution blocks, these blocks substitute the traditional convolutional layers which is usually found in the latter part of the network. Alternatively, it utilizes global average pooling to minimize spatial dimensions and create feature maps directly from convolutional layers.
d. Skip Connections: Xception comprises skip connections, alternatively referred to as residual connections (ResNet models), these connections facilitate gradient flow throughout the training process and contribute to alleviate the vanishing gradient issue. These connections grant the gradients to bypass particular layers and flow directly to earlier layers, assisting with the training of the deeper networks.
e. Regularization Techniques: Xception integrates several regularization techniques, by way of illustration, normalization which is applied for the purpose of stabilizing and accelerating the training process, dropout technique as a means to prevent overfitting and enhance generalization performance, et cetera.

The figure 2.10, provides a visual representation of the Xception architecture:

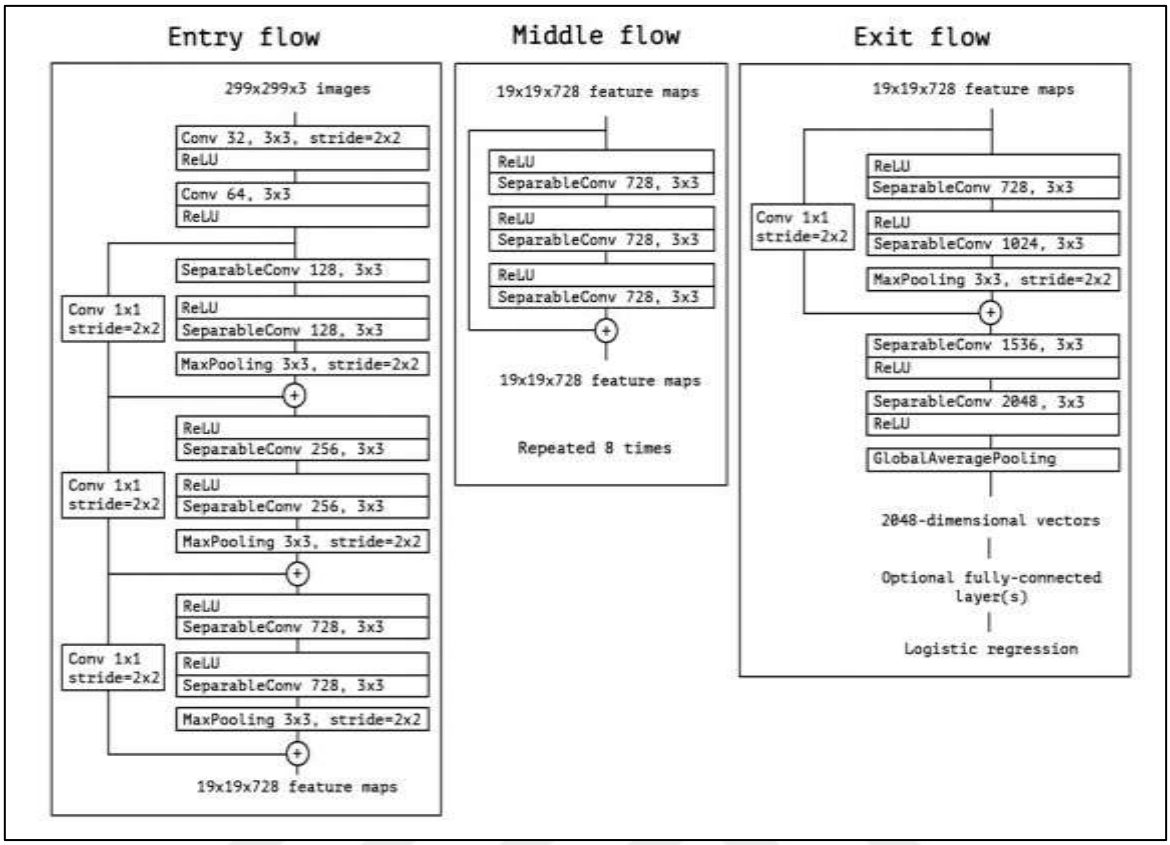

Figure 2.10: Xception Architecture [26].

2.4.3 InceptionResNetv2

Inception-ResNet-v2 is a robust deep convolutional neural network architecture that merges the capacities of both Inception modules alongside the efficiency and improved training dynamics of residual connections for the purpose of improving the performance to a greater extent [29,30]. Here's a comprehensive analysis of the InceptionResNetV2 architecture:

a. Inception Modules: Expanding on the groundwork of Inception modules utilized in the Inception architecture, InceptionResNetV2 leverages from the same building block. As stated before, the modules comprise of several parallel convolutional layers along with various filter sizes [31]. By having parallel pathways, the model gains the ability to capture features at dissimilar spatial scales efficiently [29].
b. Residual Connections: InceptionResNetV2 encompasses residual connections modelled after the ResNet architecture. The residual connections provide the model with ability to

learn residual mappings, which in turn allows for easier optimization, not mentioning that it alleviates the vanishing gradient problem during training [29].

c. Factorization Layers: With a view to achieving computational efficiency [31], InceptionResNetV2 assimilates factorization layers, therefore the model's intricacy will be reduced whereas maintaining the accuracy of the performance. These layers substitute some of the larger convolutions with smaller convolutions, resulting to minimize the number of parameters by which computations requires.
d. Batch Normalization: Utilizing batch normalization across the entirety of the architecture improves training stability as well as expediting the training process. This technique regulates the activations of each layer, decreasing internal covariate shift and enhancing convergence.
e. Auxiliary Classifiers: The inclusion of auxiliary classifiers at intermediate layers [29] in InceptionResNetV2 serves a twofold objective. Initially, the classifiers furnish supplementary supervision signals over the course of training process, and secondarily, they help in mitigating the vanishing gradient problem by presenting auxiliary gradients in the backpropagation stage.

To wrap things up, the model relies on a modular design, employing repetitive structures known as "inception modules" as a means to capture features from image data. These modules are subsequently succeeded by pooling layers for diminishing the spatial dimensions [29] and fully connected layers for performing the last classification task. The network generally terminates with global average pooling to aggregate feature information and a softmax classifier to generate class probabilities for prediction. The figure 2.11 down below offers a visual representation of the InceptionResNetv2 architecture.

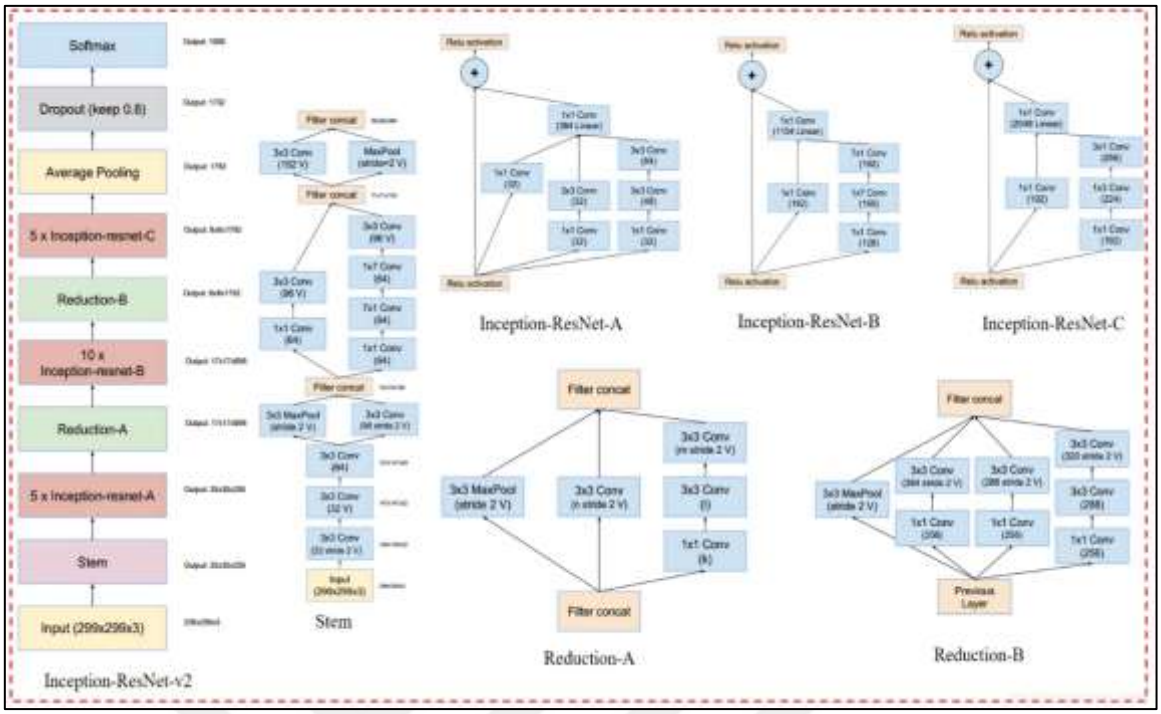

Figure 2.11: InceptionResNetv2 Architecture [22].

2.4.4 DenseNet201

"Dense Convolutional Network" forms a mighty ConvNet architecture eminent for its capability of achieving exceptional performance and efficiency in image classifying tasks [32]. DenseNet was originally developed to combat the vanishing gradient issue in deep neural networks. Noteworthy, it falls under the DenseNet family, which is marked by a unique method to how layers are interconnected inside the network. here's a comprehensive analysis of the DenseNet201 architecture:

a. Dense Connectivity: The primary breakthrough of DenseNet architecture is its dense connectivity pattern, guaranteeing that each and every layer acquires an input from all foregoing layers in the dense connectivity [32,33]. DenseNet architecture facilitates an effective feature reuse, nurturing smooth information flow across the network. This process gives the model the capability of learning more intricate-level features by way of leveraging the knowledge accumulated in formerly stages.
b. Dense Blocks and Transition Layers: The architecture incorporates dense blocks and transition layers. DenseNet engages a sequence structure of densely connected convolutional layers. These layers leverages from the aggregate features extracted

through the previous layers within the block as an input, alongside its own input, prompting feature reusability and information propagation. Subsequently, the architecture integrates transition layers deliberately situated between dense blocks. These layers fulfil the crucial role of decreasing the number of feature maps and diminishing the spatial dimensions of the data. This promotes control over model complexity, therefore alleviating the risk of overfitting.

c. Network Depth: The architecture was aptly named DenseNet-201 owing to its structure comprising of 201 layers [32], is situated in the framework of deep networks. this substantial depth, accompanied by the distinguishing characteristic of dense connectivity, grant the model the capability to capture intricate interactions within the data, eventually contributing to its high degree of accuracy.

On the whole, other architectural features was incorporated in the design of the DenseNet-201 model including batch normalization as a means to stabilize the training process and improve the gradient flow, global average pooling so as to summarize the features captured through the network, not to mention softmax classifier for the purpose of generating class probabilities for final image classification. An illustration of the DenseNet201 model architecture is provided in figure 2.12.

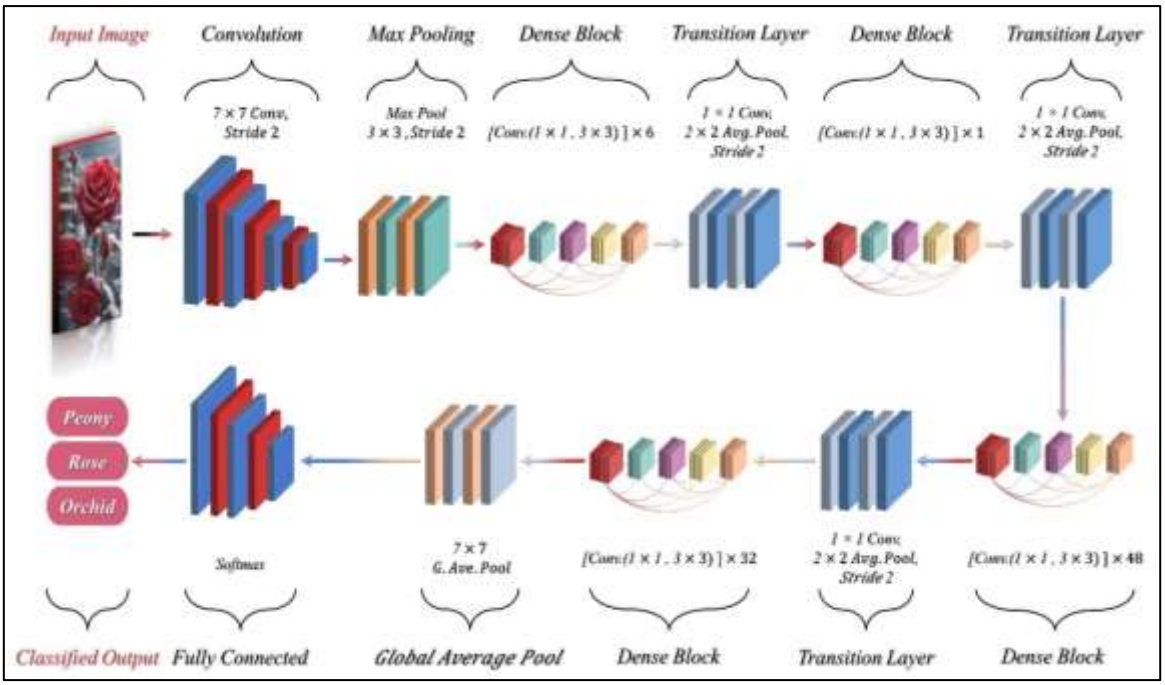

Figure 2.12: DenseNet201 Architecture.

2.5 HAAR-CASCADE TECHNIQUE

Haar-Cascade algorithm exploits machine learning techniques to discern and pinpoint the location of objects within images [34]. By virtue of its low computational requirements [35], this algorithm is ideally suitable for real-time applications.

The nucleus of the algorithm lies in the use and collection of a batch of features dubbed Haar features, haar features are a rectangular group of features that capture the variation in pixel intensity between adjacent regions at a definite location using a detection window [34]. The detection window will be divided into tinier subregions so as to start the calculation, this computation entails summing the pixel intensities within each subregion and calculating the dissemblance between the sums. Haar features comes in possess diverse sizes and shapes, among them rectangles, squares, and such [35], as depicted in the figure beneath:

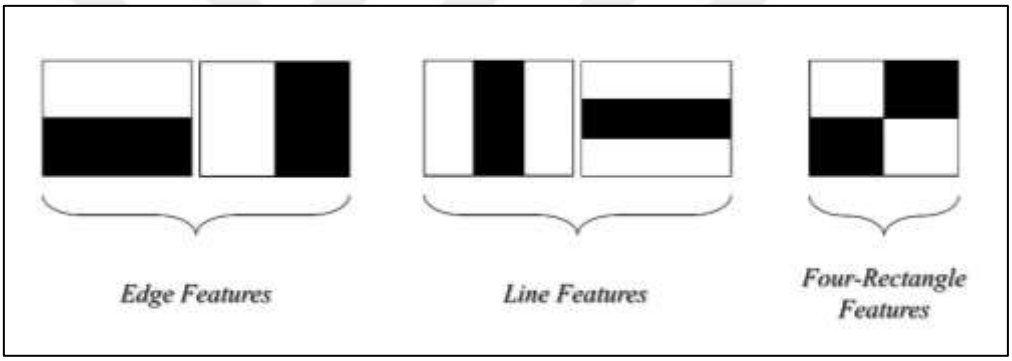

Figure 2.13: Types of Haar Features.

With regard to larger images, the computational burden affiliated with this process scales considerably, rendering it less viable in such scenarios.

For efficacious calculation of the sum of pixel intensities within a haphazard rectangular region of an image, the Haar-Cascade algorithm harnesses an integral image. On the contrary of conducting each and every pixel singularly, the algorithm segregates the image into a smaller sub-rectangle. It subsequently establishes array references for each sub-rectangle, allowing efficient computation of Haar features. It's crucial to highlight that once object detection is performed, almost all of the Haar features will be impertinent which will be excluded, and only the relevant features of the object will remain. The figure beneath showcasing the creation and utilization of an integral image:

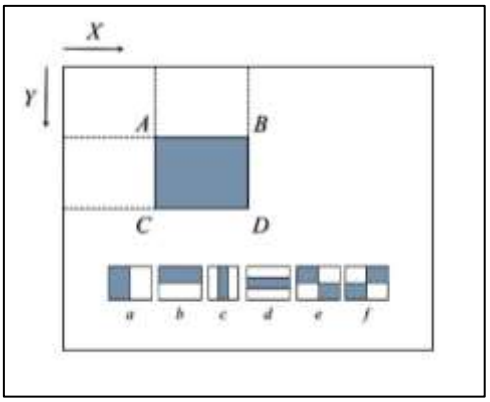

Figure 2.14: Integral Image.

Thereafter, the Haar-Cascade algorithm capitalizes on the Adaboost machine learning technique in the course of the training phase. Across the span of the training phase, a plethora of training data embracing both positive and negative samples is employed to empower the algorithm with the imperative knowledge for accurate performance [36]. In this context, the positive samples represent the object of interest (can for instance faces), whereas the negative samples represent the regions to be excluded [35]. As a means to opt for the best features and training the classifiers to utilize them, Adaboost iteratively picks a batch of weak classifiers by passing a window over the input image, computing the Haar features for each subdivision of the image. Then the dissemblance will be compared to a learned threshold which will disjoins the positive samples (object of interest) from the negative ones (irrelevant objects). It's noteworthy that with a view to form a strong classifier, abundant of Haar features is requisite. The figure below shows a visual depiction of the boosting algorithm.

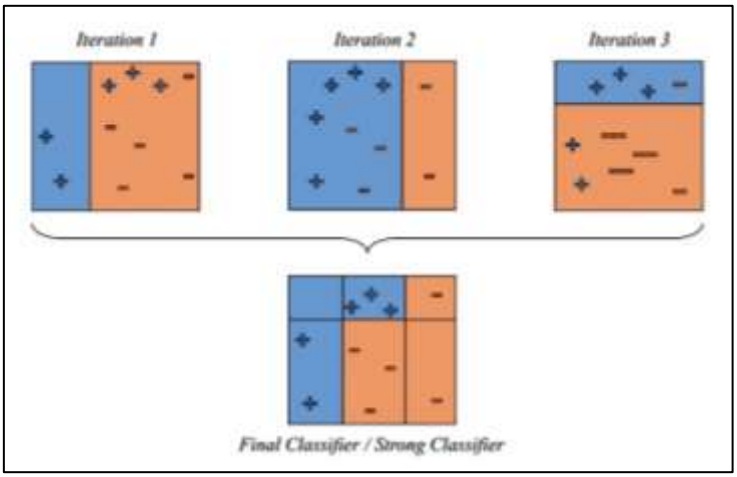

Figure 2.15: Representation of a Boosting Algorithm.

The endpoint of the process entails the combination of these weak classifiers into a strong classifier [34] by utilizing a cascading structure. The cascade classifier is constituted by a sequence of stages, each and every one of these stages enclose a compilation of weak classifiers. Given that these weak classifiers trained utilizing boosting, this triggers a highly accurate classifiers by exploiting the combined strength of numerous, individually weak classifiers, acquired via computing the average of their individual predictions. Grounded on this prediction, the classifier across every stage dictates if the object is going to be indicated as a positive object or negative object. If it is positive object, that one may proceed to the next stage. Otherwise, the object is negative and will be rejected and will skips to the subsequent region [36]. Lessening the false negative rate is critical, as misidentifying an object as non-existent eminently compromises the performance of the object detection algorithm. A breakdown of the cascade classifier's structure is demonstrated in the flowchart below:

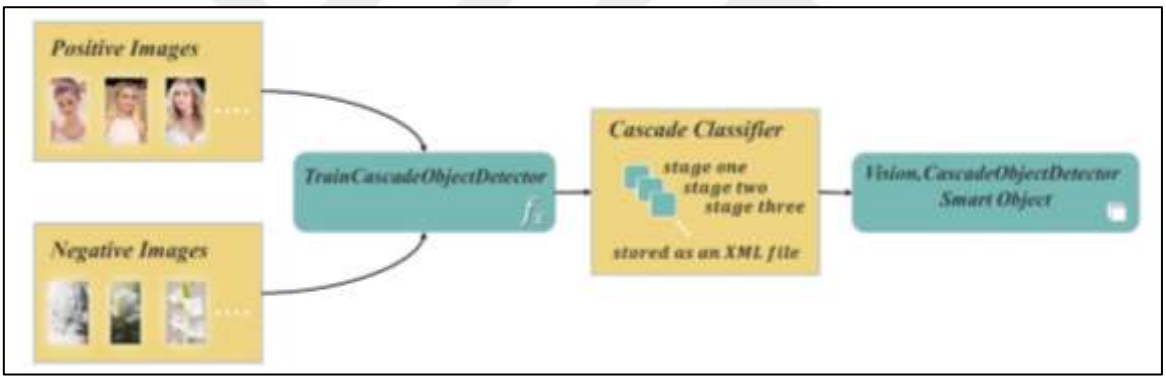

Figure 2.16: Cascade Classifiers.

2.6 EVALUATION METHODS

Vigorous evaluation methods are mandatory for gauging the performance and efficacy of the CNN trained models. The determination of a suitable evaluation method is critical, as it is contingent upon both the particular problem being tackled and the type of model utilized. Below, an illumination over the most utilized evaluation techniques for image classification:

a. Accuracy: An extensively employed metric in images classification. estimates the proportion of instances accurately categorized contrasted against the aggregate number of instances [27,30,33]. It is imperative to acknowledge that, in spite of the fact that

accuracy is a momentous metric, it can be misguiding when grappling with imbalanced datasets. The succeeding formula serves to calculate the accuracy of the specific model:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.1}$$

Bearing in mind that, TP signifies the number of true positives, TN indicates the number of true negatives, FP denotes the number of false positives and FP express the number of false negatives.

b. Precision: Precision quantifies the quotient of true positive predictions to the aggregate number of positive predictions introduced by the model [27,30,33,37]. Precision is exceptionally relevant when false positives are costly or have unfavourable outcomes. Precision can be calculated by utilizing the following equation:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2.2}$$

c. Recall (Sensitivity): Recall, stated as the metric in which estimates the proportion of true positive predictions out of all actual positive instances inside the dataset [27,30,33,37], particularly scales the model's capability to identify all relevant instances. Recall Attains paramount importance in situations where the cost of false negatives is high. The subsequent formula is used to calculate the Recall for a given model:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2.3}$$

d. F1 Score: F1Score can be measured as the harmonic mean of precision and recall [30,33], maintains an equilibrium between these two metrics. This makes it specifically beneficial in scenarios where the dataset exhibits a lopsided distribution of classes. The formula in which calculates F1 score as follows:

$$\text{F1Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.4}$$

e. Mean Absolute Error (MAE): this metric measures the average of the absolute discrepancy between the predicted values and the actual values within a uniform dataset [38]. The equation that calculates the mean absolute error is provided below:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2.5}$$

Keeping in mind that, y_i denotes the actual value, $\hat{y}_i$ signifies the predicted value and "n" indicates the number of samples.

f. Root Mean Square Error (RMSE): this metric measures the mean value of the errors between actual and predicted values as a means to assess the performance of regression models in particular [38]. The equation that computes the root mean square error as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\int_{i=1}^{n}(Y_i - \hat{y}_i)^2} \tag{2.6}$$

2.7 SCHOLARLY EXPLORATION

Diverse techniques were employed to predict an individual's gender/age founded on his/her facial features, to point out a few:

2.7.1 Gender Classification Review

Nga et al [38] presented a model for gender and age prediction based on a transfer learning pipeline which availed from the images from IMDB-WIKI dataset for training and testing stages of the models. The suggested methodology entails freezing all the layers in pre-trained ImageNet models, subsequently and upon scheduled learning rates a block of layers would be gradually unlocked, this step beneficial for feature extraction because it permits the model to leverage knowledge gained from the ImageNet dataset. In the final analysis the author stated that "The age losses and gender losses of all models are interdependent and they have the same patterns" and declared the highest accuracy that was achieved is approximately 91% for gender detection using the VGG-19 classification models.

Benkaddour et al [39] have presented a gender prediction and age estimation systems predicated on convolutional neural networks applied to face images and real-time video. Along those lines, three CNN models (CNN 1, CNN 2, and CNN 3) with diverse architectures in terms of different filter size, pooling layers, and number of convolutional layers. For this aim IMDB-WIKI dataset was used for gender and age prediction. When all is said and done the author declared that CNN3 showed the tops accuracy rate over CNN1 and CNN2, CNN3 model showed that the highest reached accuracy for IMDB dataset was

94.46% and 93.65% for WIKI dataset. The author deduce that the depth and the number of the filters used for the model had a splendid impact for building an efficient convolutional network ranking.

Singh et al [40] discusses the utilization of two disparate models for the automated detection of age, gender, and emotion in images, the first model was designed for the purpose of age and gender detection using wide ResNet architecture while the second model made with conventional CNN architecture for emotion recognition. Each model was supported by different datasets, IMDB-WIKI for age and gender detection and Fer2013 dataset from kaggle for emotion recognition. Furthermore, the author mentioned that the architecture of wide ResNet comprises five groups, average pool layer, and a classification layer incorporated at the end of convolution groups. In this article it was proclaimed that the uppermost accuracy achieved was 96.26% for age and gender detection using the residualnet architecture which is quite good and remarkable improvement in the performance of the model.

The authors [41] debate the strenuous undertaking of age group and gender classification in unconstrained images, the proffered method utilizes an end-to-end deep learning approach (precisely, a convolutional neural network (CNN)) to learn the relevant informative representations of age and gender directly from the image pixels. Furthermore, the CNN model undergoes a multi-step training strategy: at the outset, the model pre-trained on the large-scale IMDB -WIKI dataset. Next, fine-tuning performed over another large-scale facial aging dataset which is "MORPH-II" and lastly fine-tuning stage takes a place on the original dataset (OIU-Adience benchmark) with gender and age group labels. At last, the author proclaims that the model performance and/or accuracy was evaluated on the OIU-Adience benchmark dataset, the results manifest the state-of-the-art performance, with exact accuracy of 96.2% on gender.

A model [42] was suggested to extract discriminative features from unconstrained real-life face images for age and gender classification. The authors accentuate Convolutional Neural Networks (CNNs) as an essential tool for dealing with unconstrained real-world faces. Moreover, a rugged image preprocessing algorithm utilized to deal with the large variations inherent in real-life faces. Pretraining occurred on the IMDB-WIKI dataset in furtherance of enhancing the model's adaptability to real-world data. Additionally, regularization

techniques (dropout and data augmentation) appended in the interest of mitigating overfitting risk and increasing the model efficiency of generalization on test images. To wrap things up, the novelist declared that the proffered method validates its effectiveness on the OIU-Adience dataset by showing an accuracy of 89.7% for gender recognition task.

Fadhlan [43] offered a custom CNN architecture for gender classification with the view of minimizing the sophistication by enhancing the convolutional layers plus lessening the count of neurons in fully connected layers. The proposed architecture encompasses 7 convolutional layers, 2 fully connected layers, and batch normalization layers. The proposed approach marked by: the simple demands for inputted image resolutions and the lesser number of training parameters it requires. Subsequently, the authors utilized three distinct datasets: LFW, CelebA and IMDB-WIKI and disclosed that the utmost accuracy was achieved with IMDB and WIKI dataset by attaining 97% which underscore the robustness of the model and its capability of generalizing on entirely separate datasets.

Khaing and Myint [44] introduced a gender detection approach built upon fast R-CNN employing hyperparameter optimization grounded in Nelder-Mead methods as a means to enhance overall system performance. The authors exploited three different wide-range datasets which are: IMDBs, Asia Image Dataset and new Myanmar Image Dataset. The methodology of the study involved three steps, the first step was to execute the fast R-CNN algorithm, following that was employing hybrid model so as to tackle the hyperparameter tuning problem in support vector machine modelling, at last training the model with openCV2 over the aforementioned datasets. In essence, it was declared that the accuracy of gender classification task excels with Myanmar Image Dataset while achieved the lowest accuracy rate when dealt with IMDB dataset (fell below 60%).

2.7.2 Age Estimation Review

Osekhonmen and Erastus [45] have exploited deep learning methods derived from ResNet50 convolutional neural network (CNN) as a way to boost accuracy in the automatic estimation of dark skin individual ages. With the objective of training the CNN model, the authors utilized a compilation of individuals with dark skin tones from three datasets which are; UTKFace, APPA-REAL and BlackFaces. The methodology of the study encompassed many stages including the data collection, applying data augmentation, splitting the facial aging

dataset into training and testing data, employing ResNet50 CNN as a method to estimate a person age, and lastly enacting the mean absolute error (MAE) as a way to evaluate the performance of the ResNet50 CNN. The authors proclaimed that the suggested solution reached a mean absolute error of 5.21 years on the validation set plus demonstrated commendable results in the estimation of dark skin tone persons.

Osekhonmen et al [46] presented a deep learning system for age estimation over four multi-racial groups for the aim of attaining an accurate Automatic Facial Age Estimation (AFAE). In order to develop an AFAE system suitable for use in automate ages of individuals, transfer learning implemented to the ResNet50 CNN which was originally pretrained for the sake of object classification. Ultimately, the authors articulated that the model displayed a peak performance free from overfitting or underfitting, at last, mean absolute error was used to evaluate the model's performance through which revealed a MAE of 4.25 years.

Arwa and Lamiaa [47] have introduced a model founded upon VGGFace with the intention of surmounting the overfitting problem, through the fine-tuned of the CNN model which is basically was pre-trained for face detection task as a means to estimate a person's age. With an eye to bolster the model's robustness, fine-tuning the models was employed on two distinct approaches and algorithms (classification and regression), subsequently the model was evaluated for age recognition on two different datasets: the constrained FG_NET dataset and UTKFace dataset. The authors stated that a MAE of 3.446 was reached with FG_NET dataset and 4.867 with the UTKFace dataset.

Jesy and Wahyono [48] have offered a framework employing support vector regression (SVR) in conjunction with texture-based feature extraction for the purpose of developing a system for age estimation by exploiting a person's facial features. The methodology of the proposed framework enclosed three steps: preprocessing, feature extraction, and age estimation. As for feature extraction process, three methods were utilized which are: Local Binary Pattern (LBP), Local Phrase Quantization (LPQ), and Binarized Statistical Image Feature (BSIF). Later on, with a view to diminish the feature size, Principal Component Analysis (PCA) was applied, and lastly Support Vector Regression (SVR) method used to estimate the age of a person. Eventually, MAE was used to evaluate the performance of the model and UTK Face dataset was put to use so as to train the model. In the final analysis,

the authors unveiled that the combination BSIF, LPQ and PCA yielded the smallest MAE of 9.766 and 9.754.

Saber et al [49] have introduced a cross-domain multitask learning (MTL) model for the purposes of object detection, segmentation and pose estimation. The model was constructed in accordance with the Mask-RCNN object detection model whereby it computes a variety of mid-level shared features concurrently with independent neural networks so as to detect objects and estimate attributes. The authors harnessed two extensive datasets for the evaluation process, the achieved mean absolute error over the Prima dataset is 8.0 ± 8.6, and 8.2 ± 8.1 for yaw and pitch detection, as regards to BIWI dataset, MAE of 6.2 ± 4.7 and 6.6 ± 4.9 was obtained. In the grand scheme of things, the training and testing process of the presented model was performed on the public dataset UTKFace, resulting in MAE of 5.3 ± 3.2 which affirmed the commendable performance of the model and its effectiveness.

A novel approach [50] for estimating an individual's age leveraging convolution neural network was suggested by employing the deep residual network model. The assessment and evaluation of the residual network entailed initially the collection of data over the UTKFace dataset which was divided into training data (90%) and testing data (10%), in addition to that Adam (adaptive learning rate) was applied for the sake of boosting the training speed, lastly, Mean Absolute Error (MAE) held a position to evaluate the model. The authors concluded that the ResNeXt-50 (32×4d) architecture demonstrated a superior performance (minimal of 0.012) compared to Resnet-50 with linier regression model with a minimal of 1.2807.

3. METHODOLOGY

3.1 INTRODUCTION

This chapter delves into the sophisticated procedure of gender classification and age estimation resorting to deep learning techniques. Emphasizing on convolutional neural networks (CNNs), the chapter scrutinizes the systematic approach adopted for developing an accurate and sturdy model for discriminating age and gender from facial images. By elaborating on the system architecture, dataset acquisition, preprocessing steps, models selection, training procedures, and evaluation methods, this chapter provides an in-depth framework for comprehending the methodology underlying the system of age and gender detection in deep learning.

3.2 SUGGESTED SYSTEM ARCHITECTURE

In an effort to unearthing fresh perspectives through a pioneering technique, the posited methodology of this paper puts at the forefront reliability, striving to capture intricate patterns by which foster an accurate gender and age prediction to guarantee a liable development. The proposed method was constructed upon the foundation of the two most comprehensive, famous large-scale datasets. IMDB-WIKI dataset, it offers the underpinning for the CNN's training for gender classification, and UTKFace dataset, this dataset provides the bedrock for the age estimation task. Foremost, ahead of feeding the images for initiating the model training, the raw data is well-crafted by exposing it to determined preprocessing techniques. Later on, the model training will be fuelled by the transfer learning technique in conjunction with the application of data augmentation (solely during the training phase). Additionally, in pursuit of enhancing the predictions of the models for age and gender recognition, ensemble learning technique will be utilized which is a way of merging several models, the intention is to leverage each model strengths while overcoming their shortcomings, resulting in more accurate and reliable predictions. Last but not least, the performance of the model will be evaluated on the tested dataset from the perspective of accuracy and validation. Figure [3.1] serves as a decoder for elucidating the building blocks of the suggested system architecture.

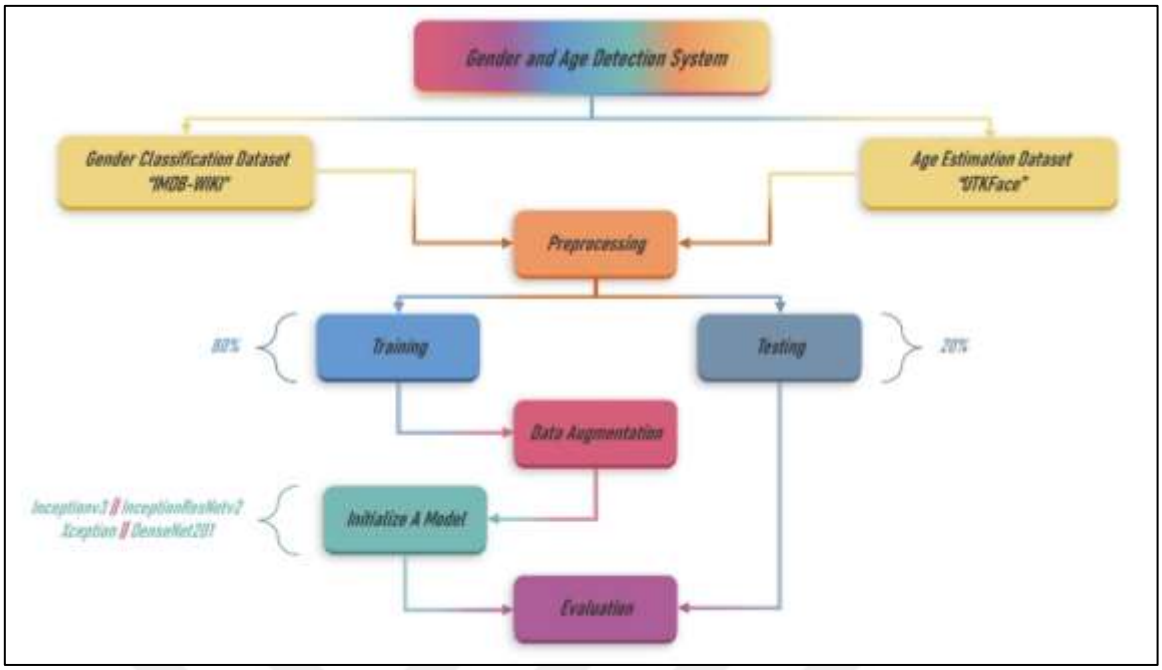

Figure 3.1: Suggested System Architecture.

3.3 GENDER CLASSIFICATION: PREPROCESSING PHASE

Preprocessing embraces a series of manipulations and adjustments enforced on input data ahead to its inclusion into a neural network for training, testing, or inference tasks. Moreover, this phase is paramount and that stems from its impact of improving the input data quality, bolstering the performance and durability of CNN models, optimize learning proficiency, all of these attributes consolidated make it seamlessly compatible with the network architecture. Developing a resilient model necessitates premium-grade data, and this study embraces painstakingly culled bundle of preprocessing strategies, to fine-tune and elevate both the learning process and the model's performance. To provide a clearer view of the process, figure [3.2] illustrates the stages involved.

Stage one: With the aspiration of developing accurate and unbiased prediction models, this study leverages the IMDB-WIKI dataset for training and testing. What discriminated IMDB-WIKI dataset from its counterparts is its intricate amalgamation of paramount metadata, extending from personal information to image specifics. With a view to train the models for gender recognition task, metadata serves as an indispensable factor, owing to the images information it provides for such a tasks, here's some of the most profound image traits that this study leveraged from for gender detection:

a. (gender): The parameter delineates the individual's gender in a certain image by expressing it as a binary value. '0' denotes females, '1' denotes males, and 'NaN' denotes unknown.
b. (face_score): This parameter serves as a detector score; it mirrors the certainty level in the existence of a face in a specific image. to simplify, the parameter scans the image seeking for faces. If no facial features are detected, it returns a value of 'Inf' to signifies that no face was found in the image. Otherwise, it returns the detected face. A higher score alludes to a higher likelihood of a face being present in the image, making it a more reliable indicator of facial presence.

Stage two: Building upon the first stage, this successive stage keeps an eye on the "face_score" metric, it will eliminate any image that was flagged by the metric given that it lacks facial presence.

Stage three: Building upon the foundation of the first stage, this stage will exclude all the images which was marked by the "face_score" metric as indicative of an unknown gender.

Stage four: At this point, and once the filtering process is finalized and all the images which does not serve the task in hand was excluded, this subsequent stage necessitates importing all the eligible curated images.

Stage five: This juncture entails utilizing the haar-cascade technique which is nothing but a method that detects specific object within an image (in this case; faces), so as to detect and pinpoint non-facial images in addition to the images by which contains multiple faces as a means to exclude them because these kind of images serves no good for the task in hand.

Stage six: Within this stage, any images which was misclassified will be eliminated, to put it simpler, instances where female image was inaccurately labeled as male, or vice versa.

Stage seven: Prior to feeding the images into the network, all images will underwent resizing to a uniform dimension of (299,299) pixels. This optimization serves to expedite training processes, decrease the hardware demands, and proven to be beneficial when dealing with extensive datasets such as IMDB-WIKI.

Stage eight: In an effort to stabilize the training process, expediting convergence speed, and reinforcing the model's generalization, this stage is devoted to normalizing the dataset by

scaling the input data to be within the scope of '0' and '1' throughout both training and testing phases.

Stage nine: At this stage, the data will be splitting into two sections: one for training, enclosing 80% of the data, and the other section for testing, comprising 20% of the data. The splitting facilitates the assessment of the model's performance on unseen data.

Stage ten: This stage necessitates incorporating data augmentation techniques so as to improve the model's capacity to learn intricate features, consequently alleviating the risks of overfitting.

Algorithm (3.1): Preprocessing Stages
Input: IMDB-WIKI Dataset
Output: Pre-Processed Data
Start **Level 1:** Launching the process by reading the IMDB-WIKI dataset. **Level 2:** Eliminating images which hold no facial features. **Level 3:** Excluding images where determination of gender as male or female was unattainable. **Level 4:** Importing the vetted images. **Level 5:** Applying the Haar-Cascade technique to exclude images which don't showcase even a single face as well as images that have presence of several faces. **Level 6:** Manually omitting the images which was observed as misclassified and/or rectifying the mislabeled images. **Level 7:** Standardizing all the images to dimensions of (299,299). **Level 8:** Normalizing the dataset. **Level 9:** Splitting the data into two sections: one for training and the other for testing. **Level 10:** Applying data augmentation techniques on the training set. **End**

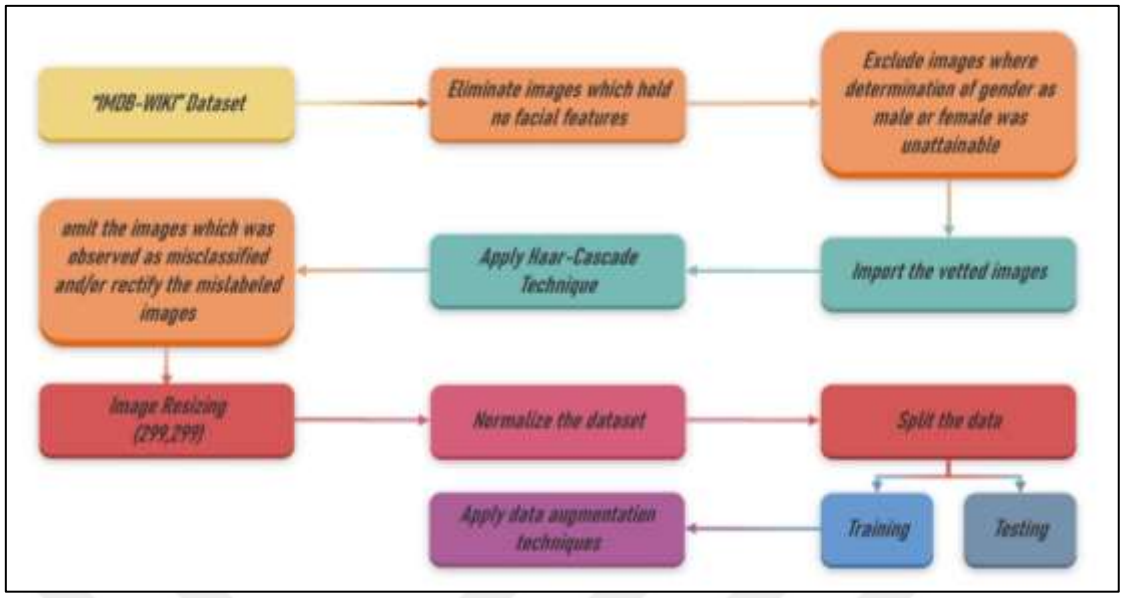

Figure 3.2: Gender Classification: Preprocessing Phase.

3.4 Haar-Cascade Technique

Although strides have been taken to grant high-resolution and crisp images, among which are, preprocessing techniques has been executed to unearth misclassified gender and to enhance images features (not to overlook that manual processing has been applied to address the issue where male images inaccurately categorized to the female class or mislabeled as a female and vice versa), on top of that, facescore has been utilized to exclude images that lacks faces within it and so forth. Yet, some images still lack clarity even after all is said and done. In light of this, with a view to overcome the potential issues arising from unclear images, haar-cascade technique has been employed to ensure high-definition images. In the Haar-cascade technique, with a view to regulate the performance and accuracy of the object detection algorithm (given this scenario, faces), two features were taken into account:

a. ScaleFactor: this parameter intention is to ascertain the extent of image size, by virtue of quantifying how much the image size decrease at each and every scale throughout the object detection process. In the context of this study, ScaleFactor was assigned a value of 1.3 and this is due to the fact that a moderate value of ScaleFactor provides both a high sensitivity to the identification of the objects and a standard processing duration thereby meeting the computational demands.

b. MinNeighbors: This parameter aims to outline the count of neighboring rectangles which supposed to pass a specific detection threshold in order to deem a particular object acceptable. Within the confines of this study, minNeighbors was assigned a value of 5, this reasonable value beneficial to the algorithm since it decreases the false positives along with expanding the prediction capabilities.

Algorithm (3.2): Haar-Cascade Technique
Input: Images
Output: Single-Face Images
Start **Level 1:** Initiating the process by utilizing pre-trained haar cascade model to identify faces. **Level 2:** Loading images from the IMDB-WIKI dataset. **Level 3:** Converting images to grayscale. **Level 4:** Setting scaleFactor to 1.3 and minNeighbors to 5. **Level 5:** Examining the detection outcomes to harness the images by which contain a face within for designated gender classification and age estimation. **End**

3.5 Data Augmentation

Data augmentation springs up as a vigorous technique in the sphere of deep learning, pursuing to amplify and enriching the training data by forming adjusted editions of the pre-existing data applying miscellaneous techniques. Acknowledging that in this research data augmentation was enacted solely on the training set. An exhaustive analysis of the executed techniques for the suggested model:

a. Rotation: With a view to enhancing the model's capability of generalizing to unseen data within images, all images were subjected to a 20-degree rotation, ultimately resulting in improving the model performance from a wholistic viewpoint.
b. Shift: In an effort to streamline the model's process in learning features in which is unaltered in minor shifts, all images undergo horizontal and vertical shifts of 20%, accordingly optimizing the model's resilience.

c. Shear: By modifying the images size with a range of 0.2, this variable facilitates learning from the distortions immanent within images.
d. Zoom: With the purpose of enhancing the model performance, a 20% zoom has been applied over all images to accommodate the features which could exists in images of diverse sizes.
e. Flip: All the images undergo a horizontal flipping for the intention of learning features which can potentially be extant solely in image reflections.
f. Fill mode: As a corollary of applying all the prior mentioned data augmentation techniques, new pixels will be generated, by which filling them requires a specific strategy, in this study the 'nearest' strategy was adopted for this aim.

Algorithm (3.3): Gender Classification: Data Augmentation
Input: IMDB-WIKI Dataset
Output: Augmented Images
Start **Level 1:** Loading IMDB-WIKI dataset. **Level 2:** Apply a rotation of 20-degree. **Level 3:** Implementing horizontal and vertical shifts of 20%. **Level 4:** Setting the images size within a range of 0.2. **Level 5:** Integrating a 20% zoom onto images. **Level 6:** Embedding horizontal flipping over all images. **Level 7:** Executing the 'nearest' strategy as a means to fill the new generated pixels resulted from the previous data augmentation techniques. **End**

3.6 GENDER CLASSIFICATION: TRAINING PHASE

This study occupied four distinct deep learning models, Inceptionv3, InceptionResNetv2, Xception, and DenseNet201, so as to develop a powerful gender recognition system. The debut step of the study implicated initializing the process by picking the desired model, each of which had been formerly trained together with the ImageNet dataset. Noteworthily, the classification part of the pre-trained model was eliminated to streamline adaptation to novel classes.

Based on predetermined settings, the model will engage with images sized at (299,299) pixels as a way for minimizing the resource usage, thereby lessening training time. On top of that, a global average pooling technique will be carried out to decrease computational demands while preserving critical information. Afterward, the architecture is further prolonged by introducing a fully-connected layer within the neural network architecture incorporating a 2000 units (neurons) dense layer in tandem with ReLU (Rectified Linear Unit) as an activation function. Moreover, a softmax activation function will be utilized to transform the concluding predictions or classifications into probability distributions for the predicted output classes, acknowledging that this task involves two classes which are “female” and “male”.

Furthermore, the study clung to a particular protocol throughout the training stages, whereat specific layers were designated for freezing, whilst the leftover layers were set as trainable layers, which grants the model the ability to adapt to the task at hand. Remarkably, Adam (Adaptive Moment Estimation) was utilized to mitigate "the loss" by which is the poor performance of the chosen models, together with a "Cross-entropy" loss function as a way to gauge the discrepancy between the model's predicted probability classes and the factual target classes.

In the long run, this study harnesses callbacks as a way to articulate and structure the results acquired from the training processes, among them tensorboard which was employed for monitoring and visualizing distinct aspects of the training stages. In addition to, checkpointer was used to store weights at specific points of the training process whereat only an improved model performance would be marked by a message illustrating that certain improvement, the determination of improvement hinge on the validation accuracy metric.

To sum up, tackling the time constraint is vital to rationalize the training duration while optimizing system performance. To accomplish this aspiration, this study harnesses the "early stopper" callback, which automatically ceases the training if no improvement is perceived over a consecutive 15 epochs. It's worth highlighting that the sum total of epochs scheduled for the training process is 20, eventually leading to amplify efficiency and augmenting overall performance.

Algorithm (3.4): Gender Classification: Training Phase
Input: Pre-Processed Data
Output: Trained Model
Start **Level 1:** Commencing the procedure by employing one of the opted models. **Level 2:** Employing global average pooling technique with the intent of reducing the spatial dimensions of the input data (height & width). **Level 3:** Constructing a fully-connected layer encompassing 2000 units (neurons) together with the ReLU (Rectified Linear Unit) serving as activation function. **Level 4:** Embedding softmax activation function so as to transform the dense layer outputs into probabilistic scores. **Level 5:** Executing the preset freezing strategy founded on the chosen model. **Level 6:** Applying Adam (Adaptive Moment Estimation) alongside cross-entropy as the loss function in which measures the distinction between the predicted outputs by the model and the actual labels. **Level 7:** Embedding the callbacks encompassing checkpointer, early_stopper, and tensorboard. **Level 8:** Designating the training process timeline to extend for the entirety of 20 epoch. **Level 9:** Evaluating the model's performance onto testing dataset and documenting all the aspects such as the accuracy, loss, precision, recall, F1-Score metrics outcomes. **End**

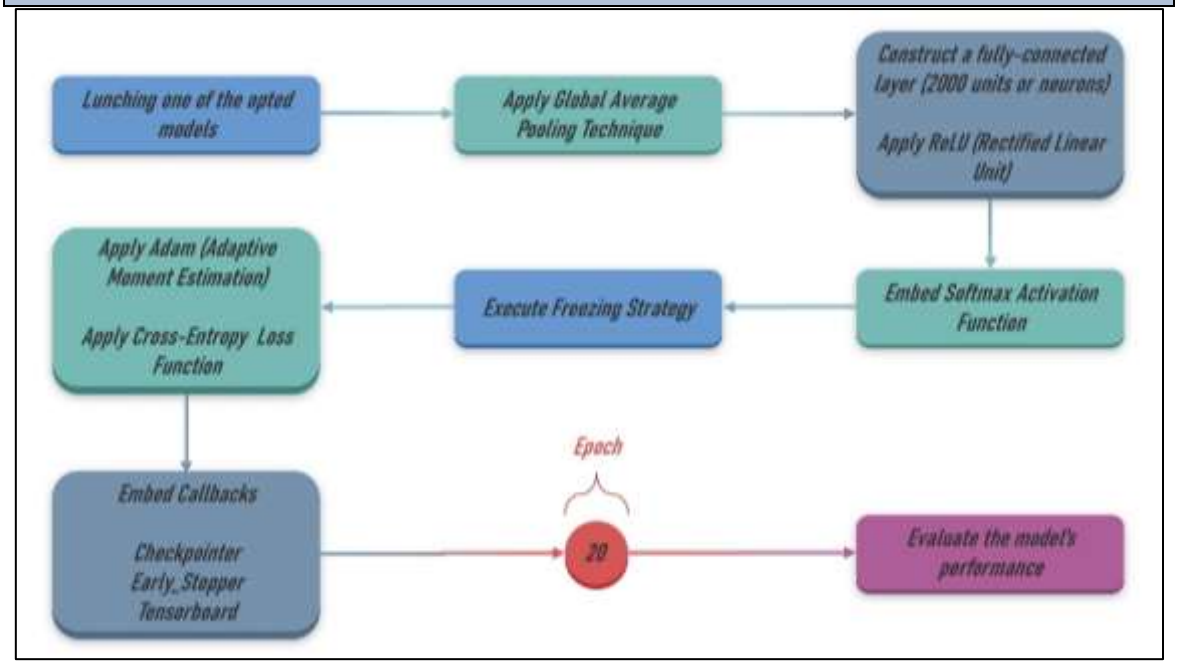

Figure 3.3: Gender Classification: Training Phase.

3.7 GENDER CLASSIFICATION: FREEZING STRATEGIES

As brought up earlier, the study adhered to a certain strategy thr oughout the training stages, freezing specific layers of the model contributes in alleviating the discrepancy between the pre-trained model and the new task data in hand, besides it mitigates the overfitting risks specifically when handling limited data.

As noted earlier, four models were employed for the purpose of developing a robust age and gender detection system so as to provide a very accurate results efficiently, accordingly there was two distinct freezing strategies was applied across these models. The first strategy was solely implemented over the inceptionv3 model, which incorporates four training levels. The initial two levels involve training only the last two layers, the distinguishing factor between the first and second level is the application of data augmentation. following that, the following duo of levels entails the training of all layers and just like previously data augmentation will be applied only in one of the levels, and the basis for that is to monitor how data augmentation could affect the accuracy of the predictions.

Moving on, InceptionResNetv2, Xception, and DenseNet201 followed a different strategy but with the same process of the distinguishing factor for all the levels. The initial duo of levels involves the freezing of the first thirty layers and training the remaining together with and without the application of data augmentation, for the succeeding couple of levels, the only difference was that this time all layers were trainable.

3.8 ENSEMBLE LEARNING

With the objective of generating extremely accurate predictions for gender and age, the study leveraged from the pluses of four models, namely, Inceptionv3, InceptionResNetv2, Xception, and DenseNet201. Needless to say, each and every one of these models have its own strengths and shortcomings. With an eye to consistently enhancing highly accurate predictions, this study offers a novel tactic for harnessing the attributes of varied models and alleviating their limitations by utilizing the ensemble learning technique as an effort to further improve the accuracy and performance of the system.

Ensemble learning is a technique entails the merging predictions emanating from numerous models as a means to foster exceedingly accurate judgments. This technique purports to alleviate noise, variance, errors, and biases that might be present within separate models by making the most of the

unified cognitive intellect of the models. The essence of this technique lies in aggregating the predictions of the opted models, these models are identified as "base model" or "weak learner" and the resultant prediction go by the name "strong learner".

In a similar vein, following acquisition of the predictions from the softmax layer out of all the engaged models, let's consider that each and every one of the models predicted that a specific image has the presence of a male at a rates of ,30,60,50, and 40%, however, It indicated that the presence of female in that certain image is at the rates of 70,60,90, and 80%, the ensemble technique will calculate the average of these predictions and highlight the highest probability, signifying the 'female' for the aforementioned instance. Below, the formulas elucidate the determination of class probabilities resting on ensemble learning:

$$P_{Female} = \frac{P_{female(Inceptionv3)} + P_{female(InceptionResNetv2)} + P_{female(Xception)} + P_{female(DenseNet201)}}{Overall\ tally\ of\ opted\ models} \quad (3.1)$$

$$P_{Male} = \frac{P_{male(Inceptionv3)} + P_{male(InceptionResNetv2)} + P_{male(Xception)} + P_{male(DenseNet201)}}{Overall\ tally\ of\ opted\ models} \quad (3.2)$$

$$Class:\ If\ (P_{Female} > P_{Male}) == P_{Female}\ ,\ \text{Else} == P_{Male} \quad (3.3)$$

Algorithm (3.5): Gender Classification: Ensemble Learning
Input: Models Predictions
Output: Gender Class
Start **Level 1:** Triggering the process through the uploading the predictions of the opted models. **Level 2:** Separating the models predictions by gender. **Level 3:** Organizing the models predictions in descending fashion founding upon the accuracy rate. **Level 4:** Calculating the average probability with respect to each gender individually, bringing about only one prediction per gender. **Level 5:** Opting the gender class by dint of outlining the prediction with the higher probability. **End**

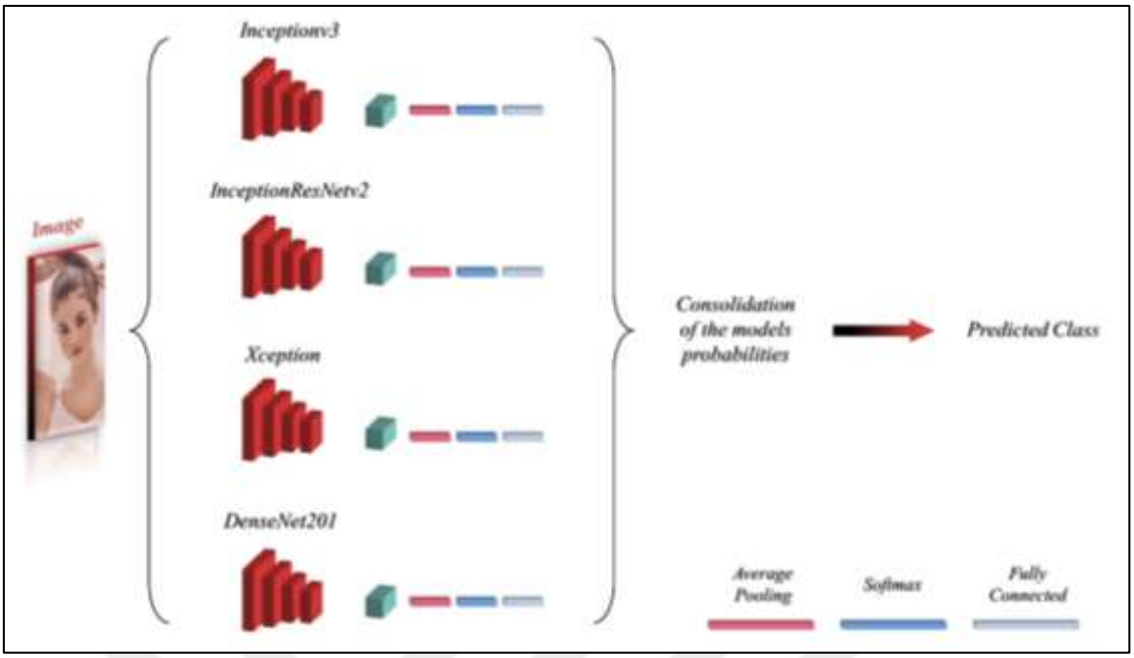

Figure 3.4: Ensemble Model.

3.9 AGE ESTIMATION: PREPROCESSING PHASE

Stage one: This study hinges greatly on the UTKFace dataset for accurate and unbiased age estimations, since this dataset backing a meaningful purpose by providing extensive age distribution which ranges from 1 to 116 years old.

Stage two: So far, and after the opted dataset (UTKFace dataset) is determined, the features must be extracted from the database including file names and age information. Notably, it does involve other information about the images such as gender, ethnicity, in addition to the date and the time of the images, but since this information doesn't help the task in hand, it was overlooked.

Stage three: At this point, the age span (which is between 1 and 116 year) will be categorized into 11 distinct labels.

Stage four: This stage implicates harnessing the one-hot encoding technique so as to convert the categorical data into a numerical format, by which allows to perform the classification task seamlessly.

Stage five: During this stage, the images will be resized to a unified format which is (299,299) in an effort to accelerate the training process.

Stage six: This stage encompasses normalization of the dataset as a means to hasten the speed of convergence together with stabilizing the training process.

Stage seven: This juncture is dedicated to split the data into two segments, the first one is for training (comprising 80% of the data) and the second one is testing (incorporating 20% of the data).

Stage eight: Finally, data augmentation will be applied in order to empower the model to learn complex features which leads to enhance the model's performance.

Algorithm (3.6): Age Estimation: Preprocessing Stages
Input: UTKFace Dataset
Output: Pre-Processed Data
Start **Level 1:** Initiating the process with reading the UTKFace dataset. **Level 2:** Extracting file names and age information within the dataset. **Level 3:** Categorizing the age span in the dataset into 11 distinct labels. **Level 4:** Applying one-hot encoding technique. **Level 5:** Standardizing all the images to dimensions of (299,299). **Level 6:** Normalizing the dataset through dividing the pixels values by 255. **Level 7:** Splitting the data into two segments: one for training purposes and the other for testing. **Level 8:** Applying data augmentation techniques only on the training set. **End**

3.9.1 Age Categorisation

Age categorization signifies the process of classifying individuals into pre-established age groups by analyzing visual data. Each one of these groups encompass distinct age ranges, accordingly the entered age will be analyzed based on the extracted features from images and assigned to the category that indicates that distinct age range. Age categorization is momentous as it facilitates the procedure of age estimation by breaking down the age range (in this context, ranging from 1 to 116) into miniature categories. Dealing with small categories instead of one giant group eases the estimation process and offers more accurate and efficient results.

The present study's algorithm involves receiving an input from the user that indicates the individual's age, thus performing the analysis so as to categorize it to a certain category. This algorithm encompasses 11 categories, each of which have a certain age range, and the entered age will be assigned to the category within which it falls.

Algorithm (3.7): Age Categorisation
Input: Age
Output: Age Category
Start **Level 1:** If the inputted age falls within the bounds 1 and 2, place the age in category 1. **Level 2:** If the entered age falls within the range 3 and 9, classify the age in category 2. **Level 3:** If the registered age falls within the bounds 10 and 20, allocate the age in category 3. **Level 4:** If the noted age falls within the range 21 and 25, position the age into category 4. **Level 5:** If the inputted age falls within the bounds 26 and 27, sort the age to category 5. **Level 6:** If the entered age falls within the range 28 and 31, situate the age under category 6. **Level 7:** If the stated age falls within the bounds 32 and 36, slot the age within category 7. **Level 8:** If the inputted age falls within the range 37 and 45, assign the age within category 8. **Level 9:** If the stated age falls within the bounds 46 and 54, classify the age under category 9. **Level 10:** If the inputted age falls within the range 55 and 65, allocate the age into category 10. **Level 11:** If the entered age exceeds 65, classify the age to category 11. **End**

3.10 AGE ESTIMATION: TRAINING PHASE

The study retained a symmetrical procedure for age estimation just as it did for gender classification, by means of the endorsed models, images resizing, incorporating global average pooling, constructing the CNN architecture in company with rectified linear unit, and wrapping up with softmax activation function. Much the same as the procedure for gender classification, Adam optimization and cross-entropy function were put to use so as to assess the models performance, the lone extra metric for the task of age estimation is MAE (Mean Absolute Error) which is particularly designed to quantify the variance between the predicted and actual values within a dataset.

With the objective of obstructing overfitting and while preserving the best performance of the models, several parameters were implemented, opening with early stopping that oversee the stipulated count of epochs (in this specific case, 10 epochs) so as to terminate training in case of no improvement is observed as per the designated metric (in this specific instance, validation loss). Afterwards, the 'ReduceLROnPlateau' callback is applied as a means to expedites model convergence and attaining the lowest minimum point during training, in this specific case the algorithm stipulates that in event of no improvement was detected for 3 successive epochs, learning rate will be reduced by 0.5 automatically.

To bring everything together, in order to train robust models with superior performance, iterations of training was set to be 100 epoch through all the training process, resulting in attaining the best performance for the models meanwhile addressing the time consumption concern.

Algorithm (3.8): Training Phase
Input: Pre-Processed Data
Output: Trained Model
Start **Level 1:** Embarking on the process by utilizing one of the opted models. **Level 2:** Applying global average pooling technique with the objective of minimizing the spatial dimensions of the input data (height & width). **Level 3:** Establishing a fully-connected layer encompassing 2000 units (neurons) in company with the ReLU (Rectified Linear Unit) serving as activation function. **Level 4:** Integrating softmax activation function in order to transform the dense layer outputs into probabilistic scores. **Level 5:** Applying Adam (Adaptive Moment Estimation) together with cross-entropy as the loss function in which gauges the distinction between the predicted outputs by the model and the actual labels. **Level 6:** Implementing MAE as the loss function so as to render the learning problem as an optimization task. **Level 7:** Ceasing the training process ahead of schedule if validation loss shows no signs of improvement for 10 consecutive epochs. **Level 8:** Employing ReduceLROnPlateau as means to decrease the learning rate by 0.5 in the event where the validation loss metric stops improving for 3 successive epochs.

Level 9: Imbedding the checkpointer callback to monitor the validation loss metric for the sake of saving only the best performance for the model.

Level 10: Setting out the training process timeline to sustain for the entirety of 100 epoch.

Level 11: Evaluating the model's performance onto testing dataset and documenting all the aspects such as the accuracy and MAE metrics outcomes.

End

4. EXPERIMENTAL OUTCOMES

4.1 INTRODUCTION

Within this scholarly inquiry, this chapter works as watershed moment where theoretical propositions are endorsed by empirical evidence. Through thorough scrutiny and elucidation of experimental outcomes, this chapter aspire to certify hypotheses, unveil insights, and exhibit discoveries arisen from a fastidiously designed series of experiments. Inaugurating the chapter by displaying the specifications of the hardware and software, expounding of the utilized datasets, and analysis of the outcomes so as to provide a rigorous examination of key findings and their implications.

4.2 TECHNICAL SPECIFICATIONS

The PC hosting the system's algorithms coheres with the specifications enumerated in table 4.1.

Table 4.1: PC Specifications.

PC Specifications	
Processor	12th Gen Intel(R) Core(TM) i7-12700H 2.70 GHz
Memory	16384MB RAM
System Type	64-bit
GPU	NVIDIA GeForce RTX 3070 Laptop GPU
Hard Disk	SSD 447 GB

4.3 DATASET SYNOPSIS

The study revolved around two primary objectives: gender classification and age estimation, each of which benefited from a distinct dataset. These datasets serve as the mainstay for the training phase since it provides an extensive metadata to facilitate model training.

4.3.1 IMDB-WIKI DATASET

IMDB-WIKI asserts itself as a wide-ranging and robust dataset, providing abundant resources for the development and evaluation of gender and age recognition algorithms

[52,53]. IMDB-WIKI dataset provides extensive variety of images and metadata reaches in excess of half a million images making it an ideal challenge to cope with so as to demonstrate the durability of the suggested model in the matter of coping with assorted data including low-quality images, sketch images, severe facial mask, full body images, multi-person images and blank images and so on. In total, IMDB-WIKI dataset comprises 523,051 images separated into two sections, first off is IMDB (Internet Movie Database) containing within 460,723 image of celebrities and secondly WIKI(Wikipedia) entailing 62,328 images from Wikipedia.

Notably, the authors acknowledge the inherent shortcomings of the dataset, as they proclaim the likely inaccuracies among the present annotations of gender and age labels. Even so, IMDB-WIKI dataset is perfectly suited for age and gender detection tasks as well as it doesn't entirely dismiss it for training and evaluation purposes, actually it underscores the importance of utilizing the dataset while taking into account these constraints. As a matter of fact, the meticulous examination of its limitations contributes toward the creation of models by which perform competently in real-world applications.

Figure 4.1: IMDB-WIKI Dataset Samples.

4.3.2 UTKFACE DATASET

The UTKFace dataset is notable as an extensive collection of facial images widely employed in computer vision area, especially utilized for tasks concentrated on gender and age recognition [54]. Encompassing myriad of multifarious images acquired from a variety of platforms embracing the internet and public databases, all the images in the dataset is annotated with labels signifying the individual's age, gender, and ethnicity. The manifold diversity of the UTKFace dataset empowers researchers to develop, train and evaluate machine learning models across a broad scope of facial features, embracing various age categories, genders, and ethnicities. UTKFace dataset enfolds tens of thousands of images,

each and every image was captured under disparate lighting conditions, poses, and facial expressions, this variability enables researchers to develop robust machine learning models and/or algorithms capable of generalizing effectively to real-world scenarios. In spite of its utility, UTKFace dataset wrestles with a set of hurdles including, the estimation of an individual's age from facial images can be inherently subjective and susceptible to errors, notably when coping with ambiguous age appearances or variations in facial aging patterns across varied demographics. Nevertheless, it remained luminous in spite of the impediments it encountered, as UTKFace dataset has garnered extensive adoption across diverse computer vision tasks, including gender classification, age estimation, facial recognition, and demographic analysis.

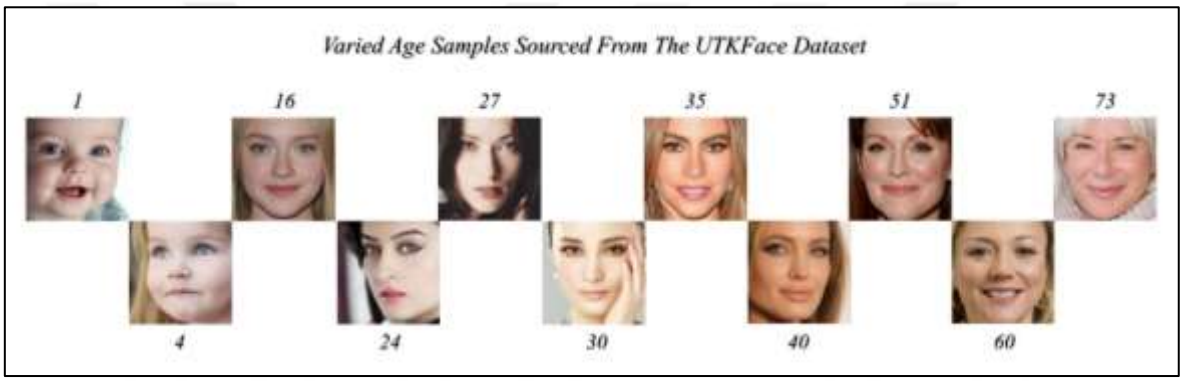

Figure 4.2: UTKFace Dataset Samples.

4.4 FINDINGS ANALYSIS

Pursuant to the previously delineated methodology of this study, four models have been utilized with the intention of developing a reliable gender and age recognition system so as to explore the impact of employing various models over the experimental outcomes. Noteworthily, the models selection process was built upon thoughtful analysis, not random selection, each one of the models was deliberately chosen due to its distinctive features that make it stand out among other viable options. The subsequent sections provide a detailed demonstration to the experimental findings of each model for gender and age recognition individually.

4.4.1 Findings from Gender Classification Experiments

Inceptionv3 model demonstrated an astonishing performance in image classification sphere spurred its adoption for this study, Inceptionv3 eases the process of feature extraction across diverse scales and abstraction layers, meanwhile it presents impressive computational efficiency. With a view to demonstrating the implications of layer trainability and data augmentation over a model's performance, several configurations were embraced and as depicted below:

As previously stated, a designated strategy was employed to train the InceptionV3 model, entailing four phases. At the outset, the initial two phases incorporate the freezing of the last two layers as the rest undergoes training. The sole contrast between the first and second phase resides in execution of data augmentation. The model showcased an impressive accuracy of 97.03% with the application of data augmentation, while gaining a tad higher accuracy of 96.75% without it. As evidenced by the data, it’s plausible to deduce that although data augmentation sparkles as a performance catalyst, it's worth bringing up that fine-tuning particular layers could at times infrequently eclipse its benefits. The figures [4.3, 4.4] provide a glimpse to the loss and accuracy of inceptionv3 model when data augmentation is implemented versus when it's not in the case where only the last two layers are trainable.

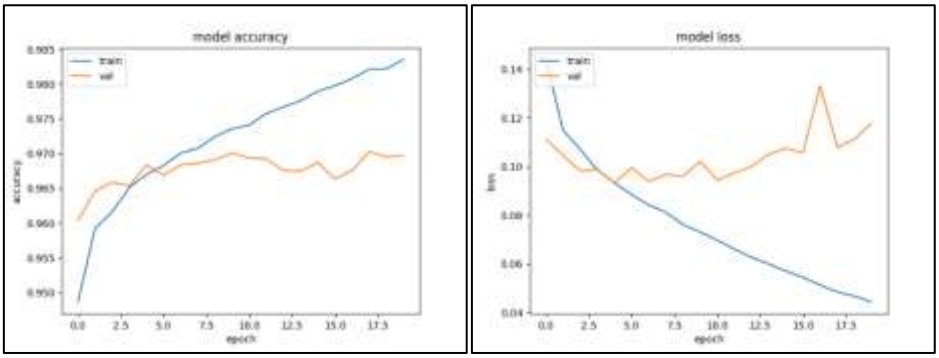

Figure 4.3: Accuracy & Loss of Inceptionv3 Model with Data Augmentation (Last 2 Layers Trainable).

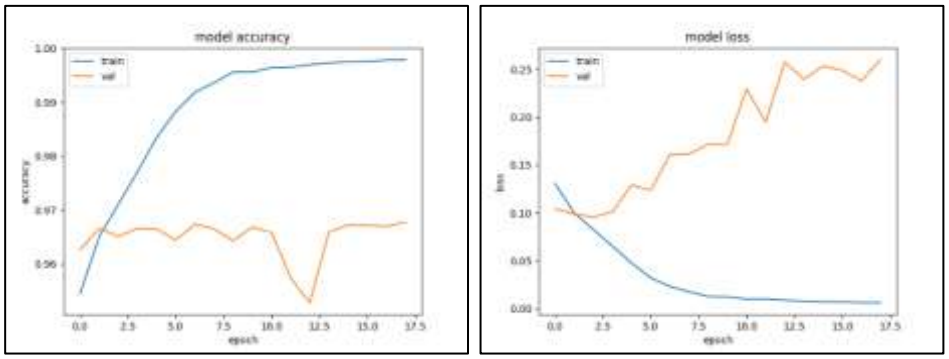

Figure 4.4: Accuracy and Loss of Inceptionv3 Model Without Data Augmentation (Last Two Layers Trainable).

When it comes to the last two phases, all layers are formed to undergo training where there is no freezing involved at all, with the same disparity as earlier concerning the decision of applying data augmentation or not. The model yielded a remarkable accuracy of 97.47% with the presence of data augmentation, while 97.18% in its absence. These results manifest that the impact of data augmentation on model performance exhibits heightened prominence when all layers are set to be trainable, this points out that the absence of it negligibly impact generalization, underscoring the crucial role of adjusting internal representations across all layers so as to effectively handle unseen data. The figures [4.5, 4.6] provide an insight into the loss and accuracy of inceptionv3 model when data augmentation is employed versus when it's not in the case where all layers are trainable.

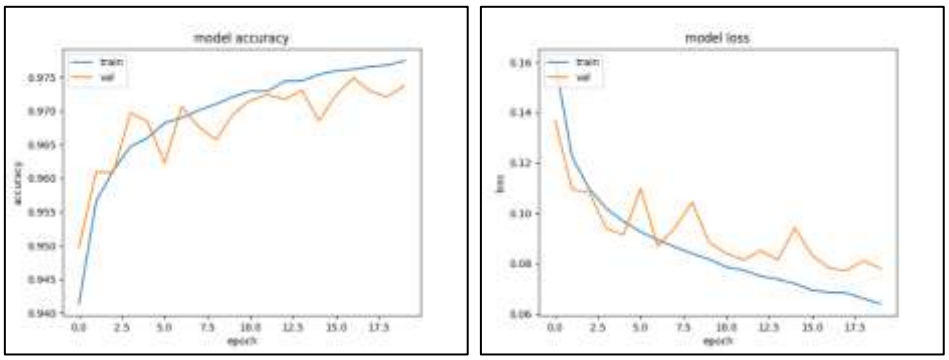

Figure 4.5: Accuracy and Loss of Inceptionv3 Model with Data Augmentation Implementation (All Layers Are Trainable).

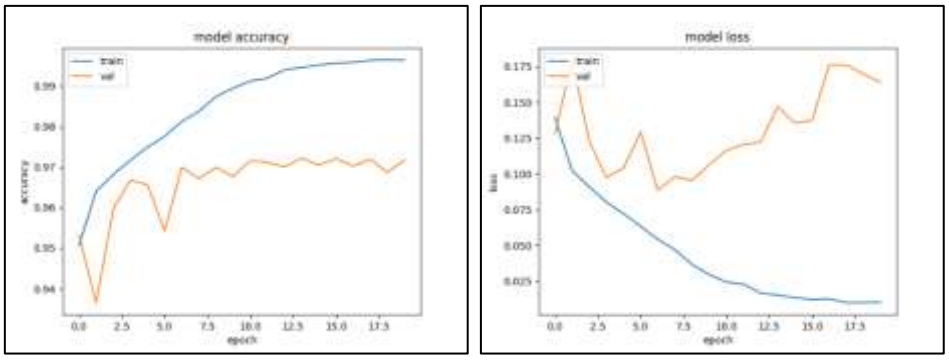

Figure 4.6: Accuracy and Loss of Inceptionv3 Model in The Absence of Data Augmentation (All Layers Are Trainable).

With regard to the remaining trio of models namely Xception, InceptionResNetv2, and DenseNet201, the study stuck to an unvarying approach. The first case depends on the freezing of the first thirty layers while the rest of the layers undergo training, and as mentioned before the first phase will be with the utilization of data augmentation while the second phase without it. In terms of the second case, it's going to follow the same old pattern when all layers are trainable and in first phase data augmentation is applied while in the next phase won't be applied.

Embarking on the analysis of the models outcomes by assessing the performance of the Xception model first, the results confirm the competence of freezing the initial thirty layers, as attested by the archived accuracy of 97.66%. In spite of what freezing status or the number of layers underwent freezing, data augmentation is persistently characterized by Xception model as a advantageous performance improvement strategy. The figures [4.7, 4.8, 4.9, 4.10] unveil the contradictory behavior of the model's loss and accuracy when accompanied by or devoid of data augmentation.

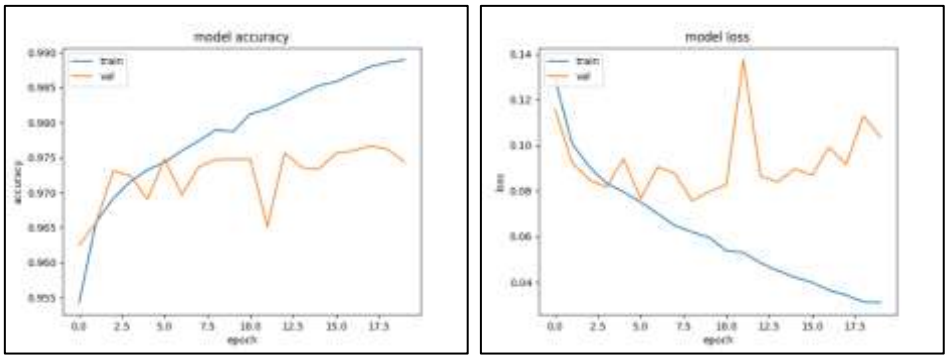

Figure 4.7: Accuracy and Loss of the Xception Model with Data Augmentation (Freezing the Opening Thirty Layers).

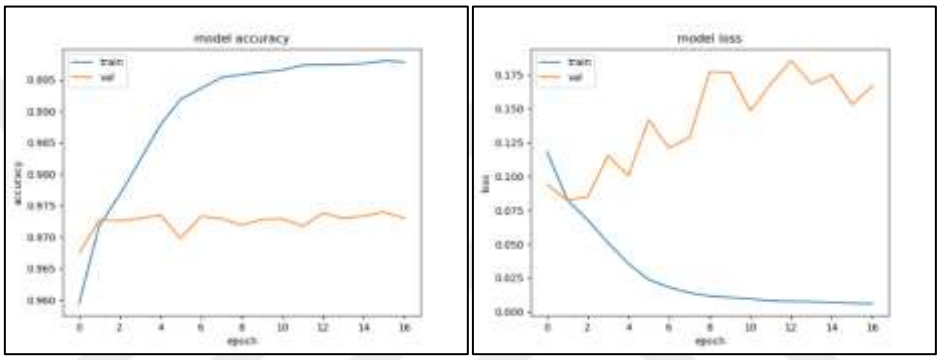

Figure 4.8: Accuracy and Loss of the Xception Model in the Case of Omitting Data Augmentation (Freezing the Initial Thirty Layers).

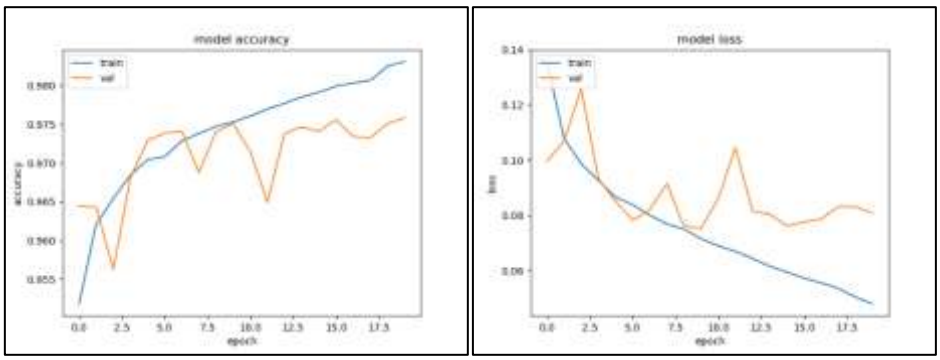

Figure 4.9: Accuracy and Loss of the Xception Model Engaging Data Augmentation (All Layers Are Trainable).

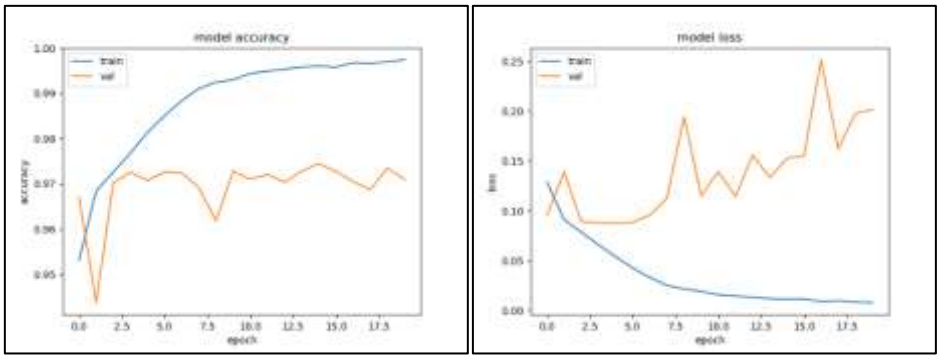

Figure 4.10: Accuracy and Loss of the Xception Model in the Case of Not Including Data Augmentation (All Layers Are Trainable).

Moreover, InceptionResNetv2 model emphasizes the previous perceptions provided by the Xception model regards the freezing status, additionally the privileges that data augmentation presents over the model performance. The analysis of validation data asserts compelling confirmation advocating for the freezing of the opening thirty layers in tandem with applying data augmentation, as evidenced by the prominent accuracy of 97.64%. Examining figures [4.11, 4.12, 4.13, 4.14] discloses a dissonant relationship between loss and accuracy over different data augmentation scenarios and alternating freezing conditions.

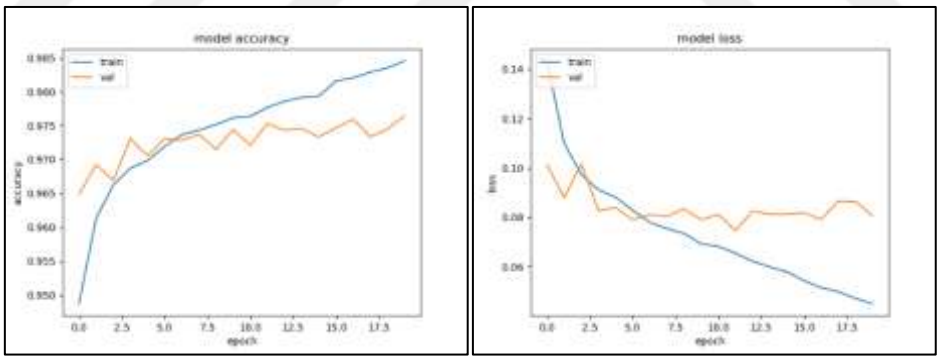

Figure 4.11: Accuracy and Loss of the InceptionResNetv2 Model Employing Data Augmentation (Freezing the First Thirty Layers).

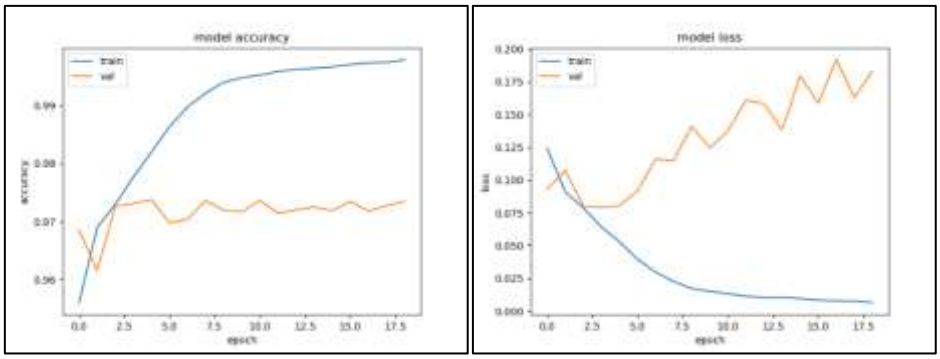

Figure 4.12: Accuracy and Loss of the InceptionResNetv2 Model Without Applying Data Augmentation (Freezing the First Thirty Layers).

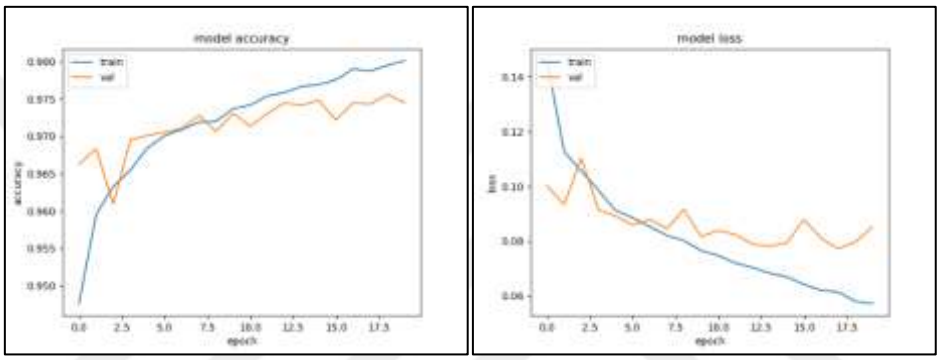

Figure 4.13: Accuracy and Loss of the InceptionResNetv2 Model Employing Data Augmentation (All Layers Are Trainable).

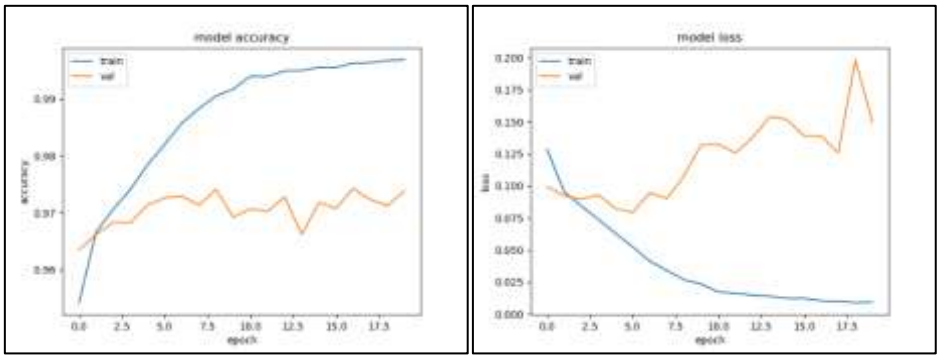

Figure 4.14: Accuracy and Loss of the InceptionResNetv2 Model Without Utilizing Data Augmentation (All Layers Are Trainable).

To bring the investigation into the experimental findings to a close, the study unpacks the outcomes emerged from the DenseNet201 model, bringing to light its outcomes. DenseNet201 accentuates the potency of data augmentation in spite of what is the applied freezing strategy, no matter whether layers are frozen or not, there's no questioning its

efficiency in enhancing accuracy and performance of the model. Noticeably, DenseNet201 affirms that freezing only the appropriate layers alongside the adoption of data augmentation helps in minimizing the training complexity, leading to achieve higher accuracy and enhance the model performance, as proven by the eminent accuracy of 97.57%. Spotlighting on figures [4.15, 4.16, 4.17, 4.18] unveils intriguing patterns in the model's performance, where model loss and accuracy shows opposing schemes over different training cases.

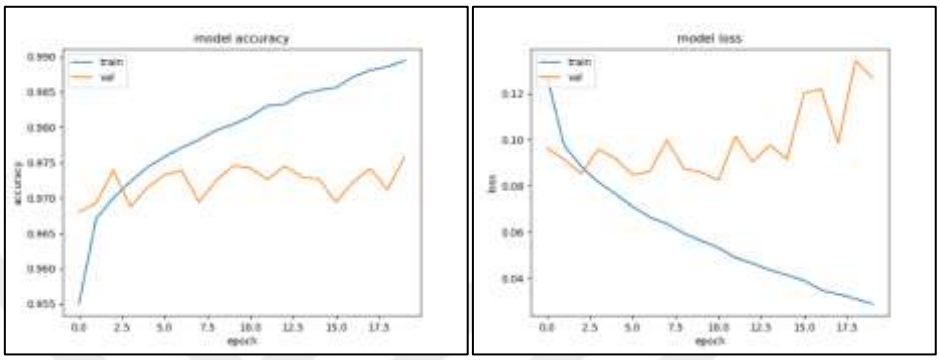

Figure 4.15: Accuracy and Loss of the DenseNet201 Model with Data Augmentation (Freezing the First Thirty Layers).

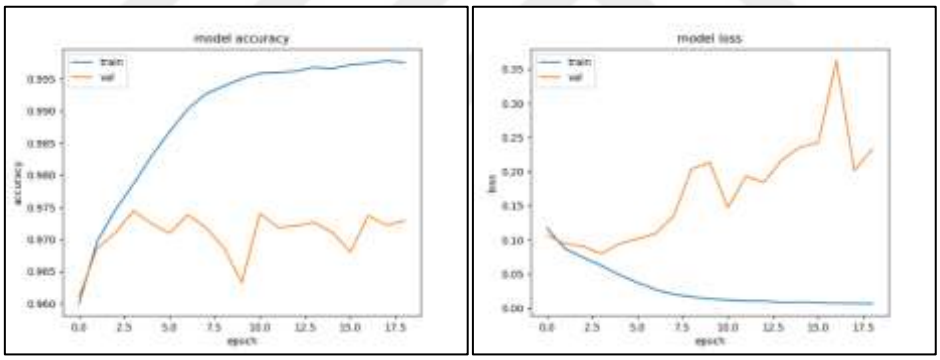

Figure 4.16: Accuracy and Loss of the DenseNet201 Model Without Applying Data Augmentation (Freezing the First Thirty Layers).

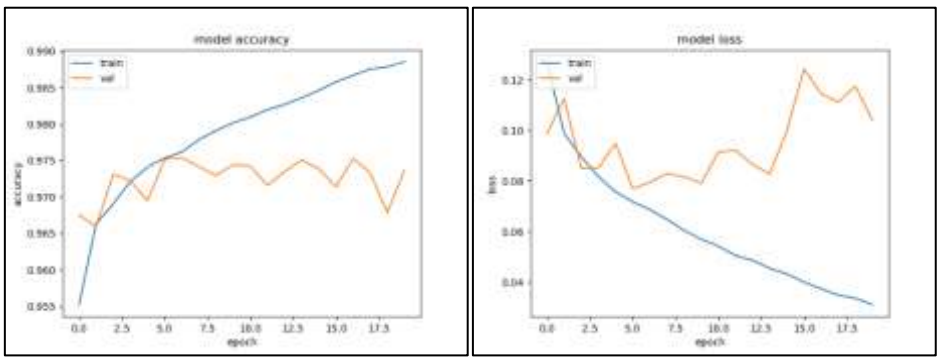

Figure 4.17: Accuracy and Loss of the DenseNet201 Model with Data Augmentation (All Layers Are Trainable).

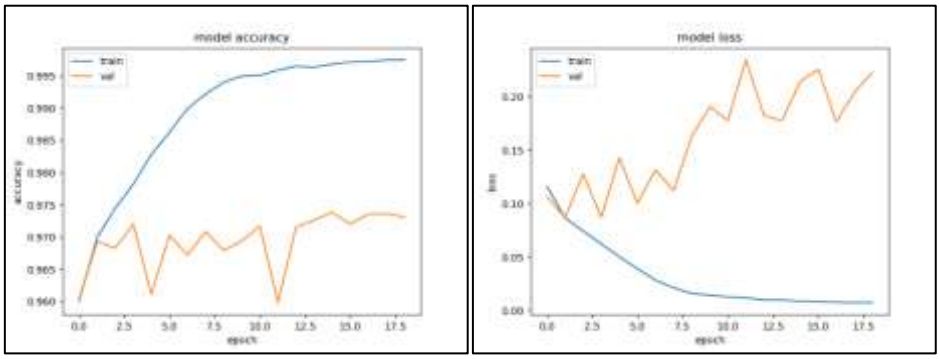

Figure 4.18: Accuracy and Loss of the DenseNet201 Model Without Applying Data Augmentation (All Layers Are Trainable).

Table 4.2: Gender Classification: Empirical Findings.

Model	Trainable Layers	Data Augmentation	Accuracy	Precision	Recall	F1 Score
Inceptionv3	All Layers	Executed	97.47%	97.53%	95.36%	96.43%
		Not Executed	97.18%	96.92%	95.14%	96.03%
	Last Two Layers	Executed	97.03%	96.01%	95.69%	95.85%
		Not Executed	96.75%	96.62%	94.24%	95.41%
Xception	All Layers	Executed	97.58%	97.3%	95.9%	96.6%
		Not Executed	97.44%	97.34%	95.47%	96.4%
	Starting From Layer 31	Executed	97.66%	97.52%	95.9%	96.71%
		Not Executed	97.39%	97.49%	95.18%	96.32%
Inceptionresnetv2	All Layers	Executed	97.55%	97.43%	95.7%	96.56%
		Not Executed	97.42%	97.26%	95.5%	96.37%
	Starting From Layer 31	Executed	97.64%	97.32%	96.06%	96.68%
		Not Executed	97.37%	96.7%	95.95%	96.32%
DenseNet201	All Layers	Executed	97.53%	96.78%	96.33%	96.55%
		Not Executed	97.38%	96.45%	96.23%	96.34%
	Starting From Layer 31	Executed	97.57%	97.46%	95.71%	96.58%
		Not Executed	97.44%	96.8%	96.03%	96.41%

Table 4.2: Gender Classification: Empirical Findings."Table Continued"

Ensemble	All Layers	Executed	97.9%	97.91%	96.21%	97.05%
		Not Executed	97.86%	97.75%	96.24%	96.99%
	Starting From Layer 31	Executed	97.97%	97.87%	96.43%	97.14%
		Not Executed	97.85%	97.74%	96.23%	96.98%

4.4.2 Findings from Age Estimation Experiments

In debates appurtenant to age estimation, the current study exploited aforementioned models assuming two contrasting scenarios as a means to scrutinize the implications of categorization over the predictive accuracy of the opted models. As a matter of course, MAE and RMSE were harnessed as an evaluation metrics for the task in hand. At the commencement, categorization was executed so as to estimate an individual's age in pursuant with the procedure delineated in chapter three and embodied in algorithm (3.7), whereas the second scenario bypassed categorization, relying exclusively on the predefined classes within the UTKface dataset. In conformity with the data presented in table (4.2) below, the findings manifest that the ensemble model outclasses in both categorized (MAE of 0.77857444 years and RMSE OF 1.23937917) and non-categorized (MAE of 5.02952341 years and RMSE of 7.91394564) scenarios, this is typical given that it combines various models' prediction so as to attain more accurate outcomes. Viewing it from a broader standpoint, utilizing categorization persistently exhibits exceptional outcomes throughout each and every one of the opted models. To put it another way, identifying particular categories facilitates the prediction process, thereby optimizing overall performance of the system. The data below indicates that the inceptionv3 model attained an MAE of 0.81358077 years with categorization whereas 5.38549135 years without. Resuming the analysis with the Xception model where it achieved an MAE of 0.81358077 years with the utilization of categorization while 5.21172501 years without it, InceptionResNetv2 yielded an MAE of 0.80725432 years with the presence of categorization and 5.29460143, finally the DenseNet201 reached 0.84542387 years with categorization and 5.3302404 years without it. The findings demonstrate that the leading determinant impacting the model's performance is whether categorization is adopted or not.

Table 4.3: Age Estimation: Empirical Findings.

Model	Evaluation Metrics	With Categorization	Without Categorization
Inceptionv3	MAE	0.81358077	5.38549135
	RMSE	1.27295457	8.39613898
Xception	MAE	0.81358077	5.21172501
	RMSE	1.26730951	8.40720805
InceptionResNetv2	MAE	0.80725432	5.29460143
	RMSE	1.27758471	8.24246394
DenseNet201	MAE	0.84542387	5.3302404
	RMSE	1.31588292	8.24388377
Ensemble Model	MAE	0.77857444	5.02952341
	RMSE	1.23937917	7.91394564

4.5 EXAMINATION AGAINST STATE-OF-ART

In pursuit of evaluating the proportional significance of this study, this section was committed to undertaking a comparative analysis between the achievements of the current study and the state-of-the-art performance results. Table 4.3 displays a comparative analysis of gender classification achieved accuracy, taking into cognizance the impact of applying diverse strategies and training datasets. Examination of the table signifies that the present study outstrips introduced state-of-the-art methods by means of accuracy. The statistics depicted in the table below indubitably highlights the supremacy of the present study over all state-of-the-art models, attaining a striking 97.97% accuracy with the ensemble model. This underlines the significance of combining predictions from varied models for the purpose of improving the accuracy.

Table 4.4: Gender Classification: Comparative Analysis.

Model / Method	Datasets	Gender Accuracy
[39] VGG-19 Model (class.)	IMDB-WIKI	91.09%
[40] Proposed CNN: CNN3	IMDB	94.46%
	WIKI	93.65%

Table 4.4: Gender Classification: Comparative Analysis "Table Continued".

[41] Proposed CNN	IMDb-WIKI & OIU-Adience	89.7%
[42] Proposed CNN	IMDb-WIKI & MORPH-II & OIU-Adience	96.2%
[43] Wide-Resnet Model	IMDB-WIKI	96.26%
[44] Proposed CNN	CelebA IMDB WIKI	96% 97% 96%
[45] RCNN Algorithm	IMDBs & Asia Image & Myanmar Image	≈ 58%
The Current Study Proposed Models		
Inceptionv3 Xception InceptionResNetv2 DenseNet201 Ensemble	IMDB-WIKI	96.75% 97.58% 97.64% 97.57% 97.97%

Yet again, the ensemble model exhibits its primacy over all existing state-of-the-art models, attaining an MAE of 97 years and an RMSE of 0.3, thereby substantiating the efficiency of the proposed system for age estimation task.

Table 4.5: Age Estimation: Comparative Analysis.

Model / Method	**Datasets**	**MAE**
[46] ResNet50 CNN	UTKFaces & APPA-REAL & BlackFaces	5.21 years
[47] ResNet50 CNN	Facial Aging Dataset (FAD)	4.25 years
[48] VGGFace	FG_NET UTKFaces	3.446 years 5.834 years
[49] Support Vector Regression (SVR)	Faceage.zip & UTKFaces	9.754 years
[50] Cross-Domain Multitask Learning (MTL)	UTKFace	5.3 ± 3.2 years
[51] Resnet-50 ResNeXt-50 (32×4d)	UTKFace	7.21 years
The Current Study Proposed Models		

Table 4.5: Age Estimation: Comparative Analysis “Table Continued”.

Inceptionv3	UTKFace	0.81358077 years
Xception		0.81358077 years
InceptionResNetv2		0.80725432 years
DenseNet201		0.84542387 years
Ensemble		0.77857444 years

5. CONCLUSION & FUTURE WORK

5.1 INTRODUCTION

Within the scope of this scholarly study, this chapter encapsulate the salient discoveries and suggested avenues for future endeavours.

5.2 CONCLUSIONS

a. Within each task investigated, the ensemble model outclassing all its counterparts, consistently manifests stellar performance by harnessing the synthesized predictions of several models.
b. With respect to age estimation, categorization has surfaced as an extremely efficient strategy. By streamlining the estimation process, it attains both efficiency and greater accuracy.
c. Within the scope of gender classification, data augmentation had negligible impact on the system's performance, seemingly by virtue of the straightforward and plain characteristics of the images in which IMDB-WIKI dataset offers.
d. Transfer learning has materialized as a potent technique for age and gender recognition by means of harnessing the pretraining models on ImageNet database in which led to accelerate the training process.
e. Data splitting served as a catalyst for many upsides encompassing permitting rapid convergence, optimal exploitation of computational resources, lessening training time and ameliorating the overall efficiency of the model.

5.3 FUTURE WORK

a. With a view to fortify the system's reliability irrespective of the circumstances. This indicates diversifying the training dataset by encompassing an extensive collection of images, including images exhibiting stringent conditions such as low lighting, individuals wearing hats or glasses, and so forth. Another avenue is to forge ahead with novel techniques that are more proficient at addressing these challenges.
b. Harnessing adaptive learning to develop a system that not only manifests adeptness at a particular juncture but also evolves and learns continuously from emerging data.

REFERENCES

[1] O. Colliot, “A non-technical introduction to machine learning,” in Machine Learning for Brain Disorders, New York, NY: Springer US, 2023, pp. 3–23.

[2] G. Rebala, A. Ravi, and S. Churiwala, “Machine Learning Definition and Basics,” in An Introduction to Machine Learning, Cham: Springer International Publishing, 2019, pp. 1–17.

[3] M. K. Benkaddour, “CNN based features extraction for age estimation and gender classification,” Informatica (Ljubl.), vol. 45, no. 5, 2021.

[4] M. Hannan, M. R. Islam, M. A. Haque, M. S. Hossain, A. Ulhaq, and J. J. Sawan, “Automated face detection, recognition and gender estimation applied to person identification,” J. Comput. Sci., vol. 15, no. 3, pp. 395–415, 2019.

[5] S. Vieira, W. H. Lopez Pinaya, and A. Mechelli, “Introduction to machine learning,” in Machine Learning, Elsevier, 2020, pp. 1–20.

[6] R. Alkhatib, W. Sahwan, A. Alkhatieb, and B. Schütt, “A brief review of machine learning algorithms in forest fires science,” Appl. Sci. (Basel), vol. 13, no. 14, p. 8275, 2023.

[7] Ssrn.com. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4458723. [Accessed: 30-Mar-2024].

[8] K. Sundar, K. Ganesh, T. Harish, S. Chaithanya, and Facial, “FACIAL AGE AND GENDER ESTIMATION USING CNN,” International Research Journal of Modernization in Engineering Technology and Science, vol. 05, pp. 2582–5208, 2023.

[9] B. Venkateswarlu, N. Sunanda, A. Y. M. Kumar, A. N. S. C. Reddy, and B. R. G. Hyndav, “AI-based Gender Identification using Facial Features,” in 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), .

[10] M. Mohtasham Moein et al., “Predictive models for concrete properties using machine learning and deep learning approaches: A review,” J. Build. Eng., vol. 63, no. 105444, p. 105444, 2023.

[11] M. T. Vi, L. T. Dat, V. T. Hoang, and T.-A. Nguyen-Thi, “Unsupervised gender prediction based on deep facial features,” in 2021 Zooming Innovation in Consumer Technologies Conference (ZINC), 2021.

[12] N. Shanthi, P. Yuvasri, S. Vaishnavi, and P. Vidhya, “Gender and age detection using deep convolutional neural networks,” in 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), 2022.

[13] M. T. Abdulhadi and A. R. Abbas, “Human action behavior recognition in still images with proposed frames selection Using transfer learning,” Int. J. Onl. Eng, vol. 19, no. 06, pp. 47–65, 2023.

[14] R. Singla and G. Singh, “Age and Gender Detection using Deep Learning,” in 2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA), 2023.

[15] P. Dönmez, “Introduction to Machine Learning, 2nd ed., by Ethem Alpaydın. Cambridge, MA: The MIT Press2010. ISBN: 978-0-262-01243-0. $54/£ 39.95 + 584 pages,” Nat. Lang. Eng., vol. 19, no. 2, pp. 285–288, 2013.

[16] M. JayaSree and L. K. Rao, “A deep insight into deep learning architectures, algorithms and applications,” in 2022 International Conference on Electronics and Renewable Systems (ICEARS), 2022.

[17] Researchgate.net. [Online]. Available: https://www.researchgate.net/publication/378435794_The_Role_of_Deep_Learning_in_Computer_Vision. [Accessed: 30-Mar-2024].

[18] K. O’Shea and R. Nash, “An Introduction to Convolutional Neural Networks,” arXiv [cs.NE], 2015.

[19] M. Krichen, “Convolutional neural networks: A survey,” Computers, vol. 12, no. 8, p. 151, 2023.

[20] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, “Transfer learning: a friendly introduction,” J. Big Data, vol. 9, no. 1, 2022.

[21] F. Raza, “Transfer learning in deep neural networks.” Unpublished, 2023.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” arXiv [cs.CV], 2015.

[23] C. Lin, L. Li, W. Luo, K. C. P. Wang, and J. Guo, “Transfer learning based traffic sign recognition using Inception-v3 model,” Period. Polytech. Transp. Eng., vol. 47, no. 3, pp. 242–250, 2018.

[24] A. E. Minarno, L. Aripa, Y. Azhar, and Y. Munarko, “Classification of malaria cell image using Inception-V3 architecture,” JOIV Int. J. Inform. Vis., vol. 7, no. 2, p. 273, 2023.

[25] A. R. Muslikh, D. R. I. M. Setiadi, and A. A. Ojugo, “Rice disease recognition using transfer learning Xception convolutional Neural Network,” J. Tek. Inform. (JUTIF), vol. 4, no. 6, pp. 1535–1540, 2023.

[26] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[27] R. Rismiyati and A. Luthfiarta, “Transfer learning with Xception architecture for snakefruit quality classification,” JAIS (J. Appl. Intell. Syst.), vol. 7, no. 2, pp. 162–171, 2022.

[28] A. Mehmood, Y. Gulzar, Q. M. Ilyas, A. Jabbari, M. Ahmad, and S. Iqbal, “SBXception: A shallower and broader Xception architecture for efficient classification of skin lesions,” Cancers (Basel), vol. 15, no. 14, 2023.

[29] J. Rashid, B. S. Qaisar, M. Faheem, A. Akram, R. ul Amin, and M. Hamid, “Mouth and oral disease classification using InceptionResNetV2 method,” Multimed. Tools Appl., vol. 83, no. 11, pp. 33903–33921, 2023.

[30] T. S. Azzahra, J. J. Cerelia, F. A. L. Nugraha, and A. A. Pravitasari, “MRI-based brain tumor classification using Inception Resnet V2,” Enthusiastic : International Journal of Applied Statistics and Data Science, pp. 163–175, 2023.

[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the impact of residual connections on learning,” arXiv [cs.CV], 2016.

[32] M. J. Akhtar et al., “A robust framework for object detection in a traffic surveillance system,” Electronics (Basel), vol. 11, no. 21, p. 3425, 2022.

[33] R. Mahum et al., “A novel framework for potato leaf disease detection using an efficient deep learning model,” Hum. Ecol. Risk Assess., vol. 29, no. 2, pp. 303–326, 2023.

[34] J.-F. Yeh, K.-M. Lin, C.-C. Chang, and T.-H. Wang, “Expression recognition of multiple faces using a convolution neural network combining the Haar cascade classifier,” Appl. Sci. (Basel), vol. 13, no. 23, p. 12737, 2023.

[35] L. Rasheed, U. Khadam, S. Majeed, S. Ramzan, M. S. Bashir, and M. M. Iqbal, “Face recognition emotions detection using Haar Cascade classifier and convolutional neural network,” Research Square, 2022.

[36] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2005.

[37] C. Goutte and E. Gaussier, “A probabilistic interpretation of precision, recall and F-score, with implication for evaluation,” in Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 345–359.

[38] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature,” Geosci. Model Dev., vol. 7, no. 3, pp. 1247–1250, 2014.

[39] C. H. Nga, K.-T. Nguyen, N. C. Tran, and J.-C. Wang, “Transfer learning for gender and age prediction,” in 2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan), 2020.

[40] M. K. Benkaddour, S. Lahlali, and M. Trabelsi, “Human age and gender classification using convolutional neural network,” in 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH), 2021.

[41] A. Singh, N. Rai, P. Sharma, P. Nagrath, and R. Jain, “Age, Gender Prediction and Emotion recognition using Convolutional Neural Network,” SSRN Electron. J., 2021.

[42] O. Agbo-Ajala and S. Viriri, “Age group and gender classification of unconstrained faces,” in Advances in Visual Computing, Cham: Springer International Publishing, 2019, pp. 418–429.

[43] O. Agbo-Ajala and S. Viriri, “Face-based age and gender classification using deep learning model,” in Image and Video Technology, Cham: Springer International Publishing, 2020, pp. 125–137.

[44] F. H. Kamaru Zaman, “Gender classification using custom convolutional neural networks architecture,” Int. J. Electr. Comput. Eng. (IJECE), vol. 10, no. 6, p. 5758, 2020.

[45] K. S. Htet and M. Myint Sein, “Effective marketing analysis on gender and age classification with hyperparameter tuning,” in 2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech), 2020.

[46] O. V. Abhulimen and O. Erastus, “Automatic Age Estimation of Persons with Dark Skin tone using Deep Learning Approach,” Int. J. Comput. Digit. Syst., vol. 12, no. 4, pp. 1184–1189, 2022.

[47] “Cross-racial automatic age estimation from facial images using deep learning,” Int. J. Emerg. Trends Eng. Res., vol. 9, no. 9, pp. 1288–1294, 2021.

[48] A. Shannaq and L. Elrefaei, “Age estimation using specific domain transfer learning,” Jordanian J. Comput. Inf. Technol., no. 0, p. 1, 2020.

[49] J. S. Amelia and Wahyono, “Age estimation on human face image using support vector regression and texture-based features,” Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 12, 2022.

[50] S. M. Bafti, S. Chatzidimitriadis, and K. Sirlantzis, “Cross-domain multitask model for head detection and facial attribute estimation,” IEEE Access, vol. 10, pp. 54703–54712, 2022.

[51] A. Fariza, Mu'arifin, and A. Z. Arifin, "Age estimation system using deep residual network classification method," in 2019 International Electronics Symposium (IES), 2019.

[52] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," Int. J. Comput. Vis., vol. 126, no. 2–4, pp. 144–157, 2018.

[53] R. Rothe, R. Timofte, and L. Van Gool, "DEX: Deep EXpectation of apparent age from a single image," in 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), 2015.

[54] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.