AGE ESTIMATION FROM FACE IMAGE LEVERAGING CONCATENATED FEATURES OF VISION TRANSFORMER ALONG WITH RESNET-50

Thesis Submitted to the Department of CSE in Partial Fulfillment of the Requirements for the Degree of B.Sc. Engineering in CSE

by

Ishra Naznin

Student ID: 18701069

Session: 2017-18

Under the Supervision

of

Dr. Md. Mahbubul Islam

Associate Professor

Computer Science and Engineering,

University of Chittagong.

1 August, 2023

Report Code:

University of Chittagong

Department of Computer Science and Engineering

8th Semester B.Sc. Engineering Examination 2021

Course No.: CSE 800

Title: Age Estimation from Face Image Leveraging Concatenated Features of Vision Transformer along with ResNet-50

Report Code:

University of Chittagong

Department of Computer Science and Engineering

8th Semester B.Sc. Engineering Examination 2021

Course No.: CSE 800

Title: Age Estimation from Face Image Leveraging Concatenated Features of Vision Transformer along with ResNet-50

Student Name: Ishra Naznin
Student ID: 18701069
Session: 2017-2018
Hall: Pritilata
Signature of Student:

_____

Submission Date: 01 Augest, 2023

Approval of Submission

This thesis titled 'Age Estimation from Face Image Leveraging Concatenated Features from Vision Transformer along with ResNet-50', by Ishra Naznin, Student ID: 18701069, Session: 2017-18 , has been approved for submission to the Department of Computer Science and Engineering, University of Chittagong, in partial fulfillment of the requirements for the 8th Semester B.Sc. Engineering Examination 2021.

Dr. Md. Mahbubul Islam

Associate Professor

Department of Computer Science and Engineering

University of Chittagong

Chittagong, Bangladesh.

ABSTRACT

Age is a significant biometric attribute that is used for many applications such as recommendation, access control, criminal detection, and so on. The human face is a vast repository of individuality and the aging upshot is easily palpable. A variety of facial features indicators and looks have built up complexity in the existing age estimation system. In recent years, deep learning techniques have shown promising results in this domain. Specially, deep residual networks (ResNets) are a key component of face-based age estimation. However, ResNet based approaches downplay the significance of some extensive face data and other facial age features. Besides, deep residual networks and Vision Transformer (ViT) models are often employed for face feature identification tasks, but their joint efficacy and contrast to estimate human chronological age somewhat have not been fully investigated. To address these challenges of age estimation from human face, this work proposes a novel hybrid model combining the strengths of two state-of-the-art architectures: ResNet50 and Vision Transformer(ViT). The hybrid model uses the ResNet50's and the ViT's features, capturing local fine-grained low-level features from ResNet50 and global contextual information from Vision Transformer, thus, enhancing robustness for age estimation from human face. Besides, extensive experiments show that the hybrid model outperforms the standalone ResNet50 and the Vision Transformer models in human age estimation, achieving state-of-the-art accuracy. In addition, this work shows the effectiveness of multi-features fusion from face image such as HOG-ResNet50-ViT features to estimate age. However, the proposed approach use the UKTFace data-set with 23708 images for training, validation and testing with splitting the dataset into 80 percent for training, 20 percent validation and 20 percent for testing. The propose approach finds MAE value of 4.88 by putting out age estimation as a regression problem in this work. The suggested hybrid model has enormous promise for practical applications, advancing human-robot interaction, biometrics, and other related domains.

Keywords: *Age Estimation, Face Image, Vision Transformer(ViT), ResNet-50, Deep Learning, Computer Vision, Machine Learning, Features Fusion.*

Acknowledgements:

This is to certify that the thesis entitled **Age Estimation from Face Image Leveraging Concatenated Features of Vision Transformer along with ResNet-50**, submitted in full result.

I am grateful to my supervisor, **Dr. Md. Mahbubul Islam**, Associate Professor of Department of Computer Science and Engineering, University of Chittagong, for his kind words of support, encouragement and his valuable time.

I wish to extend my profound sense of gratitude to **my parents** for all the sacrifices they made during my work and also providing me with moral support and encouragement whenever required.

Ishra Naznin

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

This section describes the background, motivation and problem statement of age detection.

1.1 Background:

The inexorable advancement of tools and technology that make our lives simpler is what makes up today's evolving world. The way individuals spend their lives has been completely transformed by these technological innovations. Technology now permeates every aspect of our life, from communication to business. Nowadays, practically every element of a person's everyday life involves some kind of technical instrument. In addition, humans interact with their surroundings on a regular basis by using various bodily parts and sensory organs. With the use of their human eyesight, utilizing their eyes, a person may swiftly analyze the scenario they are in by doing so. The necessity for computers to have a comparable ability of artificial vision that can resemble human vision as technology continues to grow at an astounding rate, however, emerges. Here is when computer vision is useful. Their age, gender, and other characteristics may be determined using the information which they takes to connect others. Moreover, online banking, shopping for goods and services, marketing, and many other functions have all become more dependent on the Internet. A lot of digital data has been produced as a result of the widespread sharing of private information like images and videos.It is now feasible to get useful insights about people from the massive amount of digital data accessible. This contains methods for facial recognition, which may be used to determine an individual's age, gender, and other demographic characteristics. Advertising and targeted marketing are only a couple of the uses for this data. But there are security and privacy issues that need to be addressed that are brought up by the gathering and use of such data.As humans, we engage with our surroundings via our bodily senses, such as vision, touch, and hearing. We can swiftly comprehend the environment around us with the aid of our vision. The idea of computer vision has developed as a means of re-creating this capability in machines. Computer vision is the process of teaching robots to interpret digital pictures and movies in a manner similar to that of humans. Bio-metrics, robotics, and medicine have all experienced significant developments as a

result of the development of computer vision technologies. Machines are now able to reliably recognize things in pictures and movies, spot anomalies in x-rays, and more. Computer vision technology is expected to become more crucial in determining our future as it develops.

1.1.1 Computer Vision:

Computer vision is a method used to let a computer perceive, determine, and do tasks like a person. It makes it possible for machines to complete AI jobs just like people. The capability of the computer to process data is the primary distinction between human vision and computer vision. We can train a machine to recognize and evaluate hundreds of items and things in under a second. Computer vision, in simple terms, is the eye of an AI system that gives the machine the ability to comprehend visual data. Computer vision analyzes pictures at the pixel level to extract useful data and patterns.However, It has an issue since it needs a lot of data to work well. This is mostly because of the way that computer vision works, which includes analyzing data continuously in order to recognize and understand the differences and specifics of different things. The main goal of the research area of computer vision is to use computer algorithms to mimic human vision. The systems need a large dataset to train and learn from, in contrast to people who can detect things and understand their context with ease. A large number of photos or videos have to be included in these datasets, coupled with labels or annotations that serve as a guide over the computer vision algorithms. Thus, it needs an explicit label of processing of the visual data with huge sized. Convolutional neural networks (CNNs), a class of deep learning techniques, are also frequently used in computer vision. CNNs are made particularly to extract pertinent characteristics from pictures using a hierarchical method. CNNs can learn to recognize patterns and features at various levels of abstraction by employing a number of convolutional filters and pooling layers. They can then identify items or find particular properties in a picture.

1.1.2 Computer vision and Machine learning:

A breakthrough technique called machine learning seeks to build intelligent computers that can observe, analyze, and learn from the past. It accomplishes this by using analytical, mathematical, statistical, and optimization approaches to train models. These models are made to find patterns in the incoming data, comprehend them, and then apply what they have learned to future tasks. Machine learning input data, however, frequently includes a variety of properties, some of which might not be pertinent to the current job. Therefore, it becomes crucial to carefully choose and concentrate on the pertinent elements that contribute to the intended outcomes during the training process. Machine learning algorithms may successfully learn and generalize patterns by

locating and combining the essential elements, producing precise predictions or judgments based on fresh, unexplored data. Artificial intelligence (AI) is a key component, and computer vision is one domain where machine learning is crucial. Training machines to comprehend and interpret visual input, such as pictures or movies, is known as computer vision. Machine learning algorithms, including supervised learning, reinforcement learning, and unsupervised learning, are used in the area of computer vision. In supervised learning, each input sample is linked to a matching desired output, allowing the computer to learn from labeled data. This refers to connecting each item or picture to the desired result in computer vision, such as a particular categorization or recognition. Computer vision models may accurately identify and classify things based on the structures they've learnt from the labeled data by being trained via supervised learning.Another machine learning approach is reinforcement learning, which includes teaching an agent to base judgments on rewards and penalties received from its surroundings. For computer vision applications like clustering or anomaly detection, where the models find concealed patterns or just groupings in imagery without previous knowledge of the expected results, unsupervised learning techniques might be effective. Computer vision systems can detect, examine, and interpret visual data more effectively by using machine learning techniques. Computer vision models can now do challenging tasks like object detection, picture segmentation, image creation, and even more sophisticated tasks like image captioning or video interpretation thanks to machine learning, which enables them to learn from data and adapt. It makes computer vision an essential part of contemporary AI systems by giving it the capacity to continually upgrade and increase its comprehension of visual input.

Computer vision mimics the human brain's capability to process, analyze an image, and detect objects. Computer vision tasks are done with the use of some algorithm. Several are::

- **Image Classification:** Figuring out the class or category of an object inside an image is a key challenge in computer vision. In order to forecast the class of an item in a new image, this assignment makes use of machine learning methods to gather knowledge from a library of visual images representing various object classes.
- **Object Detection:** An approach for computer vision called object detection makes it possible to recognize and localize various items in a picture. Object detection techniques are able to identify and distinguish between multiple items present in a picture, even if they are members of the same class, unlike image classification, resulting in a single label to the whole image. In addition, with the use of bounding boxes, object detection aims to not only identify the presence of things but also reveal their physical position. As a result, it is possible to comprehend the

image more thoroughly and execute activities like counting the number of items present, following their movement, or running further analyses on certain regions of interest. When numerous cats and dogs are present in a picture, it can tell whether one is a dog or cat.

- **Pose estimation:** A computer vision technique called pose estimation seeks to identify and evaluate the placements and orientations of several body components, usually in relation to a human body. The program can recognize and monitor a person's joints and skeletal structure, allowing for the identification of various postures or poses like sitting, lying down, or flying. Pose estimation's main objective is to deduce the spatial layout and relative locations of body components in a series of still or moving images. Pose estimation methods can precisely predict the location and orientation of joints including the head, shoulders, elbows, wrists, hips, knees, and ankles by utilizing machine learning and computer vision approaches.
- **Image Segmentation:** A computer vision method called image segmentation separates an image into several segments or areas according to its visual characteristics. Image segmentation seeks to accurately define the borders of individual items or areas inside an image, as opposed to object recognition, which recognizes and localizes things as a whole. The main goal of picture segmentation is to separate pixels with similar properties into groups while keeping the surrounding pixels separate. This method enables a more thorough and precise grasp of the image's information, enabling more precise analysis or processing in the future.
- **Face Detection:** Identifying and locating human faces in an image or video is the goal of the computer vision method known as "face detection." It is an essential stage in many applications, such as augmented reality, facial recognition, emotion analysis, and biometrics. In order to accurately pinpoint the location and boundaries of a face in a given picture or video frame, face detection must first ascertain whether the supplied frame contains a face. By examining the visual traits and patterns connected to human faces, face identification algorithms do this. In the past, face identification algorithms depended on methods like Viola-Jones, which identify areas of an image that are like to contain a face using cascaded classifiers and characteristics similar to those in Haar. In order to simplify the procedure, these algorithms often work on pre-processed or grayscale photos.Convolutional neural networks (CNNs) and other deep learning-based techniques have recently revolutionized the field of face identification. The well-known Single Shot MultiBox Detector (SSD) and the You Only Look Once (YOLO) architecture are two CNN-based models that can learn intricate features and structures directly from the unprocessed visual data. Through a variety of settings and variations in light-

ing, stances, and facial expressions, these models are able to recognise faces with great accuracy and durability.

Looking at computer vision from a broad perspective, we can state that it has a variety of uses in a number of different industries, including transportation, retail, manufacturing, healthcare, along with more. Besides, machines may be programmed to automatically assess and pick based on visual input with the use of computer vision algorithms and techniques, providing benefits including cost savings, higher accuracy, and better efficiency. Applications for computer vision include driver-less cars, video tracking, medical imaging, retail management, and industrial quality control, to name a few. Overall, computer vision has the promise of significantly enhancing many facets of our everyday lives by making them more automated, affordable, and effective.

1.2 Motivation:

The most crucial aspect of computer vision applications is the human face. One's age, gender, race, and other characteristics can be inferred from his or her face. Additionally, someone's mindset and state of mind may be inferred from their facial expressions, such as those of happiness, sorrow, rage, and so on. The same method may be used to determine age. The age estimation procedure uses a person's photograph to infer a certain pattern that may be used to determine their age. Age estimation from an old or current image is a tricky challenge for computer vision, though. Because a face might change as a result of a medical issue, cosmetic procedures, or other factors. It may be difficult to guess someone's age, for example, if they smoke, as their face may look older than they are. Additionally, a person's age has a big impact on a lot of serious applications, like age estimation for crime investigation, age-based access restrictions, age-based recommendations, social media analysis, ethnic background recognition, behavioral science, interaction between humans and computers (HCI), and more. A biometric identification method called face age estimate has developed dramatically in recent years (1) (2).Human Computer Interaction (HCI), security, and management software, forensic inquiries (3) (4) are just a few of the areas in which this technology may be used in the real world. Academic study on traditional face analysis covers a wide variety of topics, including face tracking, face identification, face verification, face detection, and 3D facial emotion recognition. Additionally, researchers have showed a lot of interest in the study of face characteristics including age estimation, gender differentiation, and ethnicity identification (5). Besides, men and women may nurture differently too. When people go from being children to adults, there is a chance that they may have a wide variety of images. The aging of the face occurs in predictable patterns that are defined by several age modes. For example, children's facial characteristics change significantly as they get older, with the expansion of the skull playing a major role. Adulthood's

aging process, in contrast, is marked by perceptible changes in face skin texture that affect how one looks physically. Deep wrinkles forming, the skin gradually peeling off over time, and the formation of spots on various parts of the face are all signs of this. These unique manifestations offer insightful information regarding the aging patterns seen at various phases of life. Because of this, it is difficult to determine a person's real age from the age represented by their face.In addition, aside from the subtleties of age and facial features, other variables including gender, genetics, ethnicity, health, and life circumstances can also have an impact on how old a person seems. Due to its intricacy and painstaking nature, estimating age from face photographs is a difficult task. It can be difficult to get accurate facial data for thorough study since there may be limitations and imbalances in the age, gender, and ethnicity representation of publically accessible age data. Age estimate research is significantly hampered by these issues. Despite these difficulties, face age estimation technology is used in a variety of applications, including information management, entertainment, surveillance and inspection systems, and more. Research into the driving forces behind this area of work is crucial in order to create accurate and trustworthy face age estimate algorithms.The value of estimating face age includes the following:

- **Public Safety and Law Enforcement:** Facial age estimate helps law enforcement locate missing children and senior citizens, supports search and rescue operations, and may even help reunite families.
- **Age-Related Services and Personalization:** The ability to estimate age enables the provision of personalized services, catered experiences, content suggestions, ads, and healthcare services.
- **Demographic Analysis and Social Sciences:** Using face inspects to estimate age provides information on aging patterns and trends that is useful for demographic analysis, population studies, and social science research.
- **Healthcare and Aging Research:** For age-based illness risk assessment, treatment planning, and geriatric care, accurate age estimation can be helpful in the medical field. A person's age may be determined by looking at their facial characteristics, which allows medical experts to assess possible health concerns that come with aging and decide how best to care for patients.Insights into aging patterns and trends are gained through age estimate from face photographs, which is helpful for demographic analysis, population studies, and social sciences research.
- **Forensic Investigations:** Estimating an individual's age can help forensic investigators determine the age of remains that are unidentified or suspects in criminal cases. In order to help law enforcement authorities narrow down prospective leads and increase the accuracy of investigations, it can give useful information.

- **Human-Computer Interaction:** Age estimation is essential to HCI because it enables systems to customize and change their answers based on the user's estimated age. It can improve user experience, help in designing user interfaces for different ages, and make it easier for people to engage naturally and with context awareness.
- **Biometrics and Identity Verification:** Age estimation mitigates fraud and strengthens security systems by comparing estimated age with claimed or known age during identity verification.

So, age estimate from face photos has several practical applications, but it's necessary to be aware of any potential drawbacks and moral dilemmas. Age estimation algorithms should be created with fairness, openness, and privacy in mind to avoid biases and to protect people's rights.In conclusion, face age estimate is necessary for many different applications, such as biometrics, customized services, public safety, healthcare, demography analysis, forensic investigations, and human-computer interaction. Its applications cut across a range of businesses and areas with the goal of enhancing security, user experiences, research, and decision-making processes.

1.3 Problem Statement:

The process of estimating age from face photographs is difficult and complex, especially when taking into account the differences in facial appearance caused by aging, ethnicity, and facial emotions. Face photos have traditionally been used by the majority of application including human-computer interaction, entertainment, automated surveillance systems. In addition to one's biometric traits (fingerprints, irises), one's face photograph is a fantastic resource for capturing their individuality. A face may reveal a person's age, gender, physical condition, level of health, emotional state, and a wealth of other details. However, information retrieval can also go the other way. Considering that sometimes we just know someone's age and gender but not their face. For instance, to hide their identities, thieves frequently wear masks while theft and other types of illegal activities. They also create several false profiles. The police utilize this information to examine the persons who match the suspects' age, gender, and/or assault location and time if they have any clue of their identity. By using this procedure, the actual offender can be apprehended more quickly. Additionally, sometimes we only need to know a person's age and not their face. such as an access control system, a system to prevent the use of unsuitable pharmaceuticals or medications, a suggestion system for music, pictures, or shopping lists, and a nutrition system. Besides, the difficult challenge of face age estimate in computer vision entails estimating an individual's age from their facial features. Due to the intrinsic variances in face aging patterns, the impact of outside elements like lighting and position, and the requirement to manage substantial and varied datasets,

it is still a challenging undertaking. The approaches for estimating age that are now available mostly depend on either deep learning or conventional machine learning algorithms. The use of features from the Histogram of Oriented Gradients with classifiers in traditional machine learning methods has shown some promise. individuals frequently struggle to capture complex face characteristics and have poor generalization skills. In addition, deep learning-based methods, like Convolutional Neural Networks (CNNs), have shown remarkable performance in computer vision tasks. Transfer learning and Vision Transformer (ViT) models have been widely used for facial feature recognition task, but their effectiveness and comparison with traditional CNN models remain unexplored. Besides, it is still necessary to increase the performance and resilience of deep learning models, even if they have produced encouraging results in age estimate. But Convolutional Neural Networks (CNNs) are increasingly growing in size and complexity , they still exhibit certain limitations. The drawbacks include the challenge of selecting the optimal network architecture, the significant computational resources required, the difficulty in transferring learned knowledge across various datasets, and the dependency on extensive training data. Addressing these weaknesses will be crucial for further enhancing the performance and practical applicability of CNN-based age estimation models. In order to improve the performance of the model and solve its current shortcomings, we propose a unique method for age estimate in this study that combines features from Vision Transformer (ViT) and ResNet-50.Thus, this study's main goal is to investigate the combined fusion of characteristics from Vision Transformer and ResNet-50 for estimating age from facial photographs. The suggested model seeks to overcome the shortcomings of separate approaches and produce more precise and reliable age forecasts by integrating the benefits of both designs. The goal of this research is to shed light on the efficacy of various approaches to estimating face age, with the ultimate goal of creating high-performance models that may be used in practical contexts.

1.4 Scope of the Research:

With the help of a hybrid model that combines the strength of ResNet50 and Vision Transformer (ViT), the goal of this research is to create a novel method for age assessment from face photographs. The main goal is to investigate the possible advantages of combining these two deep learning architectures to provide age prediction results that are more reliable and accurate. The project will analyze a dataset for estimating face age in-depth and look into appropriate preprocessing and data augmentation methods. The technique of optimizing the hybrid model's performance on age prediction tasks will also be covered in depth by the research. Comparative evaluations against state-of-the-art age estimation methods will also be done in order to confirm the validity of the proposed hybrid model. Besides, this study aims to investigate, analyze, and

compare numerous procedures, strategies, and models for reliably estimating an individual's age based on their facial features. While placing a strong emphasis on transfer learning and the use of pretrained Convolutional Neural Networks (CNNs). The study's scope includes looking at how well transfer learning CNN models perform in modes of pretrained and fine-tuning as well as looking at how well HOG features work when combined with various machine learning techniques.This study expands on previous work by integrating the most effective CNN and HOG characteristics to create a hybrid model for estimating face age. By undertaking a thorough comparative investigation of deep learning and machine learning approaches in the context of face age estimate, the research intends to add to the body of information already in existence. Performance indicators including MAE, MSE will be used in the evaluation to give a complete picture of the models' potential. However, age progression analysis, age-based access management, and targeted marketing are just a few of the many real-world uses for facial age estimate. Nevertheless, despite the increasing interest in this area, further study is required to fully examine and evaluate the effectiveness of deep learning and machine learning algorithms for face age estimate. Existing research has concentrated on one of these methods, but both methods have not been thoroughly analyzed. By carrying out this type of comparison research, our work may fill a sizable vacuum in the literature and provide valuable knowledge on the benefits and drawbacks of deep learning and machine learning models for assessing face age using while face. In order to choose the best strategy based on the particular needs of applications, researchers and practitioners may be guided by our work. This will ultimately result in more accurate and reliable face age estimate systems. In addition, as age estimation from voice is a fascinating research area in speech processing and machine learning. we provide a small analysis of estimate age from human voice and human face image.

1.5 Contribution:

We seek to discover and offer a high-performance model for face age estimate based on the thorough examination and comparison, which may be an invaluable resource in a variety of real-world applications. The main objectives of this thesis are as follows:

- **Transfer Learning with CNNs:** Evaluate and compare the performance of transfer learning CNN models with pretrained and fine-tuning modes for facial age estimation.
- **HOG features Utilization:** Explore the effectiveness of utilizing HOG features with various machine learning algorithms for age estimation.
- **Vision Transformer:** Investigate the applicability of Vision Transformer models for age estimation and compare their performance with previous models.

- **Hybrid Model:** Propose a hybrid model by combining ResNet50 and vision transformer and evaluate its performance.
- **Comparative analysis between ML and DL:** Experiment on the effect of age estimation using both machine learning and deep learning algorithm with combination of features extracted from hog, resnet50 and ViT.
- **Age Estimation from voice:** Analyze the effectiveness of voice and picture features to determine age groups..

1.6 Outline of The Thesis:

There are six primary chapters in the thesis report.The chapter's overall thesis is as follows:

- **Chapter 1:** This chapter describes the motivation, scope, contribution and outline of the research work.
- **Chapter 2:** Literature review of relative work has discussed on age detection from the facial image.
- **Chapter 3:** Preparation of the datasets has been descussed here.
- **Chapter 4:** A brief discussion on proposed methodology has beed given here
- **Chapter 5:** Experimental results and comparison is given in this chapter.
- **Chapter 6:** This chapter concludes the whole work, and future work has been discussed to improve the performance of methodology.

CHAPTER 2

Literature Review

Age estimate has evolved as an important study subject with extensive implications in a variety of disciplines, including biometrics, surveillance, and human-computer interaction. Accurate age assessment from face photos has the potential to transform various industries, including marketing, healthcare, and law enforcement. It can provide customised services, content filtering depending on age, and forensic investigations. Advances in computer vision techniques, along with the availability of large-scale face picture datasets, have driven significant development in age estimate algorithms in recent years.

The purpose of this literature review is to critically examine the existing corpus of research on age estimate from face photographs. This study aims to highlight the strengths and limits of present techniques, examine issues, and provide prospective areas for future research by reviewing the many approaches, procedures, and assessment measures used in this subject. Understanding the intricacies and developments in age prediction from face photos is critical for establishing strong and accurate age estimate algorithms that can be used in real-world settings with high accuracy. In addition to studies on age estimate, earlier works on feature extraction and selection methods specific to face processing were examined.

2.1 Related Work on Age Estimation:

Various human bodily parts or activities can be used to establish age. For instance, a voice signal, a face picture, a fingerprint, iris,and so on. Recent advancements in computer vision and machine learning approaches have significantly increased our capacity to infer age from physiological and behavioral markers. Physiological measurements frequently center on biological indicators including bone density, hormone levels, and telomere length. These indicators reveal information about the aging process at the cellular or molecular level. Behavioral measurements, on the other hand, take into account variables such as gait analysis, cognitive function, and response time. These tests give information on a person's functional age, which reflects the influence of aging on motor

skills and cognitive capacities.

2.1.1 Different methods of age estimation:

The reasearcher of (6) describes a unique method for calculating age from **iris** scans that applies to both children and older individuals. The strategy entails categorizing the age groups into three major categories: children, adults, and elderly persons. The iris pictures are used to extract statistical parameters such as median, standard deviation, and variance. The dataset is divided into training and testing sets, and classifiers are trained using 80% of the samples from each age group. The remaining samples are utilized to assess categorization accuracy. The experimental findings show that the suggested technique achieves excellent accuracy in age group categorization. The outcomes demonstrate that the suggested strategy performs better than the traditional methods in terms of classification accuracy and F1-score.The five distinct classifiers utilized to assess the proposed methodology are KNN, Fine Gaussian SVM, Decision Tree (Medium), Bagged Ensemble, and Linear Discriminant. With an F1-Score of 0.8284, the ensemble bagged tree classifier had the maximum testing accuracy of 83.7%. The study implies that future work may involve categorizing people into smaller age groups (1 to 10 years, 11 to 20 years, etc.) and analyzing data from a bigger dataset with more people of all ages. The suggested approach is constrained by the quality of the iris pictures, and elements like occlusion, lighting, and image resolution may have an impact on how accurate the system is, according to the article.

M. Rajput et al. (7) introduce a revolutionary method for **iris** identification that not only identifies a person, but also offers information about their gender and age. In order to estimate gender from iris pictures, the authors employed deep convolution neural networks (CNN) for age estimation and pre-trained deep CNN models AlexNet and GoogLeNet for gender. To extract features from iris images, they used deep learning pre-trained networks. For the multi-class SVM model performance evaluation, these features were trained and classified.Two different test datasets were used to evaluate the performance of the suggested strategy by the authors. 262 photos of 54 persons ranging in age make up the initial test dataset, which is separate from the training set. 37 respondents' iris photos were taken with the goal of doing research on "iris-based person identification" in the second dataset. The suggested method produced excellent rates of accuracy for both gender identification and age estimation. The suggested method, which utilized characteristics taken from AlexNet, had a gender identification accuracy of 95.34% overall. The trained CNN model was able to estimate someone's age from their iris with an error range of +/- 5 to 7 years, with the majority of individuals having their actual age and a few having a projected age that was close to their actual age. The

authors propose a number of potential enhancements to the system's functionality, such as the use of a bigger dataset, extending it to biometric modalities like face and fingerprint identification, and including demographic information like ethnicity, weight, height, and body mass index.

Another study's (8) key contribution is the suggestion of a strategy for the age prediction challenge that uses just twelve straightforward geometric characteristics taken from **iris** pictures and is computationally simpler and less costly than existing approaches. The study also provides a framework that divides the test population into three age groups, which correspond to "younger," "middle-aged," and "older" categories. It then shows how to accurately estimate an individual's age from their iris patterning by employing more than two potential age-related categories. The study also illustrates how adopting a suitable combination structure to enable a multiclassifier setup gives significant advantages in this kind of situation, offering potential answers to a variety of significant practical issues. The BioSecure Multimodal Database's (BMDB) commercially accessible data Set 2 (DS2) is used in this study. A setup based on intelligent agents allows the suggested method for age prediction from iris pictures to reach an accuracy of up to 75%. In order to enhance feature definition and feature selection for age prediction, the study makes recommendations for future research. It recognizes the drawbacks of the suggested methodology, which ignores texture-based data and solely employs basic geometric characteristics. Additionally, it admits the small dataset and advises evaluating robustness and generalizability using a bigger sample. To increase age prediction accuracy, the report also recommends looking at a variety of biometric variables.

In A. K. Saxena et al. (9), **fingerprints** are used to define age groups. The outcomes of their trials demonstrate the value of the Gabor-based system that these researchers created for categorizing fingerprints into three age groups. For this study, they collected 360 fingerprint pictures (180 males and 180 females) for each age group. The methodology may be expanded further for more exact age estimate methodologies by thoroughly analyzing the work, and the findings can be improved with more samples from other geographical regions. This shows that one constraint of the study might be the need for a bigger and more varied dataset to improve the accuracy and generalizability of the age-group estimation.

A. Abbes et al. (10) also employed **finger vein prints** to identify people's genders and age ranges. The system gets a gender recognition rate of 90%. When it comes to age estimation, the system outperforms itself, with an accuracy of 92% in determining an individual's age range. A number of tests are performed to assess the proposed system on two finger vein datasets, MMCBNU-6000 and UTFVP.

In this (11) study, the researcher presented a **gait-based** human age estimate approach using manifold learning and regression that is dependent on age group. To balance the tradeoff between age group categorization and age group-dependent age regression, they specifically identified five ideal age groups in the beginning. Following that, they used machine learning algorithm is called the directed acyclic graph support vector machine (DAGSVM). It creates a rooted binary directed acyclic network, where K is the total number of age groups, with K(K-1)/2 internal nodes and K leaves. For the ith and jth age groups, each node in the network represents a binary SVM classifier, and each leaf represents an age group selection. While the root node is the simplest to classify between young people and the elderly, the fourth-layer node is the trickiest to classify between two age groups that are close in proximity. By building binary classifiers for each pair of age groups, this method enhances the categorization of age groups by taking into account how difficult it is to classify each pair of groups (from easy to tough). In terms of the mean absolute error (MAE) between an estimated age and a ground truth age, the suggested age group-dependent technique performs better than earlier age group-independent methods. The tests were performed using the OU-ISIR Gait Database, Large Population Dataset with Age (OULP-Age), which includes the largest population (more than 60,000 participants) of any dataset with ages ranging from 2 to 90. In comparison to other cutting-edge methods, the suggested technique delivered the best results. In order to further increase the accuracy of age estimate, the authors recommend that future work may consider merging the suggested age group-dependent framework with deep learning-based techniques. The suggested approach might also be utilized in other fields, such as Parkinson's disease prediction, where typical symptoms are indicated by gait patterns.

To determine age and gender based on **gait**, 18 teams from 14 different countries were given the challenge by the 12th IAPR International Conference on Biometrics (ICB) 2019. The researcher of (12) examines the processes and outcomes of this event. The task comprised a sizable wearable sensor-based gait dataset with 745 participants in the training dataset and 58 subjects in the test dataset, ranging in age from 2 to 78. According to the study, handmade traditional approaches underperform deep learning-based solutions. A total of 67 algorithms were submitted by the 10 teams for age and gender estimation. For age estimation, the algorithms used include Random Forest, SMO regression, Random Subspace, and Kstar. The paper also provides details on the specific algorithms used by each team, such as Label Distribution Age Encoding (LDAE) and Deep Label Distribution Learning (DLDL) for age loss function, and different accumulation methods. The best results were achieved with angle embedded gait dynamic image and temporal convolution network, achieving 24.23% prediction error for gen-

der estimation and 5.39 mean absolute error for age estimation. The work has some limitations, such as a potentially unrepresentative dataset, an emphasis on age and gender estimate based on gait, a failure to take health problems into account, a lack of comparison with other approaches, and a lack of a thorough review of computational complexity and effectiveness.

Using a person's stride, which can be seen even from a distance, the study in (13) present a unique method for estimating age. For an accurate age estimation, their method uses both a global and local convolutional neural network and takes into account the ordinal relationship between different ages. With 63,846 GEIs (age range: 2–90; sample size: 128–88), OULP-Age is the biggest gait dataset ever created. It includes 31,093 male and 32,753 female GEIs. Another well-liked dataset for estimating age from gait analysis is USF but it cannot be utilized to assess the effectiveness of deep learning-based gait age calculation, though, as there are only 122 participants, and the age labels are just approximate. The USF dataset was solely utilized for testing; the suggested approach was trained using the OULP-Age dataset. In order to estimate age based on gait, the suggested method in this research employs the ordinal distribution regression (ODR) algorithm, which integrates the innovative network GL-CNN and a practical loss function ODL. To extract more representative gait characteristics, the GL-CNN comprises of one global and three local sub-networks. The ODL uses cross-entropy loss and EM D2 loss to understand the ordinal correlations between various ages.The described ODR-GLCNN algorithm yields the best outcomes of any method on the OULP-Age dataset, serving as a benchmark for gait-based age estimation. On the OULP-Age dataset, the proposed method outperforms state-of-the-art gait-based age estimation methods having a mean absolute error (MAE) as 3.68 years and a cumulative score (CS) as 83.33 percent. Future research is advised to assess the viability of the suggested approach for gait-based age estimation, investigate its potential for facial age estimation and human identification, incorporate a variety of information, such as clothing and accessories, and create effective feature extraction techniques.

Azhar, Seha et al. (14) proposed a three-dimensional real-time machine learning system for gait-based age detection is proposed. Testing and training are its two key components. The suggested training phase comprises several elements, including the extraction of skeletal data using Microsoft Kinect (MS Kinect), dataset creation, preprocessing, feature selection, best model selection, and age estimation using the best-fit model. Using the information gleaned from the training phase, the suggested testing phase automatically calculates the age of a walking person in real-time. In the study, the approach is referred to as a classical linear regression model (CLRM). Data on the 3D joint positions of 273 individuals, ranging in age from 7 to 70, made up the dataset

utilized in the study. The most accurate age detection performance was reached with a 98.0% accuracy rate utilizing CLRM and 9 features: right shoulder, right elbow, right hand, left knee, right knee, right ankle, left ankle, left foot, and right foot. The greatest accuracy attained by earlier methods is 96.0%, hence this result is contrasted with other current techniques. The research focuses on how the proposed system outperforms the state-of-the-art approaches with regard to of dataset size, few features and outstanding accuracy. The research addresses difficulties in age identification by gait, including constrained scope and considerable estimation errors as a result of constrained age-based categorization. It suggests overcoming these limitations to provide precise age estimates in realistic settings. However, the research lacks precise information on the dataset utilized and does not address the viability of adopting the suggested method in real-world contexts, which might be a significant drawback.

Atsuya, Yasushi et al. (15) introduced a technique for DenseNet-based gait-based age estimate. In order to estimate a person's age, they suggested a deep learning framework that performs better than the current approaches to gait detection. By offering a precise and effective approach for age assessment based on gait, the study hopes to contribute to the sectors of digital signage, consumer analysis, and surveillance situations. The technique utilized in this study uses DenseNet, a cutting-edge network design, as part of a deep learning framework. Employing the OULP-Age database, the suggested technique is assessed. There are two sets in the database: a training set and a test set. With a batch size of 64 and 100 epochs, the network was trained using stochastic gradient descent (SGD). The study used a standard approach that uses GPR with a radial basis function and necessitates extensive computations on a gram matrix. Gait-based age estimate approaches were tested using traditional machine learning methods, and the authors used an active set method with k closest neighbors for each test sample. As a deep learning-based method for age estimate, they also used a modified version of GEINet. The results demonstrated that approaches based on deep learning outperformed those not based on deep learning, and the suggested method performed better than GEINet because of its more dense connection and greater collective information. The suggested strategy had the highest accuracy since it suppressed variances for kids under 15 years old. with an MAE of 5.79 years (mean absolute error). Due to age range restrictions, performance variances, and dataset restrictions, the suggested approach for gait-based age estimate has several challenges. The generalizability of the procedure may be impacted by elements such as walking pace, footwear, clothes, and the surroundings. The development of reliable techniques to handle variability in gait patterns, cross-dataset analysis, and integration with other biometric modalities should be the main goals of future research. Research into appropriate network topologies like ResNet, PyramidNet, and ShakeNet as well as multi-task learning frameworks can enhance accuracy and resilience.

Sakata, Atsuya et al. (16) provided a label distribution learning framework and presented a technique for uncertainty-aware **gait-based** age estimate. In order to be more specific, they created a network that accepts an appearance-based gait characteristic as input and generates discrete label distributions in the integer age domain. Tests on the largest gait database in the world, OULP-Age, show that the proposed method performs as well as or better than the state-of-the-art techniques. This method bypasses the Gaussian process regression (GPR) framework's inability to adequately handle the uncertainty resulting from identical gait characteristics in individuals of various ages. Compared to previous techniques, the uncertainty-aware gait-based age estimate method has a number of benefits. It has more potent expression capabilities, handles comparable gait characteristics with various ages, and reduces outliers by ascribing probability to several age labels. The technique has been demonstrated to perform better than or on par with cutting-edge techniques. The approach makes use of a combined loss function made up of two loss functions: the MAE-based loss function and the KL divergence-based loss function. Both of these loss functions play significant and complementary roles in the method. A standard deviation of 1 produced the best accuracy and the best performance of the suggested approach. Despite utilizing a very straightforward backbone network with an MAE of 5.43, the method outperformed or was at least competitive with cutting-edge deep learning-based alternatives. The uncertainty-aware gait-based age estimate approach has issues, such as the need for huge quantities of labeled data for training and the fact that it is only usable in enclosed spaces. Further studies needs to focus on outside areas and various walking situations, expand the system to handle other sources of uncertainty like clothing or walking speed, and take into account applications beyond age estimate, such as gender recognition or aberrant gait detection.

Qaiser,Zeeshan et al. (17) established the outcomes of age estimate using the study of **inertial data from human walking**. While 86 volunteers performed typical gait activities, we used chest-mounted inertial measurement devices to monitor their 6D accelerations as well as angular velocities. In order to break up the lengthy signal sequences into discrete segments, the information captured were segmented. 50 spatio-spectral characteristics, derived from 6D components, are computed at each step.They trained Random Forests, Support Vector Machines, and Multi-Layer Perceptrons, three different machine learning classifiers, to identify an individual's age. Two different cross validation techniques—ten fold and subject-wise cross validation—were employed to assess the performance of the estimators. The results demonstrate that a subject's age can be predicted more accurately when trained and verified using hybrid data. For example, a random forest regressor produced an average mean square root error of 8.22

years on subjectwise cross validation and 3.32 years and 1.75 years on tenfold cross validation, respectively. They revealed on a larger empirical foundation earlier results indicate that age information can be found in the human gait since their individuals come from two distinct demographic regions, namely South Asia and Europe. Due to the utilization of data obtained on various ground surfaces and instructions given to the participants to walk while feigning various emotions, their suggested technique enables rather reliable age predictions based entirely on the inertial data of only one stride. A person may be identified by their gait if the data used to identify them is not too old, according to the observations on the available data, which also show how gait changes as people age.

Chi Xu; Atsuya et al.(18) conveyed a label distribution learning approach for uncertainty-aware **gait-based** age estimate. They create a network that minimizes the loss function and traditional mean absolute error and provides discrete label distributions in the integer age domain. This approach is useful for age-based person searches and age-based population counts. According to experiments performed on OULP-Age, the largest gait database in the world, the proposed method outperforms cutting-edge methodologies. The database trials demonstrate the use of the uncertainty aware framework in person search and people counting applications. The MAE of the GEINet backbone was decreased from 5.43 years to 5.41 years by the suggested technique, trained utilizing supplemented training sets, while the age estimate accuracy of the GaitSet backbone somewhat improved. The GaitSet backbone achieved roughly 95% IOU, making the estimated persons figures closer to the actual numbers. By preserving the standard characteristics of age labels and changing the uncertainty of the ground-truth age distribution, future study will attempt to enhance the estimated age distribution. Insufficient training samples for age will be addressed as a result. A robust age estimator should also be able to account for factors including view, carrying status, and walking speed.
Ilyas, Alice et al.(19) investigated a novel biometric characteristic **auditory perception** for determining and classifying human ages. Auditory perception, that allow us to hear and locate sounds by receiving and interpreting audible frequency waves through the ears or other methods, is one of the many ways in which humans are able to comprehend information from their dozens of senses. To estimate the participants' ages, a number of machine learning techniques are used, including Random Forests (RF), Support Vector Machines (SVM), Linear Regression (LR), Polynomial Regression (PR), Ridge Regression (RR), and Artificial Neural Networks (ANNs). The outcomes of our investigation have been assessed using an 837-test dataset with participants aged 6 to 60. The results show high accuracy for acceptable categorisation (86%–92%) and age estimate (98.2%), with a root-mean-square error of 2.6 years. The results, which are thought to be significant, show that auditory perception is one of the practical interests in daily

applications. concentration on two crucial areas can improve their method which are improving procedures for increased effectiveness and user comfort, and growing the dataset to produce more accurate tools for age estimate and categorization. In order to create sophisticated anti-spoofing biometric systems,there also need to investigate the possibility of leveraging age-based aural perception to assess prospective assaults. This viewpoint is essential for resolving potential flaws in biometric applications that rely on aural perception for age estimate.

Ioannis et al.(20) offered an approach for age identification using **keystroke analysis** that can be used to give circumstantial evidence for identifying suspects in situations of account or identity theft, as well as masquerade assaults. In a warning system, it may also be used to identify persons taking part in online discussions who are of a different age. This is crucial for spotting cases of child grooming, in which an unwary juvenile may speak with an older person who is pretending to be younger. Overall, beyond criminal profiling, age identification through keystroke analysis offers potential uses. In order for people to provide their ages, an affordable text keylogger named "IRecU" was created. Due to their superior performance over other N-grams, the study discovered that digram intervals were selected for age categorization. 120 digrams were categorized by the majority of participants. For collecting parameters and determine latency averages, "ISqueezeU" was built as a software component. This study achieved average success rate to classify images of 70% while for 26-35 age groups of people of 71.8%.Their work can be achieved high performance result if they incorporated other deep neural networks instead of single three layer neural networks. Besides, larger collection of log file and other parameter fusion can improve their work.

To categorize users according to their age Ioannis, Shahin et al.(21) used **keystroke dynamics** analysis. With a larger dataset, more keystroke characteristics, and a process for choosing the best features for age categorization, they have improved upon their earlier work. In contrast to other forms of information like face photos and social media postings, this study focuses on user categorization using keyboard dynamics, an area that has received less attention. It is distinctive for emphasizing the usage of the results in the area of digital forensics. Their findings demonstrated that they had a 66.1% accuracy rate when determining a user's age group.A three-phase technique was used for the investigation. Free-text information was gathered from consenting participants in the initial phase. The most informative characteristics were given priority in the second stage employing a feature selection algorithm. In the last stage, the authors trained and improved five well-known machine learning algorithms to forecast age ranges of the individuals whose age had earlier been unknown: SVM, Simple Logistic, a Naive Bayes Bayesian Network, and RBFN. An enhancement in the overall effectiveness of

the suggested system has been made possible by these modifications to the research methodology. Keystroke lengths and down-down digram latencies was the among the characteristics that was evaluated. As a result of the high temporal complexity that might result from the vast number of potential features, a feature selection method as required to limit their number. This procedure entails determining the traits that are best at differentiating users depending on age, which was be done by determining the information gain (IG) for each feature. The enhanced system has an 89.7% accuracy rate. By increasing the dataset with users from other ethnic groups, examining the detachment of keystroke pattern features from language, and evaluating additional user variables including race and education level, the research hopes to enhance keystroke identification systems. It also investigates how far users may adjust their typing habits in order to fool recognition systems and proposes the application of other keystroke dynamics characteristics for user categorization.

Buriro, Bruno et al.(22),prposed a method for estimating soft attributes—such as gender, operating-handedness, and age—from **touch strokes** entered on smart mobile devices is presented. When a user enters her confidential PIN or password, they specifically created a technique to estimate the user's gender, age, and operating hand using information from keystrokes. In order to verify whether the feature description matches the actual class, the study concentrated on utilizing keystroke vectors for features from text input. For the purpose of assessing their approach, they used the TDAS the database, which is provides soft biometrics data. The collection contains information about the size, touch pressure, and timing of keystrokes for both 4- and 16-digit number lengths. The Synthetic Minority Oversampling Technique (SMOTE) was employed to extract timing-based keyboard characteristics for various PIN/password lengths in order to overcome restrictions. According to the PIN/password's length, different size feature vectors were retrieved. Various applications that call for various PIN/password sizes were accommodated by using diverse feature lengths. When using timing-based keystroke attributes to estimate age, gender, and operating-hand, the suggested technique had the best accuracy—87.7%, 82.8%, and 95.5%, respectively. by inspecting the study, the accuracy of smartphone authentication will be improved in the future by investigating robust characteristics and fusing the outputs of soft biometric estimators with authentication algorithms.

Using an analysis of the typing habits on computer keyboards and touchscreens, S Roy, U Roy et al. (23) were hoping to point out the age group (0 to 17 and 18 to above) . With a few modest adjustments, this technology is simple, economical, and straightforward to integrate with current systems. Through this method, Internet users and minors may be distinguished. The next level of security, an auto detecting firewall suited for the

users, will be activated the instant a user is determined to be a kid or minor. The goal of the current project is to automatically protect kids from online risks and misuse of their skills in settings for desktop and Android devices. By examining a user's typing habits on a keyboard and on a touch screen while typing, it has been found that users in the age bracket of children may be distinguished from adults. The article talks about a number of timing characteristics that may be derived from keystroke dynamics data, including KD, RR, PP, RP, PR, T-time, Tri-graph-time, and Four-graph-time. With the use of these characteristics, the user's age group may be accurately determined by analyzing their typing habits 92% of the time. By examining children's typing habits on keyboards or touchscreen devices, the suggested method seeks to shield kids from internet dangers. The use of machine learning algorithms for classification is also mentioned in the article. Examples include FRNN with quantifiers and SVM with Radial Basis Function Kernel. The parameters of these algorithms are adjusted to get the best outcomes possible, including the penalization coefficient and the number of nearest neighbors. The enrollment step must, however, be exceedingly exact for this strategy to be accurate. Additionally, the procedure may be impacted and the failure to enroll rate may rise due to factors including mouse dynamics, pressure, and hand weight. In order to investigate more elements and increase accuracy, the essay contends that more study is necessary. It also emphasizes that although turning 18 is the official age at which someone becomes an adult, other factors like mental maturity and ability should also be taken into account.

Ioannis, Cagatay et al.(24) described a method that analyzes a dataset of 387 log files to determine the gender and age categories of users. Over 110,000 characteristics are derived as features from the lengths of **keystrokes** and down-down diagram latencies. These traits were examined by the study, which found that just a tiny number of them comprised a sizeable share of the top 100 features with the largest information gain. Performance is assessed using accuracy, training time, F-score, and AUC for five machine learning models that are tested using all accessible KDs and DDDLs. The top 100 features with the highest IG are used as a starting point for experiments with other feature sets, and then the number of features is gradually increased up to 800. The 10-fold cross-validation method is used to provide objective statistical indicators. According to the study, keystroke dynamics are a reliable way to recognize certain user traits. The concealed nature of user typing patterns and the potential for users to purposefully alter them, however, are causes for worry. As it raises concerns about potential exploitation by malevolent users, the gathering of keystroke dynamics data is sometimes viewed as a restriction. In order to extract more traits and create systems that integrate them, the study will be expanded. Among these characteristics are trigrams, tetragrams, typographical halts, typing corrections, and others. Applying the Dempster-Shafer theorem to solve the issue is still another expansion that might be made. To explore the possi-

ble identification of this trait and assess the method's sturdiness using various keyboard dynamics datasets, the present dataset will be augmented by recording volunteers who speak various native languages.

With the goal to improve speaker age prediction systems based on **speech** signals, Gil, Ron et al. (25) offers a novel dimension reduction technique. Age-group categorization and accurate age estimate using regression are the two age estimation methods that are the subject of the study. The methods use supervectors from a support vector machine (SVM) model called the Gaussian mixture model (GMM) as features. Using an RBF kernel instead of a linear one increases processing cost while improving accuracy. This study showed that traditional dimension reduction techniques, such as the principle component analysis (PCA) and linear discriminant analysis (LDA), may exclude essential feature information, reducing the precision of the model. Weighted-pairwise principal components analysis (WPPCA), based on nuisance attribute projection (NAP), is a unique dimension reduction technique developed in this study. This method did not result in duplicate within-class pairwise variability in the supervectors. The approach was compared to a base system without dimensionality reduction. Smaller feature vectors were found to greatly reduce the training and testing times for SVM. The proposed method also improved system accuracy by 5% and 10%, respectively, for classification and regression. Voice data that was used to train the UBM model were taken from the LDC Switchboard corpus and had age and gender labels applied. There were 2430 presenters in the training sessions, giving six-minute chunks that lasted a combined 50 hours. The testing sessions made use of LDC's Fisher corpus, which comprises of 11,699 unscripted phone calls that lasted 10 minutes each. 5000 male and 7000 female speakers, ranging in age from 15 to 85, are included in the 12,000 recordings in the Fisher corpus. A 26-dimensional vector for features was produced by the study's auditory feature extraction utilizing MFCC. A 13,312-dimensional supervector was produced when 512 Gaussians were used to train a UBM model. As well as using the WPPCA approach for classification tasks, several dimension reduction strategies were utilised. With a parameter value of 100 for each classifier, the preprocessing function parameters for the Young-versus-all learner and the Seniors-versus-all classifier were set to 25 and 55, respectively. Future research can be focused on non-linear kernel-based approaches for dimension reduction.

An another innovative method for age estimate from **voice** signals utilizing i-vectors is suggested in this research of Mitchell, Hugo et al. (26). Each utterance is represented in this manner by the associated i-vector. The session variability is then compensated using a technique called within-class covariance normalization. The age of speakers is lastly estimated using a least squared support vector regression (LSSVR). In the field

of speech recognition, the i-vector, a compact feature vector, is used to represent each utterance for the purpose of estimating speaker age. The utilization of i-vectors in this study has demonstrated a notable enhancement in the accuracy of speaker age estimation. Using telephone calls from the NIST speech recognition assessment datasets, the present inquiry trained and tested a technique. The findings demonstrated that the suggested strategy worked better than traditional methods for determining speaker age, with a smaller mean absolute error and a higher Pearson correlation coefficient. The gains were about 5% for mean absolute error and 2% for correlation coefficient when compared to the best baseline system. The study also examined how spoken language and utterance length affected the age estimate algorithm. Because they contain a huge number of speakers and conform to the i-vector structure, NIST databases were chosen for this project. Its efficacy in other speech recognition databases requires more investigation. The paper discusses difficulties in estimating speaker age, taking into account elements including speech length, content, language, recording technology, and channel circumstances. The effect of speech time on the suggested automated system has to be further investigated.

Sriram, Jason, et al.(27), for the problem of age estimation, employed the same method of a phonetically aware i-vector extractor for age estimation from **voice** as well. The speaker identification sector has shown success with such senone i-vector-based techniques. Fixed-length, low-dimensional i-vectors that have previously undergone conditioning are used to train a support vector regression (SVR) model using a linear discriminant analysis (LDA) transform. In order to more harshly penalize estimation mistakes for those who are younger compared to older speakers, we utilize the logarithm of age as the objective in training the SVR. With telephone speech data from the NIST SRE 2008 and 2010 assessment corpora, the proposed system is put to the test.The study shown that the DNN-based age estimation system consistently outperforms the GMM-UBM-based age estimation system. A mean absolute error (MAE) of 4.7 years for each of the male and female users on the NIST SRE 2010 telephony test set indicates good age estimate ability according to experimental data.The study discusses difficulties in estimating speaker age, including speech length, content, spoken language, and technical aspects like recording equipment and channel conditions. Future research of their work can be examined how utterance duration affects the suggested automated speaker age estimate technique. Future directions should focus on dataset extension, feature design, model optimization, cross-domain assessment, error analysis, and interpretability. The age estimate system's accuracy may be increased by investigating cutting-edge machine learning algorithms and raising the dataset to incorporate a variety of voice data.

Arafat, Zakariya et al.(28) presented an innovative method for identifying the speaker's gender and age from speaker's **voice**. In the suggested approach, bottleneck features are produced using a deep neural network with a bottleneck layer. The Gaussian Mixture Model-Universal Background Model (GMM-UBM) classifier is then given these characteristics in order to classify the data. Transformed mel-frequency cepstral coefficients (T-MFCCs) are created by fine-tuning the deep neural network, which is first trained unsupervised to determine the initial weights across layers. In order to classify the speaker's age and gender, the GMM-UBM classifier build a GMM model for each of the classes. The efficiency of the suggested categorization approach was evaluated using the age-annotated dataset of German telephone conversation (aGender). The results showed that the newly created T-MFCCs have the potential to significantly improve its accuracy of person speaking age and gender categorization when combined with the GMM-UBM classifier. The suggested system's overall accuracy is 57.63%, with adult female speakers having the greatest accuracy (72.97%). The essay speculates that the evaluation of the suggested system's performance on a specific dataset may have certain limitations. The system must be put to the test on numerous datasets and in different languages in order to confirm its dependability. The classification accuracy may also be improved by adding further feature sets, such as prosodic or linguistic characteristics. The system's classification skills may also be improved further by experimenting with other models or by using ensemble approaches.The presented approach could potentially be improved upon by integrating spectral and morphological data with complementary sets, increasing classification precisions and offering a more thorough representation of presenter age and gender. To determine how well it works with various linguistic varieties, regional accents, and demographic groups, it might be tested across a number of datasets. The best size for collecting the temporal information in voice signals might be found by doing window size analysis. To determine if the system is applicable in real-world contexts, robustness analysis might be performed in difficult situations, such as loud surroundings or with non-native speakers.

Another work of the use of **vocal** signals for determining aging and cognitive impairment is discussed by Y Pan, VS et al. (29). In order to estimate age and MMSE, embeddings such x-vectors and i-vectors have proved effective.The Mini-Mental Status Evaluation (MMSE), a test that is frequently used to gauge cognitive deterioration, and age are both estimated jointly in this research using multi-task learning. Two neural network architectures—one based on SincNet and the other on a feature extractor followed by a shallow neural network—are compared to examine the association between age and MMSE. Single-task or multiple-task goals are used to teach both. To put it in perspective, a single-task configuration is used to train an SVM-based regressor. We investigate the characteristics of i-vector, xvector, and ComParE. The explained

Sinc-CLA architecture can be prominence as an effective end-to-end approach for classifying neurodegenerative disorders. An SVM-based regressor was trained in a single-task configuration to compare the performance. The i-vector, x-vector, and Compare attributes were investigated. An internal dataset as well as the ADReSS dataset were used to analyze the trials, which were carried out on the DementiaBank dataset. In the ADReSS dataset's acoustic-only task, the findings demonstrated that multi-task learning enhanced age and MMSE prediction, attaining cutting-edge performance.The best results acquired with speaker embedding features on single-task/multi-task pipeline system estimation is for Age of multi-category 4.64±(0.11) 5.47±(0.29) (x-vector) and (i-vector) respectively. The constraints in the work might include a dearth of real-time updates, a lack of real-time interaction, and compatibility problems. Limited interaction can make cooperation and effectiveness more difficult, while real-time updates may rapidly become obsolete.

Voice categorization is essential for creating more intelligent systems that help with student exams, criminal identification, and security systems, according to A Almomani, M Alweshah, et al (30). Their research's main objective is to develop a method for categorizing and evaluating accent, age, and gender. In light of this, the classifying **voice** Gender, Age, and Accent (CVGAA) method is proposed. In order to improve speech recognition systems that use sensory voice qualities, for instance rhythm-based features that train the system to distinguish between the two gender groups, backpropagation and bagging algorithms are used. The Voice Gender Experiment utilized a dataset to determine an audio's gender on the basis of voice and speech traits. Accuracy rates for the backpropagation NN and bagging were 98.10 and 98.10, respectively. The Voice Common Experiment employed CommonVoice, which demonstrated Bagging's greatest accuracy for age categorization at 55.39%. Bagging has the highest accuracy of 78.94% when utilized in the Speech Accent Archive Experiment to classify languages. Future studies might investigate novel architectures such wav2vec2.0 allowing embeddings from initial waveforms and expand the approach to additional languages utilizing CommonVoice or MLS datasets.

B Wu, H Lu, Z Chen, (31) addressed the problem of missing modality in age data by proposing an Embedding-Regularized Double Branches Fusion (ERDBF) structure for the simultaneously single-modal and multi-modal age estimate. The dearth of multi-modal age estimate methodologies is addressed by this method. They propose a double branches fusion network that comprises of an embedding regularization component and an information interaction component to address the modality that is absent. The former seeks to increase the single-modal representational capability of of features. The latter picks up supplementary intermodal information. Through their cooperation

during the fusion process, the network is able to extract discriminate age representations. Even in the absence of a specific modality, robust age estimate may be accomplished utilizing augmented single-modal information. To the best of our knowledge, the suggested model is a cutting-edge deep learning-driven multi-modal age estimation framework that can easily benefit from the single-modal advancements developed and has more applications. Utilizing embedding regularization and information interaction modules, the ERDBF consists of dual age embedding extractors and a double branches fusion network. The information interaction component combines the Interaction-S and Interaction-M sections to fuse regularized embeddings, improving the resilience, efficacy, and efficiency of the embedding process. For the final age embedding, the two branches are fused together by a fusion layer. The offered information exchange module is plug-and-play, facilitating simple integration with pre-trained models with no trainable parameters. A 5.05 mean absolute error (MAE) and a 0.88 Pearson's correlation coefficient (PCC) on AgeVoxCeleb were obtained from the tests using our multimodal system. Our framework's MAE is 23% better than the public's state-of-the-art on AgeVoxCeleb, which is a relative improvement. Our system performs similarly to the single-modal framework when a modality is absent. The ablation experiment proves the two modules' and the fusion network's efficiency. Researchers would need to test its success in more multimodal tasks in further research. The difficulty of high computational cost, however, frequently hampers multimodal fusion. They'll also concentrate on creating computationally effective fusion techniques to fulfill the real-time requirements of applications like video multi-modal age estimation.

Amruta, Elie et al. (32) descrived in their work that with increased voice channel usage, age estimate by **speech** is becoming more crucial. Based on how it compares with the speaker's recorded age, customer service centers can employ age estimations to modify call routing or offer security. For parental control purposes, voice-controlled devices can use it. It's common to think of the issue of age estimate from speech as a classification or regression issue. The ordinal ordering or age estimation uncertainty that people frequently use is not explicitly accounted for by these systems, though. In this study, the idea is that age implies a normal distribution with a certain confidence interval, centered on the genuine age. KL divergence, GJM distance, and mean-and-variance loss are three alternative distribution learning losses that we examine. On both the NIST SRE08/10 along with AgeVoxCeleb datasets, cross-dataset tests were done, and the results suggest that the distributional learning techniques are extremely competitive and, in most situations, superior than conventional methods.

The significant application of speaker identification algorithms in a variety of industries, including speech interaction, residential services, privacy and access management, and

smart terminals, has attracted a lot of attention. Short-duration speech parts must be handled by today's interactive gadgets, such smart speakers and phone assistants. Existing speaker identification software, however, struggles to identify speakers from brief utterances and instead excels when speech is rather extensive. In order to address this issue, Chakroun, M Frikha et al. (33) provided a unique technique in their research that will improve speaker identification performance for applications that recognize brief utterances. Convolutional neural network (CNN) along with recurrent neural network (RNN)-based novel deep neural network designs were taken into consideration for this aim. The suggested technique is assessed using a probabilistic linear discriminant analysis (PLDA) methodology based on the common i-vector. When significant and brief syllable durations are employed, the experimental findings demonstrated that our model can outperform the i-vector -PLDA baseline system and improve the speaker identification capabilities.

Age determination from speech has lately attracted more attention due to its various uses in user profiling, targeted advertising, and customized call routing. These applications might considerably gain from real-time capabilities and need to estimate the **speaker's age** rapidly. Long short-term memory (LSTM) recurrent neural networks (RNN) have been shown to outperform state-of-the-art techniques when an accurate real-time response is required for pertinent speech-based tasks like language identification or voice activity detection.A novel age estimation method based on LSTM-RNNs was presented in this work by R Zazo, PS, et al. (34). This system is straightforward to construct in a real-time framework and can handle short utterances (between 3 and 10 s). The suggested system has been assessed and compared with a cutting-edge i-vector approach using data from the NIST Speaker Recognition Evaluation 2008 and 2010 data sets. They found that the train MAE/p was 5.62/0.79 and the test MAE/p was 6.62/.73 using conventional MFCCs with 20 static features and a frame size of 20 ms with 50% overlap. However, when they used zero mean and unit standard deviation, the Pearson correlation coefficient (p) was found as 4.58/0.87 as the train MAE/p. The results show that our LSTM-based system consistently beats the reference system when dealing with short utterances, with test utterances of 3, 5, and 10 seconds. The MAE on predicted age varies from 18 to 28 percent improvement. In comparison to experiments that employed LSTM neural networks that were gender dependent, we notice very minimal degradation when using the gender independent system, making it far more suitable for practical, online applications. Additionally, we want to emphasize that the suggested approach does away with the need for a predetermined test utterance length because it enables us to start gradually estimating the age as soon as we get even the first few seconds, and we can keep improving our forecast as we learn more.

A recently suggested time-delay neural network (TDNN)-based age estimate system has produced state-of-the-art results in speech-age prediction challenges. However, when the recording circumstances of each phrase in the testing and training stages are different, the effectiveness of this TDNN-based system can significantly deteriorate. In order to solve this issue, N Tawara, A Ogawa et al.(35) evaluated the effectiveness of many unsupervised domain adaptation (UDA) techniques in order to get the framework consistency against the difference of these circumstances. To investigate an ordinal connection between age labels, they specifically suggest utilizing local maximum mean discrepancy (LMMD) using soft-target labels. In the majority of UDA techniques, the model is trained to produce representations that are domain invariant by minimizing the statistical disparity between the distributions of the labeled origin and unlabeled target data by not taking into account the labels of the age classes. By applying soft-target labels and taking into account nearby age classes, our LMMD-based strategy reduces local discrepancies between the variations on each age class. With mismatched noise from the background, reverberation, and microphones, we ran speech-age estimation studies on internal datasets. The experimental comparison showed that, compared to previous UDA approaches like MMD and reverse gradients, the LMMD-based method effectively reduced the impact of input data mismatches, producing noticeably better results.

Since different people utilize different writing styles, different scripts, different alignments, etc., it might be difficult to determine a writer's age from a handwritten image. Age classification based on handwriting is accurate and reasonably priced because of the papers' plain backgrounds. Using the phase spectrum of the Harmonic Wavelet Transform (HWT), Z Huang, P Shivakumara, et al. (36) develop a novel model for age classification of **handwriting** drawings from 11 to 65 years old. There were 11 classes overall, with a 5-year lag between each. In contrast to the Fourier transform, which produced a noisy phase spectrum because time variations were lost, the proposed HWT-based phase spectrum preserved both temporal variations in phase and magnitude. The proposed HWT-based phase spectrum preserved this essential information despite the fact that handwritten drawings change over time. In order to gather this data, they developed a new phase of statistics-based qualities for age classification based on the hypothesis that writing style also differs with aging. The traits and the input images were fed into a VGG-16 model for categorization of aging. The proposed method was tested on their own dataset as well as the IAM-2, KHATT, and Basavaraja et al. datasets to demonstrate how effective it is in terms of classification rate when compared to the existing approaches. The proposed method outperformed the existing methodology in terms of classification rate, according to results from the proposed and current approaches on a variety of datasets.

V. Basavaraja, P. Shivakumara et al. (37) descrived that forensic and information security professionals are increasingly turning to handwriting analysis for precise indicators, such as human age estimation for immigration inspections. In this article, they provided a novel technique for age estimate using **handwriting** analysis that makes use of disconnectedness and Hu invariant moments features. They showed how Hu invariant moments may be used to extract disconnectedness characteristics for each pair of text component pairings. Hu invariant moments measure multi-shape components based on distance, shape, and mutual position analysis of components. For the classification of distinct age groups, iterative K-means clustering is also recommended. According to experimental results on their dataset and several common datasets, such IAM and KHATT, the recommended approach is effective and outperforms state-of-the-art algorithms. According to the experimental findings, the proposed and actual age categorization rates for their dataset were 66.25 percent and 64.37 percent, respectively. For the IAM dataset, these rates were 63.6 percent and 61.9 percent, and for the KHATT dataset, they were 64.44 percent and 60.3%, respectively. They would like to expand the suggested concept for calculating different ages with a one-year difference because the provided features are script and content-resistant. A qualifying factor is also taken into account when the job is extended in order to improve outcomes.

Through characteristics like slant, pen pressure, and word spacing, handwriting analysis can identify personality. Najla, C. Suen et al.(38) used fixed feature extractors in their work employing **handwriting** samples and the ResNet and GoogleNet CNN architectures. Using the retrieved characteristics, SVM was used to categorize the writer's gender and age. To conduct the analysis, they created the FSHS Arabic dataset.2428 scanned pages make up the dataset, which was written by 43% men and 57% women. With 15 occurrences written in English, Arabic is the language used the most. The majority of them have 12140 rows with 121400 words per page and are written in Arabic.Use the suggested system, and test it. Since features are collected from a dataset using a pre-trained model architecture, the fully-connected layer is not utilized in their study. SVM classifiers and deeper CNN layers are employed. Using ResNet and GoogleNet, the automatic feature extraction approach applied to the FSHS dataset resulted in gender detection system rate of accuracy of 84.9% and 82.2%, respectively. Using ResNet and GoogleNet, respectively, the age identification system's automated feature extraction approach had performance rates of 69.7% and 61.1%. To determine the author's age and gender as a next work, new languages will be studied and also identifying novel writer traits from handwritten manuscripts will also be explored.

Rabaev, Izadeen et al.(39) addressed difficulties in forensic investigations and histor-

ical document analysis by concentrating on automated gender and age prediction activities from **handwritten documents**. Deep neural networks have not been used for age categorization from handwritten documents and current approaches perform poorly, underlining the need for more study. English and Arabic are mostly the focus of the works that were released in this field. This work also takes into account Hebrew, which was learned considerably less than Arabic and English. In response to the success of the bilinear convolutional neural network (B-CNN) towards fine-grained categorization, they suggest a special implementation of a B-CNN employing ResNet blocks. According to their understanding, this is the first instance of categorizing writers' demographics using the bilinear CNN. It has never been done before to categorize ages specifically using a deep neural network. They carry out tests on articles from each of the three baseline data sets that are written in three different languages and provide a thorough comparison with the results reported in the literature. In their work, four groups of ages are given as age labels: (1) "¿15", (2) "16–25", (3) "26–50", (4) "¿50". In all challenges, B-ResNet came out on top. On the KHATT and QUWI datasets, B-ResNet fared better for gender categorization than other models. The proposed method acquired accuracy of 66.65% from paragraph images along with the full dataset. As a future work, they will eventually include the attention mechanism in the B-ResNet model. The attention mechanism greatly lowers the classification error on fine-grained general picture datasets. They also want to investigate the relationship between gender and age classification issues, such as if a certain gender group can more properly classify age and the opposite is true for age.

2.1.2 Facial age estimate based on conventional learning

Based on feature space and feature extraction technique, facial age identification is divided into two approaches: appearance based feature extraction (Global Features) and geometry based feature extraction (Local Features). The entire face is used as feature space in the appearance-based technique, whereas geometric-based approaches extract features from major facial areas such as the brow, nose, lip, and so on.Besides, prior to the deep learning revolution, age estimate techniques relied on manually created characteristics that were derived from face photos.

The earliest approach is based on traditional learning for categorizing face age into three groups: infant, young adult, and senior adult. Based on an examination of skin wrinkles, they classified statistically. The ratios of the spacing between the facial elements, including the wrinkles, are taken as features when using Sobel edge detection to identify wrinkle patterns on the skin of the face. On a tiny dataset, this approach is quite difficult. Different edge extraction approaches, such as the Canny Edge detector

for boundaries global features on the face, the Gabor filter edge extraction, and the Gabor Wavelet at various scales and orientations, were employed as a continuation of the wrinkle extraction methodology.

P. Pirozmand, et al. (40) carried out feature extraction process using the Gabor wavelet, a potent mathematical and biological technique. These Gabor faces can accurately simulate the visual cortex's capacity for pattern recognition. The suggested approach used photos from the training along with testing classes to extract comprehensive face characteristics for age prediction utilizing just the magnitude sections of Gabor wavelets. The magnitude component uses convolution to create an augmented Gabor feature vector from a sample picture.Principal Component Analysis (PCA) as well as Linear Discriminant Analysis (LDA) were employed to reduce the number of dimensions and enhance class separability. Finally, the photos were divided into one of three primary groups using Euclidean distance. Group1 (ages 0 to 3), Group2 (ages 5 to 10) and Group3 (ages 20 to 80) are these groupings. The proposed system's accuracy and robustness were evaluated using the public face aging datasets FG-NET [1] and MORPH [2]with 90% accuracy and obtained a 4.715 inaccuracy on the MORPH-II dataset was attained by this technique.

N. Mehrabi a et al. (41) proposed an age estimate system whaich was broken down into four separate steps in their research. In the first step, it was possible to extract local properties such Gabor wavelets (GW), local binary patterns (LBP), local phase quantization (LPQ), including histograms of oriented gradients (HOG). In the following stage, these properties are integrated using a feature fusion approach that incorporates four separate feature extraction techniques. The next phase involves reducing the quantity of characteristics and identifying the best features using the meta-heuristic optimization approach known as Particle Swarm Optimization (PSO). Finally, they estimated the ages and age groups using classification and regression techniques. The age class was first identified using support vector machines (SVMs), and the ages within those categories were then estimated using support vector regression (SVRs).However, images were changed to grayscale, normalized by eye center, and scaled to inter-pupillary distance to save time and memory. A median filter was used to eliminate noise and rotate the face angles horizontally. The face was located, cropped, and the image was scaled to 240x300 pixels using Viola-Jones face detection. The efficacy of their system in estimating aging was then tested using two popular datasets, the FG-NET and the MORPH. They were able to classify data in the FGNET dataset with an accuracy of 75.69% and an MAE of 3.34 years, and data in the MORPH dataset with an accuracy of 81.66% and an MAE of 3.21 years. Despite the suggested approach's high performance, several elements, such as a lack of data in certain age groups and different environmental circumstances like light

and expression, may have an impact on how effective it is.
The bulk of conventional hierarchical categorization methods used support vector machines to first classify age groupings, and then support vector regression to infer ages within those groups. A distinct hierarchical Gaussian process framework for age estimate automatically was offered by M. M. Sawant et al. in their research (42). A warped Gaussian process regression is used to simulate group-specific aging trends after a multi-class Gaussian process classifier divides the input photos into several age groups. In their study, they individually adjust the hyper-parameters during the regression step for each age group. Additionally, a group-specific WGP regression that effectively learns aging patterns seen at various phases of life had been developed. Due to its adverse effects, use of GPs is typically restricted in many practical applications with the difficulty of calculation. However, at both levels of hierarchy, their approach was computationally effective. The cost of learning kernels for a multi-class issue was equal to that of binary classification at the first level. Their model, which is computationally more effective than the current GP-based age estimate algorithms, learns distinct hyper-parameters for each age group at the second level. Experimental results on two frequently used databases, FG-NET and MORPH-II, demonstrate the effectiveness of the proposed hierarchical architecture. It is gratifying and intriguing to observe that their strategy outperforms cutting-edge techniques like deep learning. An effective expansion of the suggested work is the application of non-parametric transformation functions for WGP regression.

To address the issue of the lengthy processing times and low accuracy of existing age estimation approaches, D. Lu, D. Wang, et al. (43) suggested a new age estimate methodology integrating Gabor feature fusion with an improved atomic search algorithm to acquire feature selection. In order to extract texture information from the face area, the Gabor wavelet transform was first used. The statistical histogram was created to encode and combine the largest feature value on Gabor scales with the directional index. The second strategy is known as chaotic improved atom search optimization with simulated annealing (CIASO-SA), which is a novel hybrid feature selection method based on the improved atomic search algorithm and the simulated annealing methodology. By including a chaotic mechanism during atomic initialization, the CIASO-SA technique also significantly accelerated the system's accuracy and convergence time. The final phase was using a support vector machine (SVM) to categorize the age group. The effectiveness of the recommended strategy was examined using three different resolutions of facial images from the Adience dataset. The Adience dataset includes 26,580 images of 2284 different objects. The dataset divides age labels into eight categories: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, and 60-100. Additionally, changes in light, angle, and location are included. Using the Gabor real part fusion feature at a resolution of

48*48, the maximum average accuracy and one-off accuracy for age classification are, respectively, 60.4% and 85.9%. The collected results show that the proposed algorithm outperforms state-of-the-art methods, which has substantial referential value when used with mobile terminals.
By seeing the works on traditional machine learning approach, the most popular options were biologically inspired features (BIF) and histogram of oriented gradients (HOG) (42) or their variants were used.

2.1.2.1 Handcrafted Facial Feature Extraction:

Facial feature extraction is the process of identifying and removing certain facial characteristics from a picture of a person's face, such as the eyes, mouth, nose, and other regions of interest or even the whole face. Numerous applications, such as facial recognition, emotion recognition, age estimation and human-computer interaction, utilise this method. For the extraction of facial features, a variety of methods can be utilized, including HOG, LBP, PCA, SURF, and HAARIS, among others. Machine learning techniques are typically used to process and evaluate the retrieved data in order to categorize and identify various faces or expressions. Reducing the size of data that represent a big set of data is part of feature extraction. As the number of variables rises, serious issues arise when analyzing complicated data. Analysis with a larger number of variables typically calls for a lot of processing power and memory. When classifying data, effective feature extraction techniques prevent the overfitting of testing data over the training data curve. Construction of variable combinations to accurately describe actual data while also representing it is known as feature extraction.The use of facial recognition technology has grown, but training deep CNNs can take a while and requires a lot of labeled data. This article examines feature extraction, fine-tuning pre-trained models, and transfer learning strategies to increase effectiveness and accuracy.

2.1.2.2 Anthropometric Models:

Human body mathematical models are referred to as anthropometric models. They are used to predict the size, shape, and proportions of the human body based on a number of anthropometric measurements. N. Ramanathan et al. provided a craniofacial development model that explained growth-related shape changes noticed in growing human faces (44). The model was based on the "revised" cardioidal strain transformation model proposed in psychophysical studies of craniofacial development. Because the model took anthropometric data on facial growth into account, it was compatible with the patterns of growth that have been seen in human faces across time. They defined growth criteria spanning major anthropometric landmarks on the face to represent facial growth. Furthermore, they described methods for calculating the optimal

growth parameters and described how age-based anthropometric limits on face proportions translate into linear and non-linear restrictions on facial development parameters. A person's face may be recognized as they become older and their appearance can be predicted using the suggested craniofacial growth model. An age-separated collection of face images of individuals under the age of 18 was used to demonstrate this.

In order to demonstrate the similarities in craniofacial morphology of persons at the same age, P. Koruga et al. (45) appled the anthropometric model of facial aging to photographs of 20 people at various stages of their lives. The first stage in this research wass to establish the face landmarks required for ratio calculation. Euclidean distances between those points of interest are then determined. The next stage is to identify ratios crucial for age estimation. Presentation of study findings and issues found is the final phase. The investigation will make use of the FGNet database, which includes two-dimensional portraits of people from childhood to adulthood. In order to better understand the differences in the structure of the face of infants, children, teenagers, and young adults, future research need to focus on age estimation based on facial images of children and young adults and explore options for categorizing age groups to which a person belongs based on the middle part of the person's face (the upper point of the eyebrows to the bottom of the lip). Additionally, future studies would used a bigger sample and automatically identify landmark sites. Additionally, based on the data gathered in this study, categorization groups would be established using pattern recognition and clustering.

S. Kumar et al. (46) proposed a fast independent component analysis (Fast ICA)-based feature extraction method based on the Kande-Lucas-Tomasi (KLT) with the anthropometric model as the distance measure. Each face is dissected, and the facial organs are marked for distance measurements using the anthropometric model (AM). They acquired the low- and high-frequency KLT facial coefficients at various scales and angles. For subsequent processing, the coefficients were used as a feature vector. Consider the extracted face image and use the Fast-ICA technique, which was entropy-based, to extract the information on the features of the extracted face. The Anthropometric distance to identify facial features and Artificial Neural Network (ANN) utilized to estimate age for all types of facial databases, respectively. Experiments were conducted using the YALE and FERET datasets. The efficiency of the suggested method for extracting face features was supported by an experimental finding that showed the recognition rate Mean Absolute Error (MAE) of the proposed algorithm to be both acceptable and exceedingly promising. Compared to other dimension reduction techniques like PCA, ICA, and MPCA, Fast ICA provides an optimal feature selection at a cheap computing cost. Comparing the Fast ICA to more traditional approaches like PCA and ICA, the

recognition rate is boosted by 90–95%. Additionally, MAE is decreased by 30–33% when adopting the proposed strategy compared to the earlier methods. This method can be used for any other categorization and is appropriate for real-time applications in robotics and visual surveillance systems. They achieved MAE of 7.65 at YALE dataset and 5.20 at FERET dataset.

2.1.2.3 Active Appearance Models (AAM):

Active Appearance Models (AAMs) are statistical models that are applied to the analysis and synthesis of facial pictures in computer vision and image processing. For tasks like face alignment, face tracking, and facial expression analysis, they represent the many shapes and textures of a particular object class, like the human face. AAMs are able to manage complicated variations in position, illumination, and facial expression because they separate the appearance and shape information.

(47) demonstrated an effective direct optimization method that concurrently matched shape and texture, producing a quick, precise, and reliable algorithm. Each time they want to fit the model to a new image, they didn't try to solve a general optimization problem using their method. Instead, they took advantage of the fact that the optimization issue was consistent each time in order to discover these parallels offline. Despite the search space's extremely high dimensionality, this enables them to uncover pathways of quick convergence. To converge quickly and consistently, the approach mads use of a learned correlation between model parameters and residual texture flaws. Finding examples of objects with variable shape and texture, such as tree-like structures or organs with variable shape but constant topology, wass made easier with its help. With a focus on faces and medical images, the algorithm has undergone considerable testing and has produced encouraging results in other fields. Due to the utilization of all available image evidence, it wass more robust than Active Shape Model search, albeit being a little slower. By taking a sample of each color at each sample point, the method can be expanded to handle color images. It offered reliable results when used to track objects in image sequences.

S. Feng et al.(48), presented a novel approach that combined the potency of low-rank matrix recovery theories and cost-sensitive label ranking techniques. Instead of requiring the user to select a single option for each age label, their system prioritized age labels in decreasing order based on their assessed relevance to the presented face image. In order to explicitly capture the correlations between different age labels and control model complexity, the proposed method also integrated the linear prediction functions for different ages into a matrix and incorporated matrix trace norm regular-

ization. The success of kernel techniques for nonlinear generalization also led them to extend the trace norm regularization from a finite dimensional space to an infinite dimensional space.In order to guarantee the practicality of the offered approach for addressing severe prediction challenges, they also incorporated theoretical research on the efficiency of the suggested kernelized trace normalization. Extensive evaluations on a variety of well-known facial image datasets revealed that the proposed framework for age estimation performed better than the state-of-the-art approaches. They discovered 6.02 MAE on the WebFace dataset and 4.59 as MAE or MORPH.

K. Chen et al. (49) created a cutting-edge cumulative attribute notion for learning a regression model when only sparse and unbalanced data are available. To be more precise, low-level visual features were derived from sparse and unbalanced image samples and mapped onto a space of cumulative attributes where each dimension had a label that captured the way the scalar output value, such as age or the number of people, changed continuously and cumulatively. A two-layers regression framework was used after cumulative attributes are created from the scalar values of training samples. In order to transfer the feature inputs to an intermediate attribute space, they first developed a multi-output regression model given any low-level feature presentation of the image. In order to accomplish this, a single structured output model was trained to explicitly associate several properties. Second, using the attribute representation as input, a different regression model was learned to estimate the scalar output. Extensive testing revealed that theur cumulative attribute approach outperforms traditional regression models in terms of accuracy for both crowd counting and age estimation, particularly when the labelled training data is scarce and the sampling is unbalanced. They acquired 4.67 as MAE on FGNet dataset, 5.88 MAE on MORPH dataset.

2.1.2.4 Texture-Based Features:

Texture-based models, sometimes referred to as appearance-based models, are statistical models that capture the variations in the texture or look of a target item or area of interest. For tasks like picture recognition, object detection, and image synthesis, these models are frequently employed in computer vision.

P. Yang et al. (50) suggested to pick features with haar-like properties using a ranking algorithm. Age sequences were arranged by personal aging pattern within each individual in order to construct the paired samples for the ranking model. The paired samples were taken out of each subject's sequence. As a result, the paired data naturally carried the order information. Based on the paired data, a ranking model was used to choose the discriminative characteristics. The ranking model and individual aging pat-

terns worked well together to help choose the discriminative features for age estimate. Different types of regression models were used to construct prediction models based on the chosen features. They achieved 5.67 MAE value on FGNet dataset with 4 fold cross validation.

M. Hajizadeh et al. (51) proposed a face description method called Histograms of Oriented Gradients (HOG) for age-group classification. The suggested strategy divided the participants into four age groups. Pre-processing, feature extraction, and age-group categorization make up the system's three basic stages of operation. HOG features were computed in a number of areas, including the forehead, eyecorners, around the cheekbones, and below the eyes. To create a feature, these local features are joined together for each face image's vector.They used the Iranian Face Database (IFDB) in their study since it contained information on the subjects' true ages. The IFDB had pictures of people ranging in age from 1 to 85. Pre-processing was used to recover the HOG features from the faces, and a PNN classifier was used to divide the images into age ranges. Color frontal photos were converted to grayscale, and face regions were manually trimmed, in order to localize and recognize faces. Manual cropping was used to concentrate on face feature extraction and age assessment because complex methods struggle with lighting, shade, and skin tone. They used age group of 0–2, 3–7, 8–12, 13–19, 20–36, 37–65, 66+ total 7 group of the Images of Groups database. According to experimental findings, the proposed technique had an age-group classification recognition rate of 87.025%. Their future research would be concentrate on including more facial characteristics like anthropometric and Gabor traits to get more accurate findings.

The more difficult issue of automatically estimating age, gender, and race from facial photos taken in the field under unrestricted settings was explored by (52) . They first applied posture and photometric modifications to normalize the input facial image.The center face region as well as the surrounding context region were both included in the extraction of biologically inspired features (BIF) from the normalized face image. Three separate Support Vector Machines (SVM) are employed in this representation to estimate a subject's age group (or exact age), gender, and race. According to experimental results on two substantial public-domain unconstrained face databases (Images of Groups and LFW), the proposed methodology performs noticeably better than the state-of-the-art methods. They used the well-known leave-one-person-out (LOPO) procedure to accurately estimate age using the FG-NET database. State-of-the-art techniques reported a lower MAE (4.1 years with LOPO protocol) than the proposed approach (4.5 years MAE), even if the proposed methodology consistently outperformed in an operational absolute error range of 0–5 years.

O. F. W. et al. (53) looked at the issue of determining a person's age from their face image utilizing a GroupWise age ranking approach combined with learning about ageing pattern correlation. In their suggested GroupWise age-ranking method, they created a reference image set with each person in the reference set grouped by age, and then we utilized this to determine age-ranks for each age group in the reference set. By transforming LBP features into age-rank-biased LBP (arLBP) features and attaching ageranks to them, the built reference set was utilized to train an age estimation function for determining the ages of test images. Their research employing the leave-one-person-out method on the publicly accessible FG-NET dataset and a locally acquired dataset (FAGE) demonstrates the best age estimation accuracy ever. The MAE on FG-NET was 2.34 years. Due to its poor ranking they intended to investigate the suggested approach with some other age ranking and estimation functions besides LSBoost in subsequent works; yet, when rankings were correctly predicted, their method predicts ages pretty well. They also plan to expand on this work by doing experiments with a larger multi-ethnic dataset and a wider age range in order to enhance the generalizability of the suggested approach across ethnicities. They also intended to experiment with trimming this tactic in order to improve the accuracy of age prediction.

2.1.2.5 Multi-feature Fusion:

The fusion of data from many models is a recent development in biometrics. The data fusion approach is the name of this technology. This method has been used by numerous researchers to improve the performance of their systems. Therefore, the method is built on combining many feature representations into a single model.

A deep learning technique for accurate age estimation and a simple yet effective fusion of descriptors based on texture and local appearance were the two novel approaches put forth by I. H. Casado (54).Furthermore, by concurrently optimizing both processes, they also provided a strong deep learning architecture that connects the extraction and regression of key cues. Both methods can be applied to eye-aligned 50/50 photos, unlike many other traditional techniques for estimating age, and they don't require complex statistical face models for precise alignment or additional clues. The trials show that early HOG, LBP, and SURF fusion improves MAE scores by 4.25 years when compared to hand-crafted BIF at 60x60 pixels. This distance can be increased even farther if you utilize larger photos. While requiring more accurate parameter tuning, the deep learning architecture lowers the minimal error to date from 3.98 to 3.88 without restricting the amount of training samples. The durability of these tactics is examined in their work with regard to parameter choices and regularization.In-depth tests performed on two substantial databases (MORPH and FRGC) produced results that were state-of-

the-art compared to earlier research, and these techniques had already been tested in a number of different situations. To validate or enhance performance, different feature fusion techniques should be created and tested, such as feature pooling, dimensionality reduction methods, and late-fusion. Testing of additional aspects should be done both separately and in combination. Future directions include developing deeper network designs, adding a frontalization stage during preprocessing, and improving deep learning methods with additional data augmentation. Additionally, evaluation should look into how errors vary by age, gender, and ethnicity, as well as how well cross-database validation strategies can generalize results.

The grouping estimation fusion (GEF) multistage learning system is introduced by K. Liu et al. (55) for the aim of predicting a person's age from face pictures. Three phases comprise the GEF: Age grouping, age estimate within age groups, and decision fusion for final age estimation were the first three processes. Faces were first separated into a variety of groups, each of which contained a range of ages. The extraction of global features from the complete face and local features from the facial features (such as the eyes, nose, and mouth) took place using three techniques in the second stage. To estimate age within each group, each global or local attribute was employed independently. Choices can be made as a result at the second stage and then transmitted to the third stage for fusion. Six distinct fusion schemes, including maximal diversity fusion, composite fusion, intra-system fusion, intersystem fusion, intra-inter fusion, and intra-inter fusion, were developed and assessed. The performance of the GEF system was evaluated using the Face and Gesture Recognition Research Network and the MORPH-II databases, and it was found to perform noticeably better than the most recent state-of-the-art age estimation methods. In other words, the average age estimation errors on the FG-NET are down from 4.48 to 2.81 years to 2.97 years on the MORPH-II.

2.1.3 Face age estimate using deep learning:

A deep learning-based framework for age classification problems is presented by Dong, Y., Liu, Y., et al. (56) in their research, with a focus on **facial** age prediction utilizing basic handcraft characteristics and tagging faces for age ranges. The paper suggests an algorithm to enhance age categorization by considering the link between age labels and recommending reduced gaps between related categories. In order to outperform more established multi-class classification algorithms like Soft-max, which do not take label connections into consideration, the proposed approach incorporates a new loss function that contains a distance term to account for label relationships. In order to improve performance, the authors of the research suggested a technique that mixes the findings of numerous separate Deep ConvNets models with patches cut from facial photos. When

compared to current best practices, the suggested algorithm performs superbly. Their works followed by developing face detection networks, extract differentiating characteristics, and adjust parameters for the age estimate job. The suggested method produces 10 patches by cropping two centered parts of a facial image on two scales. The input data is utilized to train 20 Deep ConvNets using different patches, and the output from the 20 ConvNets is merged. For the purpose of developing and testing their age categorization system, the authors used the Images of Groups dataset. The authors improved their Deep ConvNets using a randomly chosen sample of 3500 face photos from the dataset, which covers 7 age groups. The age ranges in their works are 0–2, 3–7 ,8–12, 13–19 ,20–36 ,37–65 ,66+ The validation dataset consisted of an independent collection of 1050 face photos. It took an average of 120 ms to predict one image in the testing on the NVIDIA GPU-K40. Works achieve best performance in single and multiple models with AEM (56%) and AEO(92%) scores. Future work should focus on gathering a bigger labeled picture dataset, creating a loss function for chronological age classification, and overcoming occlusion by utilizing the segmentation of images to identify viewable facial regions.

A ordinal deep learning technique for predicting face age was reported in the work by Liu, Hao, et al. (57). They recommended an approach called ordinal deep feature learning (ODFL), which differs from conventional handmade feature-based approaches that require prior knowledge and skill. Their method for directly learning feature descriptors from unprocessed pixels for face representation is as follows. Due to the temporally related nature of age labels and the fact that age estimate is an ordinal learning problem, their proposed ODFL imposes two restrictions on the descriptors, which are obtained at the top of the deep networks: Their study used the topology-preserving ordinal relation first, followed by the age-difference cost information, to make use of the order information in the learned feature space, which was then used to dynamically assess face pairings with different age value gaps. However, with ODFL, age estimation and feature extraction are learned separately, which may not be the best scenario. They also proposed an end-to-end ordinal deep learning (ODL) architecture to address the issue, which would boost our model by utilizing the complementary data from both approaches. On five widely used face aging datasets, including MORPH (Album2), FG-NET, FACES, LIFESPAN, and the apparent facial age estimator, they evaluated their suggested ODFL and ODL. Particularly, the FACES and LIFESPAN datasets are subjected to a range of expressions on the face that produce notable differences in the appearance of facial aging. Each time, they trained their models by randomly removing a particular number of age categories (0-10, 10-20, 20-30, 30-40, 40-50, 50-60, and 60+). Their investigation used the L2 loss function-tuned VGG-16 Face Net and deep regression (also known as Deep Reg.) as a starting point. As a novel method for de-

termining face age, the authors proposed ODFL and ODL. Because ODFL and ODL learn how to represent features directly from raw pixels, they are more resilient to a range of face expressions, size ratios, and complicated backdrops. Because ODFL and ODL can benefit from the nonlinear relationship between face samples and age labels, the ordinal relation to aging pattern is included in the learned feature space. On five face age estimate datasets, ODFL and ODL outperform leading techniques the best. On the Morph Dataset, they increased the MAE with quadruplets and triplets by 3.12, on the E FG-Net Dataset by 2.92 using ODFL + ODL + Cross Entropy approaches, on the Appearing Age Estimation Dataset by 3.95, and on the Lifespan Dataset by 4.51 using ODL + Cross Entropy techniques. By examining their work, we discover that facial age estimation should be dealt with using feed-back deep networks in order to further profit from the supplementary information for the particular aging pattern. In the future, it will be interesting to examine how order information for the face-aging problem might be used to boost the efficiency of age-invariant face identification.

Comparative Region Convolutional Neural Network (CRCNN), which is inspired by how humans think, first compared the input face against nearby faces with a known age in order to create a set of clues (comparative connections, i.e., the input face is either younger or older than every single reference), according to Abousaleh, Fatma S et al (58). Then, in an estimation stage, the age of the person is approximated by merging all the signals. The strategy they came up with has a number of advantages: first, they divided the age estimation task into several juxtaposed stages, making it easier than just figuring out the subject's age; second, in addition to the face being input itself, additional information (comparative relations) can be clearly involved to benefit when performing the estimation task; and third, a few incorrect comparisons won't significantly affect the result's accuracy, making their method more accurate. They asserted that the presented strategy was at the time the first comparative deep learning method for determining facial age. Furthermore, we recommended training the deep network using the Method of Auxiliary Coordinates (MAC), which mitigates the ill-conditioning problem and enables an efficient and dispersed optimization. With relative improvements of 13.24% (on FG-NET), 23.20% (on MORPH), and 4.74% (IoG), the CRCNN significantly outperformed the best results from the state-of-the-art approaches on the respective benchmarks. Examining their efforts, we discover that while aging processes vary between individuals and social groups, further CRCNN study on these themes might be recommended as a "baseline bank" of calculated groupings of social consistency.

In this paper, Zaghbani et al. (59) described a unique autoencoder-based technique for age prediction from face photographs. An artificial neural network called an autoencoder is utilized for the unsupervised acquisition of effective coding. To learn the

form of representation for a collection of data is what it seeks to do. The goal of their research is to leverage autoencoder performance to learn properties in a supervised way to estimate user age. To evaluate the effectiveness of their suggested strategy, they utilized the MORPH and FG-NET datasets. The MAE (Men Average Error) rate, which has a value of 3.34% for the MORPH dataset and 3.75% for the FG-NET, illustrates the resilience and efficacy of the suggested strategy in experimental findings. The work is divided into two sections. The initial step in improving the classification task is to remove the facial region from the image and adjust the user's head position. The second module offers an off-line learnt model and an age prediction based on face traits. It is clear that the effectiveness of each of the modules individually affects the system's total performance. the KNN along with WN classifier were used in the study on age groups of 0–15, 16–30, 31–50, 5.12, and 51 and beyond. Aside from the performance for smaller (0.01,0.02, or 0.03) further emphasizes the significance of the term for similarity preservation. They didn't do anything complicated while optimizing the objective function; they just utilized the SCG. Various optimization techniques will significantly impact the performance of DNN, according to a number of previous research. Due to the restricted availability of labeled data in publicly accessible datasets like FG-Net and MORPH, the suggested technique has shortcomings. This could make it harder to generalize to new photos. Future research will concentrate on utilizing DSSAE to estimate age by integrating speech and face characteristics. Additionally, the approach is not resistant to occlusion and is unable to take into account changes in appearance brought on by things like cosmetics, lighting, and position. Despite these drawbacks, the suggested approach is a promising first step in the direction of creating a more precise age estimation algorithm.

For face-based age estimation, Sendik et al. (60) introduced a combination Convolutional Neural Network (CNN) with Support Vector Regression (SVR) method. Their approach is to first training a CNN for representation learning and then by using metric learning, the learnt features are then subjected to SVR. By initial training the CNN on face recognition, it is possible to get around the dearth of big datasets with age annotations. As with face recognition tasks, they begin by teaching a CNN to categorize face photos to one of C individuals. They refer to this CNN as the Face Classification Network.In the subsequent stage, they applied transfer learning to improve the FCN network by applying an L2 loss to train it for an age regression problem. They called this network the Age Regression Network (ARN).The main goal of their technique is to demonstrate that the ARN's age estimate accuracy may be enhanced by employing one of the layer's Fully Connected (FC) layers of to serve as an age-related face descriptor, even if the ARN can be directly used to age estimation. By employing metric learning with age-tagged face photos, which explicitly optimizes the difference (in

form of L2 sense) between the depictions connected to various ages, this representation is further refined. To account for aesthetic and geometrical variables, CNN application needs preprocessing. In order to find facial landmarks in face photos, the Steepest Descent Method (SDM) was applied in this work. An affine transform was used to align a selection of nine feature points to a canonical face, and then they subtracted average color values per pixel to normalize the result. From the MORPH-II and FG-Net datasets, they looked at age-related face photo databases. The recommended methodology performs brilliantly on the FG-Net and MORPH-II datasets when compared to current state-of-the-art techniques. Domain adaptation is necessary for the study of small datasets like the FG-Net, and we show that this may be done by retraining its SVR layer rather than the CNN. According to reports, the MAE and CS Accuracy were attained in the MORPH-II Datasets at 2.87 and 84.18%, respectively, and at 9.10 on the FG-NET Dataset. In publicly available datasets like FG-Net and MORPH-II, labeled data are few, which limits the applicability of the recommended technique. This might make it more challenging for the algorithm to generalize to recently taken photos. As a sign that they are aware of the drawbacks of their present strategy, the authors note this one and suggest domain adaptation utilizing the SVR layer rather than CNN.

With the aid of deep learning techniques, Xing, Junliang et al. (61) also examined this problem of age estimation from a face photo in their study. They did in-depth analyses of the deep learning models' evaluation and training procedures on two of the largest datasets. Their diagnostic included three alternative types of formulations for the age estimate problem using the five most representative loss functions, together with three distinct architectures to combine multi-task learning with race and gender categorization. Having the mean image removed, the data is a color image measuring 227 by 227. Every one of the practice photos made into the average image. With the appropriate issue formulation and loss function, they achieved cutting-edge performance with the simple baseline architecture. With the help of their recently developed deep multitask learning architecture, the performance of age estimation was further enhanced using high-accuracy race and gender categorization findings obtained concurrently. Using all the information gathered during the diagnosis process, they created a deep multi-task age estimate model that produces MAE of 2.96 for the Morph II dataset and 5.75 for the WebFace dataset, both of which greatly surpass the previous best results. However, the goal of future research is to investigate age estimation issues for certain age groups or people applying deep learning models. In order to transfer information from generic models to particular person-specific models, it is necessary to develop a transfer learning based mechanism and create a principled technique for learning age-related optimization targets. By inspecting the work closely we find some restrictions of the investigation on deep age estimate methodologies include a lack of repeatability, am-

biguous external validation, and data restrictions. The dataset's quality, variety, and representativeness determine the performance of the suggested model, which may not be transferable to other datasets or real-world settings. Progress and acceptance may be hampered by the absence of accessible code, tested models, or resources. Fairness, avoiding prejudice, and fostering trust all depend on an understanding of the model's decision-making procedure. The research also emphasizes the significance of moral issues including data gathering, informed permission, and biases in terms of demographics.

Computer vision confronts a huge difficulty in age estimate founded on real faces, as present approaches require massive face data sets including age labels, restricting the use of unmarked or poorly annotated training records, like as social network photographs. By using advanced convolutional neural networks and using weakly labeled data, Hu, Z., Wen et al. (62) offered a unique learning technique to increase the accuracy of age estimate. 4000 participants are included in the study's face collection of over 150,000 photographs from Flicker that were automatically retrieved and sorted using face identification and alignment techniques. They developed a deep age estimator based on common aging datasets and investigate age disparities employing deep convolutional neural networks. The age discrepancy information for each picture pair is embedded using the Kullback-Leibler divergence. In order to make the distribution on each image show a single peak value, the entropy loss and the cross entropy loss were both adaptively applied. It was intended for the neural network to eventually grasp age based solely on age difference information thanks to the aggregation of these losses. Additionally, they provided a data collection with more than 100 000 facial photos that is accompanied with the dates that each photograph was taken. A timestamp and a person's name are both written on each photograph. They discovered information on age differences utilizing loss functions including entropy loss, cross entropy loss, along with K-L divergence distance. In order to handle faces of people with arbitrary postures, age, and ethnicity, this method is essential for use in real age estimation systems. Using the FG-NET and MORPH aging datasets to examine the accuracy of age estimation, they discovered MAE of 2.8 and 2.78, respectively. In their next study, they expected to investigate other biological traits of humans, such look, hairstyle, height, stance, and gait.

In their study, Liu, H. et al. (63) a "group-aware deep feature learning (GA-DFL)" technique for determining face age. By utilizing deep convolutional neural networks to mine raw pixels for face representation, their GA-DFL algorithm creates a discriminative feature descriptor for each image as opposed to the majority of existing techniques, which use hand-crafted descriptors. They divided up the ordinal ages through a collec-

tion of discrete groups and learned deep feature transformations across ages in order to project each instance pair into the new feature space, where the intra-group variances of positive face pairs compared to the training set are reduced and the inter-group variances for negative face pairs are minimized. They also employed an overlapping linked learning strategy to benefit from the smoothness for surrounding age groups. In order to further enhance the discriminative capabilities of face representation, they created a multi-path CNN technique to incorporate the supplemental input from various scale viewpoints. Experimental results on three public face aging datasets known as FG-NET, MORPH (Album2), and Chalearn Challenge Dataset demonstrate the efficacy of the presented GADFL that were captured in both controlled and uncontrolled environments. These datasets show that our strategy gets outstanding performance compared with most state-of-the-arts. Their study used resized face photographs with 224x224 training data and the OHRank age estimator to more accurately predict facial age. We may infer from this work that individualized face descriptor learning for age estimation may be a promising subject for future study.

The age group-n encoding (AGEn) approach was suggested by Tan, Z. et al. (64) in their study to investigate the link between actual age and nearby ages as a result of sluggish, non-stationary processes. They once splites face photos into overlapped age groups using their age grouping approach. By decoding group categorization findings in accordance with a mapping relationship between age and age groups, this technique assured that each age belongs to a distinct set, enabling accurate age recovery. This technique was more effective than earlier ones, enabling a more precise categorization of faces.Therefore, rather than a group of networks or a cascaded network, their strategy might be used to a single network. A Local Age Decoding (LAD) technique was suggested to quickly determine the expected age by decoding the binary classifiers' outputs locally. In this work,Multiple output layers that individually handled a binary classification problem to ascertain the age of a group of input samples make up the MO-CNN network design for age group classification. Only unclear age ranges may be acquired by the classification system. This implies that while it can recognize age categories like "young adult" or just "senior citizen," it is unable to establish a precise age. A decoding phase is added to the process to solve this constraint. In order to extract an accurate age from an uncertain age range, this phase leverages mapping interactions across ages and age groups. They extended the cost-sensitive strategy of learning used in conventional methods—namely, Cost-Sensitive Dataspace Weighting with Adaptive Boosting and Cost-Sensitive Decision Trees to their designed objective function of the pitched CNN structure for age estimation, which was successful in addressing the imbalanced data issue brought on by age grouping.Their method generates findings that are state-of-the-art on a variety of datasets, including the FG-NET, MORPH II, CACD, and Chalearn

LAP 2015 databases. Age estimate requires accurate face alignment. For this reason, a face detector is used to analyse photos and remove non-facial elements like tattoos. Besides, Active shape models (ASM) identify facial landmarks, and the upper lip and eye center are used to align faces. Then, 224*224-sized cropped photos are put into a network. The suggested technique obtains MAE 2.94 on the Chalearn LAP 2015 with an error rate of 0.263547 and MAE 4.68 on the CACD dataset. Deep learning algorithms for estimating age, however, have a number of restrictions and downsides. The reliability of the system may be impacted by previous methods' real-world limitations, which are mostly brought on by variances in the shooting environment for particular photographs and this can be faced with this problem.

The major goal of the work of Rizwan, S. A. et al. (65) was to create a system for age and expression identification that can accurately identify people's expressions and ages in indoors as well as outdoors.A combined model for age estimate and identification of many facial emotions was suggested in their works. The suggested approach preprocesses an input image before detecting faces throughout the full image. The creation of an artificial beings face mask prediction was then aided by landmark localisation. For the precise categorization of sentiments and age group, a fresh collection of characteristics were retrieved and sent to a classifier. Two benchmark datasets, the Gallagher collection person dataset as well as the Images of Groups dataset, are used to test the proposed method. The following describes the main distribution of their suggested model: First, the YCbCr complexion segmentation model was used to detect faces. Second, using the linked components approach, landmark spots were plotted on the face. Third, a synthetic face mask is mapped onto the face using localized landmark points. Fourth, characteristics were separated into two groups after being retrieved. In order to extract features for age assessment, energy-based point clouds, wrinkles, and anthropometric models are all utilised. Energy-based point clouds, HOG-based symmetry identification, and geodesic distances between reference points are all recovered for expression recognition. Recurrent Neural Network (RNN) technology is utilized to correctly identify age and facial expressions. Ages 0-2, 3- 7, 8–12, 13–19, 20–36, and 37–65 are the target audiences for their works. Additionally, using the Gallagher collection person dataset and the images of groups dataset, respectively, their technique achieved higher classification accuracy rates of 85.5 percent and 91.4 percent. The proposed method can be applied to emotion robots, security systems, video games, consumer applications, and online learning. However, the shortcomings of the works can be include the system's inability to identify people from photographs taken too far away from the cameras by their precise facial characteristics. In the future, they would work on the system's computational time complexity

Due to differences in face features and looks, as well as the absence of significant datasets, age estimate automation has been hampered. Besides, The efficacy of traditional strategies, such as handmade techniques for precise age assessment, has also been hampered by these restrictions.Because of this, Convolutional Neural Network (CNN) approaches have recently been used to Age estimation and picture categorization, both of which have benefited. To the best of their knowledge, Aruleba, I. et al. (66), took advantage of the CNN-based EfficientNet architecture for age estimation. In this work, the EfficientNet architecture was used to age-classify people in accordance with the pertinent age range using the UTKface and Adience datasets. This article offered seven enhanced and tested EfficientNet variants (B0-B6) for age classification. The studies' findings demonstrated that the EfficientNet-B4 variant performed best on both datasets, with accuracy on UTKFace and Adience of 73.5% and 81.1%, respectively. In accordance with the UTKFace datasets, they divided the ages into the following five (5) age groups: 0–13, 14–23, 24-39, 40–55, and 56. The 2284 photos in the Adience dataset are divided into 8 age categories, ranging from 0 to 2 to 6 to 20, 25 to 32, 38 to 43, 48 to 53, and 60 years old. In their methods of work, they followed by drawing a four-sided bounding box and identifying characteristics like the mouth, nose, and eyes, the facial identification method locates faces in datasets. To ensure accuracy across multiple sizes and occlusions, it cut out faces that were recognized and eliminated those that weren't. Various sizes and rotations are compatible with this method. Then, the Facenet technique is used for both face and landmark identification, recognizing spatial locations for annotations and calculating their distances. By taking pictures from various perspectives, face alignment may be accomplished and normalization with image augmentation are performed. The models demonstrated an effective method for resolving issues with learning global features while using less training time and computing resources.In their upcoming work, they would take into account other pre-processing methods that can raise the model's accuracy.

A method for conditional multitask learning that structurally factorsizes an age component into gender-conditioned age likelihood within a deep neural network was presented by B. Yoo et al. (67). It is unclear how the auxiliary activities enhance the model with the primary aim; although they addressed the link for the primary and auxiliary tasks, it is difficult to explain in traditional multitask learning to obtain age estimation. The training of age estimation models is also significantly hampered by the inability to precisely train labels containing discrete age numbers. They offered a method for label extension that increases the number of trustworthy category labels that may be produced from labels with little supervision. They carried out comprehensive testing to validate the applicability of the proposed method on the publicly available MORPH-II and FG-NET datasets. The suggested techniques were more accurate at correctly iden-

tifying gender and age than current approaches. To demonstrate the universality of the suggested methodologies, these performance benefits are validated on popular deep network designs, including VGG-16, CASIA-WebFace, and Alexnet.The suggested multitask deep learning approaches may have boundaries in this assertion because they can only be used to predict a discrete age value which is dependent on with a gender likelihood using face photos. The techniques might not be relevant to tasks or other types of data. Additionally, the label expansion approach may still have accuracy limitations with bigger datasets, even if it can enhance performance with just a couple of precise labels. Utilization of a generalized mean picture from a separate dataset may not always be acceptable and may result in mistakes or inaccuracies in testing, to sum up.

Age-related impacts of aging are a non-stationary phenomenon. While adulthood is mostly impacted by skin texture, skin aging during childhood is predominantly influenced by face shape. Ordinal regression-based techniques can be used to model age prediction. To utilizing these disparities, W. Cao, V. Mirjalili et al. (68) suggested the implementation of the CORAL framework, which ensures consistent rank ordering and reliable confidence scores, backed by robust theoretical assurances. The suggested approach is independent of architecture and may be used to improve any current deep neural network classifier for ordinal regression challenges. When aligned to the reference ordinal regression network, the suggested rank-consistent technique significantly reduces the prediction error in empirical tests on a variety of face-image datasets for age prediction. They collaborated on the Asian Face Database (AFAD), CACD dataset, and MORPH-2.Each picture database was split into 20% of test data and 80% of training data at random. To improve the model training, all photos were downsized to 128 * 128 * 3 pixels followed by randomly cropped to 120 * 120 * 3 pixels. The 128 * 128 * 3 RGB face photos were center-cropped to an ideal input size of 120 * 120 * 3 for model assessment. They selected the ResNet-34 architecture and replaced the final output layer with the appropriate binary tasks to test CORAL's performance for age estimate from face photos. They refer to this implementation as CORAL-CNN. Similar to CORAL-CNN, they updated ResNet-34's output layer to adopt the ordinal regression reference technique, also known as OR-CNN. On the MORPH-2 dataset, they attained 2.64 as MAE and 3.65 RMSE, 3.47 as MAE and 4.71 RMSE on the AFAD dataset, and 5.25 as MAE and 7.48 RMSE on the CACD dataset. The work's issue is that it only considers ordinal regression issues and the expansion of typical CNN architectures to handle them. Different strategies or adjustments to the CORAL methodology can be necessary for other kinds of regression issues or neural network topologies. Additionally, rather than a broad variety of activities, the findings of the experiment only show enhanced age estimate predicted ability on three specific datasets. It's also feasible that the methodology won't produce reliable outcomes for specific datasets or situations.

While deep learning-based age estimation frameworks such as convolutional neural network (CNN), multi-layer perceptrons (MLP), and transformers have demonstrated exceptional performance, they have limitations when modeling complex or irregular objects in an image with a high amount of redundant information.To solve this issue, Y. Shou, X. Cao et al. (69) employed the resilience characteristic of graph representation learning in dealing with picture redundancy information and presents a unique Masked Contrastive Graph Representation Learning (MCGRL) technique for age estimate. Specifically, they used CNN for acquiring semantic characteristics from the picture, which are then segmented into patches that function as nodes in the network.Then, they employed a masked graph convolutional network (GCN) to generate image-based representations of nodes that capture rich structural information. Finally, they added various losses to investigate the complimentary link between structural data as well as semantic features, which enhances GCN's feature representation capabilities. The results of experiments on real-world face picture datasets show that their suggested method outperforms current state-of-the-art age estimate algorithms. They built a self-supervised masked graph autoencoder (SMGAE) that executes mask reconstruction on nodes in order to increase the fusion representation ability of node characteristics and structures in graphs. SMGAE encoded and decoded feature vectors had improved semantic representation capabilities. Furthermore, to broaden the gap between various classes while narrowing the gap among the same classes, they used a contrastive learning technique to increase the model's generalization performance. The MORPH-II, FG-Net, and CACD databases are commonly used for age estimate.The MORPH-II, FG-Net, and CACD databases are commonly used for age estimates n. As a result, this research chooses these three benchmark datasets to validate the efficacy of our MCGRL technique. They obtained 2.39 as MAE 89.9 as CS on the MORPH-II dataset, 2.86 as MAE 88.0 as CS on the FG-Net dataset, and 4.03 as MAE 80.1 as CS on the CACD dataset. The work's shortcomings is that it primarily focuses on age estimate tasks, and the efficiency of the MCGRL structure may vary substantially for different tasks or datasets. Furthermore, the model's performance may be influenced by the quality and variety of the training data, and the suggested self-supervised masked graph autoencoder as well as contrastive learning processes may not be ideal for all datasets or picture characteristics. Finally, because the studies were only conducted on a limited amount of benchmark datasets, the generalization potential of MCGRL to additional datasets is not yet evident.

The age estimation based on images is based on multiple latent heterogeneous variables, such as gender, however exploitation without reliability analysis could reduce accuracy due to uncertain noise. Chen et al. (70) proposed a Feature Constraint Re-

inforcement Network (FCRN) to take advantage of the constraint gender influence on age estimation, which was motivated by the fact that gender had a significant impact on face at a particular age stage. The model estimates the confidence of effect on age estimation techniques for multi-scale latent heterogeneous features. It particularly collects the gender and age characteristics using both classification and regression. The age prediction result is then improved by the model's use of gender variables obtained by constrained gender features to enhance and compute the effect of different genders for age predictions among different age groups. Extensive tests were carried out using available public aging datasets. It utilized UTKface, Morph, AFAD, IMDB WIKI datasets for their works.The findings demonstrate the usefulness and effectiveness of the suggested technique.

Age-invariant face recognition (AIFR) and face age synthesis (FAS) are two methods that have been used in the past to remove age variation. The former, however, lacks visual results for model interpretation, while the latter successfully removes age variation by transforming the faces of different age groups into the same age group. The result was the introduction of MTLFace by V. L. C. Wang, et al. (71), a unified, multi-task framework that can acquire age-invariant identity-related information while providing pleasing face synthesis. To be more precise, they employed an attention technique to separate the mixed face data into two unrelated parts—the identity- and age-related aspects—and then decorrelate these parts using multi-task training and continuous domain adaptation. In contrast to the usual one-hot encoding that accomplishes group-level FAS, they offered a novel individual conditional module that achieves identity-level FAS, along with a weight-sharing method to increase the age smoother of synthesized faces. In addition, they gathered and disseminate a huge cross-age face dataset with age and gender annotations to enhance AIFR and FAS. Extensive studies on five benchmark cross-age datasets show that their proposed MTLFace outperforms state-of-the-art approaches for AIFR and FAS. They further validate MTLFace on two major general face recognition datasets: g CACD-VS, CALFW, AgeDB, and FG-NET, demonstrating competitive performance for face identification in the field.With leave-one-out cross validation, they achieved 96.23% accuracy on the AgeDB-30 dataset, 95.62% accuracy on the CALFW dataset, 99.55% accuracy on the CACD-VS dataset, and 94.78% accuracy on the FG-NET datasets.

face age estimate is a common face analysis subtask wherein a model learns the various facial ageing aspects from multiple facial photos. Despite several research establishing a link between age and gender, relatively few studies investigated the possibility of introducing a gender-based system comprised of two independent models, each trained on a different gender group. V. Raman, K. E et al.(72) paper tried to overcome that

gap by providing an age estimate algorithm with two key components. The first part is a custom-built gender classifier which differentiates between females and males. The second module is an age estimate module comprised of two models. Model A had only been trained on female photographs, whilst Model B had only been taught on male images.The system starts with a picture that is supplied, extracted the facial gender, and then sends it to the proper model based on the expected gender label. In addition, to detect faces in a picture, the technique employs the C++ Deep Learning Library (dlib) and OpenCV. Cropped faces were isolated from the remaining ones, and dlib was used to distinguish left as well as right eye locations. For alignment and rotation, coordinates ware extracted. Their age estimate models were based on networks developed by the Visual Geometry Group (VGG16) and have been tweaked to meet the nature of our situation.The gender estimation classifier was a trained binary classifier that produced a sigmoidal output around 0 and 1. Labeling male photos with ”0” and female ones with ”1” The output was generated once the model was fed a particular picture x. The gender categorization model takes a 96*96 RGB picture as input and employs four hidden layers alongside two fully linked layers. Model A and Model B are VGG16 networks that have been pre-trained against the ImageNet dataset as well as the experimental dataset, respectively. The model changes the input size and includes a dense layer with 512 neurons driven by the ReLU function, as well as a dropout layer. The output softmax layer corresponds to four age groups. Individually, the models attained accuracies of more than 85%, while the system achieves an overall accuracy of 80%. To evaluate performance on unseen data, the suggested approach has been trained and evaluated on the UTKFace dataset as well as cross-validated against the FG-NET dataset.Childhood (0-12), Teenage (13-19), Adulthood (20-59), and Senior Citizens (60+) are the four age groups considered. On the UTKFace and FGNet datasets, this grouping of age labels was favored to cover all conceivable age categories, with 80.76% and 66.50% accuracy, respectively. They also tested the results without the gender model and obtained 70.46% and 48.00% accuracy on two datasets, respectively. According to their findings, decreasing the age difference and increasing the number of age groups decreases accuracy. This problem arises because certain age groups may overlap.One of the key constraints of the work is the lack of adequate data for exploring the integration of generative adversarial networks (GANs) to produce more training examples.

S. K. Gupta et al. (73) evaluated the Single Attribute either Gender or Age as well as Multi-Attribute both Gender and Age prediction models to assess the gender effect on age estimation. They reviewed the face age estimation and gender classification techniques developed to date using traditional and deep learning methodology, as well as an evaluation of their benefits, drawbacks, and implications for further research. The databases used for benchmarks results and their characteristics for both controlled and

uncontrolled contexts were also given in this research. Gender categorization and age regression in a same network are combined in multi attribute heterogeneous prediction. It is more difficult than the single attribute prediction issue, but it has many applications. Because of the smaller size of the weights model, feature extraction takes less time and consumes less memory. The multi-attribute prediction models incorporate both attribute heterogeneity and attribute correlation in a single network and provide category-specific feature learning for heterogeneous attributes as well as shared feature learning for all attributes.The review results show that deep learning (Alexnet)-based techniques outperform handmade feature engineering; nevertheless, they need massive computational and data regularization resources. There are still significant obstacles in real-time facial gender categorization and age estimate, such as face localisation in the wild, feature identification of the juvenile age group or children, blur and de-focused faces, human expression, occlusions, and ethnicity (race). Humans are employing face masks in the current Coronavirus epidemic, creating extreme obstacles for face localisation and concealed facial feature extraction for occluded faces, making it exceedingly tough to identify. Future research on these difficulties is required to develop stronger and more powerful face characteristic identification for use in real-life settings.

The image resolution will degrade, however, if a low-resolution camera is used to take the pictures or if close-up facial images are taken of people. In this case, it is impossible to determine the age due to the loss of information regarding wrinkles and the texture of the skin on the face. Research on estimating age before did not account for resolution loss and only used high-quality facial images. S. H. Nam et al. (74) provided a deep convolutional neural network (CNN)-based age estimate method that overcomes this limitation by first reconstructing low-resolution face images as high-resolution ones using a conditional generative adversarial network (GAN), and then using the images as inputs. The PAL and MORPH databases, two open datasets, were used in an experiment. The results demonstrated that the proposed method performs better in high-resolution reconstruction and age estimate than the most recent methods. Furthermore, the suggested SR based on conditional GAN outperformed the current VDSR and DCSCN in terms of SR performance. When the suggested method's processing time was assessed on a desktop computer as well as an embedded system, the real-time processing speeds were 27.8 frames/s and 3.8 frames/s, respectively. A further research will look into how video pictures might be used to improve SR and age estimation performance. The suggested method's applicability to low-resolution photos with both optical and motion blurring must be investigated. Furthermore, it is required to test if the suggested technique is useful in reconstructing face pictures taken in low-light or at night, as well as estimating the ages in such images.

Access control, social robots, corporate intelligence, and digital signage can all benefit from age estimate using facial photos. Convolutional neural networks, which are used in deep learning, have produced reliable approaches, however they are time and resource constrained and need a sizable, consistently annotated training dataset. A. Greco et al.(75) suggested an alternative to the usual knowledge distillation process. The used the knowledge to overcome the particular restrictions of the age estimation issue, notably the lack of a sizable dataset with trustworthy annotation and the absence of a convenient method for training powerful CNNs for age estimation that can be used in practical situations.Two steps make up the process: In the first phase, they used a teacher technique to automatically annotate the VGG-Face2 Dataset using a group of CNNs trained to estimate age. They refered to this dataset as VGG-Face2 Mivia Age (VMAGE). Then, they utilized VMAGE to train a range of smaller student models, thereby transferring the teacher's expertise to these pupils. All the CNNs trained using this method performed remarkably well on LFW+, LAP, and Adience, outperforming the similar models built using IMDB-Wiki, the dataset that has historically been used by the scientific community. In instance, the student model built on SENet was able to achieve the third-highest rank over Adience and the greatest performance over LAP 2016 (apart from the instructor model). The most recent experiment using LFW+ also showed that student models built using SENet, MN3-Small, and VGG were more resistant to picture corruptions than instructor models. All of these achievements were attained with a training effort that was significantly lowered and an inference time that was around 15 times quicker than the instructor model.by inspecting their works we might say that, extension of the training set to include additional photographs of young children and elderly people might help to balance the dataset in the future. Additionally, the age group classification fine-tuning technique could make use of the ordinal ordering of the age groups. Without making the training process more difficult, these gadgets may significantly enhance the performance of the student models over LAP 2016 and Adience.

Directed Acyclic Graph Convolutional Neural Networks (DAG-CNNs), which utilize multi-stage data from several CNN layers to estimate age, were introduced by S. Taheri et al (76). In addition to merging the feature extraction and classification phases of the age estimation into a single automated learning process, DAG-CNNs also use multi-scale features and automatically combine the results from numerous classifiers. DAG-CNN eliminates the need to extract manually created features as a result. Only the upper-level characteristics that are retrieved by CNN's last layer are utilised in the majority of current techniques. The suggested multi-scale system may automatically learn various level of features, combine them, and estimate the subject's age instead of conducting feature-level fusion explicitly and sending the results into a classifier. All of

the input photos have been downsized to 256x256 and are in RGB format. The DAG-VGG16 network is fed five different cropped sizes of 227x227 and their flip, whereas the size of the cropped input pictures for the DAG-GoogLeNet architecture is 224x224. The offline multi-stage feature fusion systems use the same techniques for data augmentation. Morph-II and FG-NET are employed with age groups of 20, 20-29, 30-39, 40-49, and ¿50 in order to explore the appropriateness of DAG-CNN, and MAE of 2.87 on Morph-II and 3.05 on FG-NET were obtained. Future study will utilize the suggested strategy on more aging datasets in various scenarios.

In order to identify characteristics related with facial photos, Z. Jiang et al. (77) suggested a reliable method for age prediction based on local binary patterns. To successfully overcome obstacles including variations in lighting, positions, and facial expressions, people's ages from face photographs must be predicted with accuracy using their study. They called the suggested technique, which combines feature extraction, feature selection, and machine learning methods, a hybrid method.First, facial landmarks were found in order to identify the main characteristics of the face and made it possible to extract the related facial traits. In order to reduce dimensionality and improve model effectiveness, these features wew then sent into a feature selection algorithm to determine which ones are the most unique. The age estimation software can be run by the authors' implementation system in around 3.5 seconds, according to them. To extract facial characteristics for this study, the LBP approach will be employed.A genetic algorithm had been used to classify the gathered information more effectively. The genetic algorithm will made use of the generated training data set to determine the fitness function for the produced chromosome.The tests made use of the databases FG-NET, MORPH, and FACES. The proposed approach is quicker and more accurate than earlier algorithms like nearest neighbor as well as supported vector machines. The ultimate sensitivity, accuracy, and precision of the suggested approach for guessing people's ages are each 97.2%, 96.8%, and 99.1% respectively. Future study may concentrate on improving preprocessing techniques, reducing computational cost, picking appropriate training standards, and using multi-objective algorithms.

As part of a research on how people perceive their age based on photographs, the visible age estimation focused on single face photos has lately attracted attention. For apparent age estimate from a single face picture, C. Miron, V. Manta et al. (78) provide an effective convolutional neural network design that achieves outstanding results without the need for pretraining. The design only needs 79k parameters and comprises of 9 convolutional layers and 2 max pool layers. Additionally, we enhance the findings by using a weighted class distribution, which ensures that classes with high representation do not distort the prediction results. These are contrasted with other findings in the

literature, both for pretraining-based approaches and those that don't. The suggested technique, which is distinguished by its effectiveness in both training and testing and by using orders considerably fewer parameters and a great deal less training time than previous systems, produces estimation errors equivalent to the individual benchmark and to existing methods.

S. E. Bekhouche et al. (79) provided a technique for automatically determining a person's age using three different components: face alignment, feature extraction, and age estimate. Face alignment is a technique used to identify faces in photos, correct each face's 2D or 3D position, and then trim the area of interest. Due to its dependence on earlier stages and the potential impact it may have on system performance, this preprocessing stage is crucial according to their study. The processing stage might be difficult since it must deal with several changes that may arise in the facial image. The facial features are extracted in the feature extraction step. Either texture descriptors or deep networks are used to extract these characteristics. The final step was feeding the retrieved characteristics to a regressor to calculate the age. They used the Viola-Jones algorithm-based cascade object detector to find faces in the crowd. The ERT method was then used to identify each face's eyes. They used a 2D transformation according to the center of the eyes to align the face in order to correct the 2D facial position in the original image. On the PAL dataset, 4.09 on the FG-NET dataset, and 3.88 on the FACES dataset, they get an average MAE of 4.62, 4.09, and 3.88, respectively. Future work will focus on creating brand-new CNNs from scratch and optimizing existing CNN architectures with face datasets, according to the researchers.

In order to overcome the difficulty of determining the age of a human face, I. Dagher et al. (80) applied transfer learning to a few pre-trained CNNs, such as VGG, Res-Net, Google-Net, and Alex-Net. These networks were improved utilizing several testing and transfer learning in order to get the appropriate output amount and age gap. A special hierarchical network that generates a highly accurate age estimate was built on the base of these tests. This new network, composed of a number of pre-trained 2-class CNNs (Google-Net) with the ideal age gap, can better organise facial pictures according to the age group to which they belong. On the FGNET and the MORPH databases, it was evaluated against other cutting-edge methodologies to demonstrate its efficacy. Getting the ideal number of outputs (classes) from such networks to get high age estimate results and determining the ideal age gap have been their two main studies. The selection of the FG-NET + MORPH dataset was made possible by the vast age range it spans [0-77]. Changing the classifier module layers with new ones linked to the new studied tasks is one of the standard transfer learning processes. Results for accuracy for six classes. On Google-Net, the results for A:0-5 B:6-10 C:11-19 D:20-29 E:30-39 F:40-

77 are 74%.Five classes' accuracy results. The Google-Net score for A:0-9 B:10-19 C:20-29 D:30-39 E:40-77 is 85%.Results for 4 classes' accuracy. The percentage for A:0-14, B:15-29, C:30-44, and D:45-77 on Google-Net is 87%. Three classes' accuracy results. On Google-Net, A:0-19 B:20-39 C:40-61 is accomplished as 89%. Results for two classes' accuracy.

Besides, Fariza et. al(81) used the fact that the Deep Residual Network (Resnet) which is a convolutional neural network architecture is simpler to tune and can increase accuracy as a consequence of a significantly rising depth. With the deep residual network (Resnet) model, they provided a novel method for age estimation on convolution neural networks (CNN). Resnet outperforms other cutting-edge image categorization techniques according to their literature. They contrasted Resnet and a straightforward linier regression model architecture with ResNeXt. They utilized the UTKFace dataset to assess the effectiveness of the residual network for age estimate of individuals between 1 and 100 years old. In comparison to Resnet-50 and linier regression, their result demonstrated that the ResNeXt-50 (324d) architecture produces better age estimate results. They depicted that ResNet-50 and ResNeXt-50 both achieved the same training accuracy and training loss, but ResNeXt-50 achieves a greater MAE. Besides, ResNeXt-50 considerably decreases MAE from Resnet-50 by 1.53 years. In a similar vein, Resnet-50 and ResNeXt50 provided superior estimation outcomes over linear regression. Because the UTKFace data collection is composed of faces found in the wild, those techniques arranged studies on this difficult data set. Overall, it demonstrated that the deep residual network classification was suitable for resolving the age estimation problem. They achieved 7.21 MAE on UTKFace dataset.

2.1.4 Attention-based facial age estimation:

However, cutting-edge CNN-based approaches treat each face region identically, neglecting the significance of particular facial patches that may carry valuable age-specific information. The way that the attention mechanism works parallels the underlying workings of biological observation behavior, sharpening some areas of observation. Speech recognition, image processing, and machine translation all make extensive use of the attention mechanism because it can swiftly identify key elements from sparse input. A multiheaded attention mechanism has more than one attention head, each of which can focus on a different informative portion of the input. As a result, the quantity of attention heads determines the number of segments that the multiheaded attention mechanism can focus on. The self-attention layer in computer vision receives the feature map as input and computes the attention weights between each pair of features to provide an updated feature map where each position holds information about any other feature in the same image.

S. V. L. C. et al. (82) offered Attention-based Dynamic Patch Fusion (ADPF), a face-based age estimation approach. ADPF includes a bespoke feature extractor comprised of an AttentionNet as well as a FusionNet. The AttentionNet dynamically creates age-specific patches using a unique attention mechanism, whereas the FusionNet forecasts the subject's age by combining characteristics acquired from the face picture with the identified age-specific patches. To increase speed, the found patches are put into the FusionNet in descending order according to the quantity of age-specific information they contain. To that aim, they included the Ranking-guided Multi-Head Hybrid Attention (RMHHA) mechanism into the AttentionNet. Instead of employing MHSA's multi-channel feature maps, each attention head in RMHHA produces an efficient single-channel attention map, resulting in is utilized to crop the associated age-specific patch from the face picture. RMHHA ranks the relevance of the created attention maps by assigning learnable weights to them. They demonstrated that ADPF delivers state-of-the-art performance on numerous face-based age estimate benchmark datasets through extensive trials. They further show that, as compared to our previous work, ADPF significantly reduces training times. They ran studies on three widely used face-based age estimate benchmark datasets: MORPH II, FG-NET, and the Cross-Age Celebrities Dataset (CACD). They attained an MAE of 2.54 on MORPH II, 2.86 on FG-NET, 4.72 (train) and 5.39 (val) on the CACD datset. In the future, they would look at the construction of tailored estimators to improve performance even further, such as taking into account ordinal information throughout ages and decreasing the gap between label and feature distributions.

Deep learning convolutional neural networks (CNN) provide unparalleled benefits in processing visual characteristics. Convolutional neural networks are demonstrably superior to more traditional methods in practice for determining the age of facial photographs. Anyway, it gets harder and harder to install models on mobile terminals as neural networks get deeper and get bigger and more complex. Liu, Y. Zou, et al. (83) proposed an enhanced ShuffleNetV2 network based on the mixed attention mechanism (MA-SFV2: Mixed Attention-ShuffleNetV2) by transforming the output layer, combining classification and regression age estimation methods, and highlighting key features through image preprocessing and data augmentation techniques. The final age estimate accuracy can be compared to the state-of-the-art thanks to a reduction in the impact of noise vectors that include background information unrelated to the faces in the image.A 224 224 pixel RGB (Red, Green, Blue) facial image serves as the network's input. The MORPH2 dataset and the FG-NET (Face and Gesture Recognition Research Network Aging Dataset) are both used in this investigation. On the FGNet dataset, the final mean and variance of MAE were 4.12 (0.025) and 3.81 (0.031), respectively, and 2.58 on the MORPH2 dataset. They performed 33 tests. Future study would concentrate on how to

use a more compact model to achieve the necessary age estimation impact because the model described in this paper is still too heavy to deploy to a mobile terminal.

S. Hiba et al. (84) presented a cutting-edge deep-learning method for age estimate using facial image data. They started out by presenting a dual image augmentation-aggregation strategy that was focused on attention. As a result, the network was able to leverage many facial image augmentations that were aggregated by a Transformer-Encoder. The aggregated embedding that results effectively encodes the properties of the facial picture. After that, they put forth a framework for probabilistic hierarchical regression that combines an ensemble of regressors with a discrete probabilistic estimate of age labels. Each regression model was specially tailored and trained to improve the probability estimate over a variety of ages. When used to age-estimate data from the MORPH II dataset, their approach was demonstrated to perform better than existing schemes and gave a new state-of-the-art level of accuracy. The bias analysis of the most recent age estimate data was then introduced. By using the VGG16 backbone network with RS protocol, they were able to achieve 1.13 MAE on the e Morph II dataset. However, the MORPH II dataset's interpretability issues, lack of application, assessment metrics, and tiny dataset are some of the drawbacks that the deep-learning technique for estimating age from face photographs may have.

The accuracy of age estimation of face pictures found in the wild is quite low because existing algorithms only take general characteristics into account and overlook the fine-grained properties of age-sensitive regions. Using their attention long short-term memory (AL) network, K. Zhang et al.(85) proposed a novel method for fine-grained age estimation in the wild that was inspired by the fine-grained categories and the visual attention mechanism. In order to extract local characteristics from age-sensitive areas, this technique developed AL-ResNets or AL-RoR networks utilizing LSTM units and residual networks (ResNets) or residual networks of residual networks (RoR) models, which considerably improved the accuracy of age estimation. A ResNets or RoR model was selected as the fundamental model and pretrained on the ImageNet dataset before being refined on the IMDB-WIKI-101 dataset for age estimation. On the target age datasets, the general characteristics of face images were then retrieved using fine-tuned RoR or ResNets. In order to extract the regional properties of such age-sensitive regions, the LSTM unit was then employed to automatically determine the coordinates of the age-sensitive zone. On the FG-NET, 15/16LAP, and MORPH Album 2 datasets as well as directly on the Adience dataset, age group classification tests utilizing the Deep EXpectation algorithm (DEX) were completed. By combining the global and local factors, they came up with our final forecast results. Results from experiments showed that the proposed AL-ResNets or AL-RoR was effective and reliable for age estimation in

the real world, outperforming all other convolutional neural network (CNN) methods on the Adience dataset with 67.832.98% accuracy and 97.530.59% 1-off, MORPH Album 2 with 2.36 as MAE, FG-NET with 2.39 as MAE, and 15/16LAP datasets with 3.137 as MAE.

2.2 Face Detection:

A common use of computer vision technology is face recognition, which enables machines to correctly identify and validate people by analyzing their facial features. The system analyzes and categorizes facial traits using a combination of cutting-edge algorithms including machine learning models, then compares them to a database of recognized faces. Although commonly utilized in daily operations like phone unlocking, attendance tracking, and border control, facial recognition systems still have issues with occlusions and different camera types.People can be recognized using facial recognition technology in real-time or in still images and videos. It falls under the biometric security category. There are numerous face recognition libraries available, including the cutting-edge, deep learning-powered from dlib and so on.

The use of facial recognition technology has grown tremendeously, but training deep CNNs can take a while and requires a lot of labeled data. H. O. Ikromovich et al. (86) examined feature extraction, fine-tuning pre-trained models, and transfer learning strategies to increase effectiveness and accuracy. They used the Labeled Faces for the Wild (LFW) dataset for their research, which included more than 13,000 face photos of 5,749 people. In order to compare the effectiveness of their transfer learning strategy to a baseline CNN that was on the same dataset, starting from scratch. In comparison to training from scratch, their findings demonstrate that transfer learning with fine-tuning previously learned models greatly increases the accuracy of facial recognition. Though less efficient than fine-tuning pre-trained models, feature extraction also produces encouraging results. They also ran tests to look into the impact of the quantity of labeled data offered for training.

J. M. Sahan et al. (87) proposed a novel face recognition approach that integrated a one-dimensional CNN deep convolutional neural network (1D-DCNN) classifier with linear discriminative analysis (LDA) methods. The one-dimensional face feature set created by LDA from the original picture database to train the 1D-DCNN classifier was the contribution of this paper, which helped to increase the accuracy of facial recognition. The MCUT dataset, which included 3755 photos divided into 276 classes, was used to test the model. Face recognition was successfully implemented, yielding 100% accuracy, 100% precision, 100% recall, and 100% F-measure.

H. N. Vu et al.(88) suggested using RetinaFace, a joint extra-supervised and self-supervised multi-task learning face detector that could handle different scales of faces, as a quick yet efficient encoder to recognize the masked face by combining deep learning and Local Binary Pattern (LBP) features. To create a unified framework for recognizing masked faces, scientists also retrieved local binary pattern features from the masked face's eye, forehead, and eyebow regions and combined them with features from RetinaFace. Additionally, they gathered 300 individuals from our university for a dataset called COMASK20. On the released Essex dataset and our own self-collected dataset COMASK20, they conducted an experiment in which they tested their suggested system with a number of cutting-edge facial recognition techniques. These results showed that their proposed system outperformed Dlib and InsightFace, demonstrating the effectiveness and applicability of the proposed method. The recognition results were 87% f1-score on the COMASK20 dataset and 98% f1-score on the Essex dataset.

R. Hammouche et al.(89) suggested a powerful face recognition system based on the Gabor filter bank and the Sparse AutoEncoder (SAE) deep learning technique. The primary goal of the suggested system was to enhance the characteristics obtained by the SAE approach when employing the Gabor filter bank. Then, using the principle component analysis and linear discriminant analysis (PCA + LDA) technique, these improved features were submitted to features reduction.Their study described that the facial image was subjected to a gabor filter in order to extract multi-scale and multi-orientation characteristics based on various scales and orientations. Then, by encoding and decoding features, the unsupervised deep learning technique known as Sparse Auto-Encoder improved the Gabor filters' features. The feature vector was still larger since SAE maps the Gabor filter feature extracted to a hidden representation and then maps this concealed representation back into an estimate of the original input. Therefore, a compact feature vector was created from SAE's output. The cosine Mahalanobis distance was then used to complete the matching stage. Experiments on seven publicly accessible databases (i.e., JAFFE, ATandT, Yale, Georgia Tech, CASIA, Extended Yale, and Essex) indicated that the proposed approach outperforms previously described methods and yields promising results when Gabor and SAE are combined.
The following table shows the summary of some different age estimation method on different dataset:

Table 2.1 Summary of Different Age Estimation Method on Different Dataset

Feature Model	**Database**	**Descriptor**	**AE algorithm**	**Performance (MAE, Accuracy)**
Texture (51)	FG-NET, MORPH	Haar-like	Regression	5.67
Texture (52)	IFDB	HOG	Group Classification	87.025%
Active appearance (49)	FG-NET, MORPH	AAM	Regression	4.67(FG-NET), 5.88(MORPH)
Active appearance (48)	FG-NET, MORPH, WebFace	AAM	Ranking	4.35(FG-NET), 4.59(MORPH), 6.03(WebFace)
Mutli-Feature fusion (54)	MORPH, FRGC	HOG, LBP, and SURF	Classification	4.25(MORPH) 4.17(FRGC)
Mutli-Feature Fusion (55)	FG-NET, MORPH II	HOG, LBP,BIP	Hybrid	2.81(FG-NET) 2.97(MORPH II)
Training from Scratch (61)	MORPH II, WebFace	Multi-task and fusion using VGG16 CNNs	Hybrid	2.96(MORPH II) 5.75 (WebFace)
Transfer Learning (68)	AFAD, CACD, MORPHII	ResNet-34 replaced the final output layer with binary tasks.	Regression	2.64(MORPH II) 3.47(AFAD) 5.25(CACD)
Transfer Learning (80)	FG-NET, MORPH	VGG, Res-Net, GoogleNet, and Alex-Net	Classification through regression	2.97(FG-NET) 2.94(MORPH)
Attention-Based (82)	CACD, FG-NET, MORPH II	Attention-based Dynamic Patch Fusion (ADPF)	Regression	2.86(FG-NET) 2.54(MORPH II) 5.39(CACD)

CHAPTER 3

Data-sets Preparation

This chapter explains the steps involved in creating age-annotated face image datasets and their properties. The most crucial phase in developing a machine learning model is acquiring an appropriate training dataset. It is difficult to obtain a perfect dataset, nevertheless, as practically every dataset has data discrepancy or an unequal sample distribution. In reality, assembling balanced, ideal aging databases is really challenging. What's more, assembling a collection of pictures of a single person at various ages is even more challenging.

Age can be experimented as a regression along with classification problem which makes it exceptional from other image classification task. Age regression is the process of estimating an individual's chronological age from a continuous variable, usually expressed in years. This method treats age as a regression problem and seeks to determine a person's precise age. On the other hand, age classification entails classifying a person into an age range. This method views age as a categorization issue and seeks to determine which age bracket the subject falls into. Different researcher takes different age group in their according their working principle. In our work, we use age as both classification and regression task.

Two datasets are utilized in our work to show our work validation: UTKFace(90) and ApaLearn(91).

3.1 UTKFace:

In studies on facial analysis, such as age estimate, gender categorization, and ethnicity recognition, UTKFace, a large dataset of face photos, is commonly employed. Over 20,000 photos make up the collection, which spans a vast age range from 0 to 116 years old. There is a wide range of poses, facial expressions, lighting, occlusion, resolution, etc. in the photographs. Dlib is used to align, trim, and provide landmarks for the photos in the UTKFace collection. Age, gender, and racial details are all listed next to each image. Age is represented as an integer ranging from 0 to 116, gender is 0 for men and 1 for women, and race is an integer ranging from 0 to 4, designating White, Black, Asian, Indian, and Others. Only non-commercial research purposes may use

this dataset.The UTKFace dataset is available to researchers and developers for non-commercial research projects. Age, gender, and ethnicity prediction models utilizing neural networks have all been developed using it in numerous studies and projects. The dataset has also been combined with other facial picture databases for use in research and comparative analysis. In general, the UTKFace dataset is a useful tool for developing and testing facial recognition algorithms, notably in the areas of ethnicity prediction, gender categorization, and age estimate.

3.1.1 Splitting Data-set:

Our working data-set consists of 23708 images of different ages peoples. The data-set have divided into three categories:

- **Training data:** In labeled data-set training data is presented in supervised learning.When it gives wrong prediction, model parameters updates based on the loss from the actual data-sets.But, in unsupervised learning it is difficult.
- **Validation data:** It gives the performance of trained model.If a model give a good validation result, then it is expected that the model will work on uncertain situation with great performance.
- **Testing data:** Testing data is used to test the performance of proposed model.

We divide the dataset into training, testing and validation task with 80% and 20% of whole dataset.Thus, training data is composed of 15172 samples, testing data are of 4742 samples and validation data is of 3794 samples. We divide the dataset in such a way that data distribution is even across different ages people. Fig 3.1 depicts the training data distribution: Fig 3.2 gives the training sample distribution after labeling.

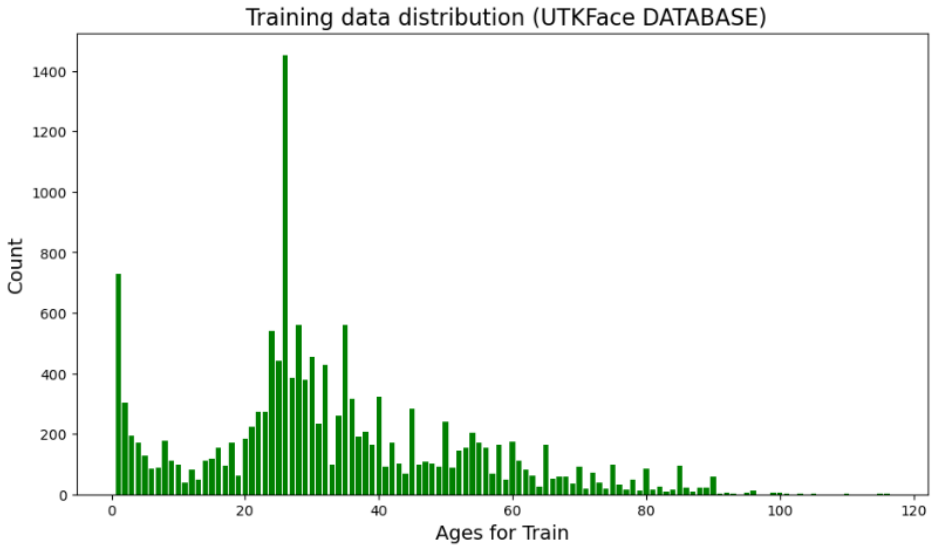


Fig. 3.1 Training data distribution of UTKFace

Fig 3.3 depicts the overall age distribution of UTKFace dataset. From this figure we see

that our train dataset, test dataset and validation dataset follows the original distribuion of the UTKFace. Some sample images of UTKFace datset corresponding to differnet classes are shown in Fig 3.4. Due to the uneven age distribution, varied picture brightness, and irregular image placements, this data set is quite difficult and can be a great benchmark for assessing the efficacy of the age estimation method.

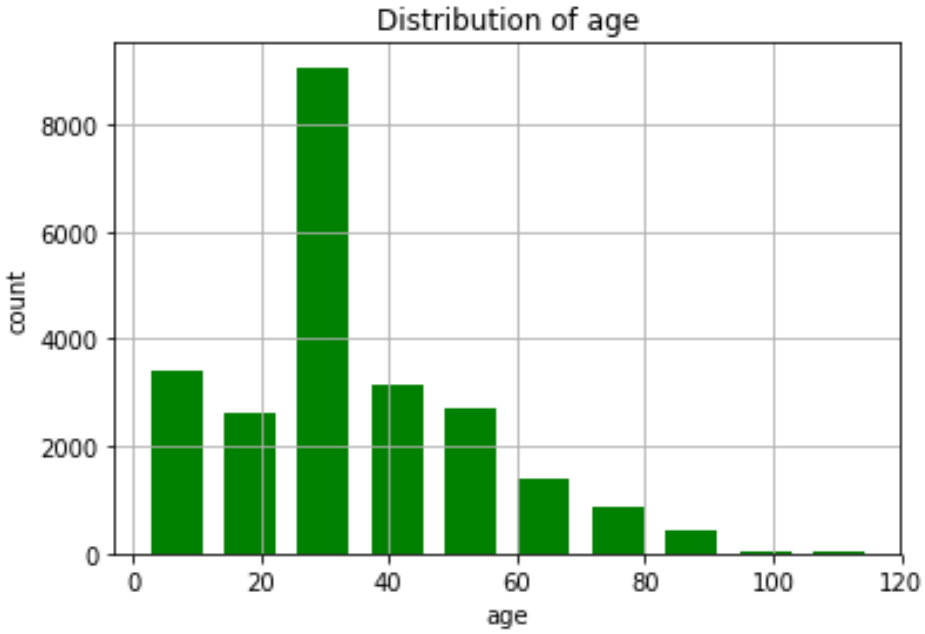
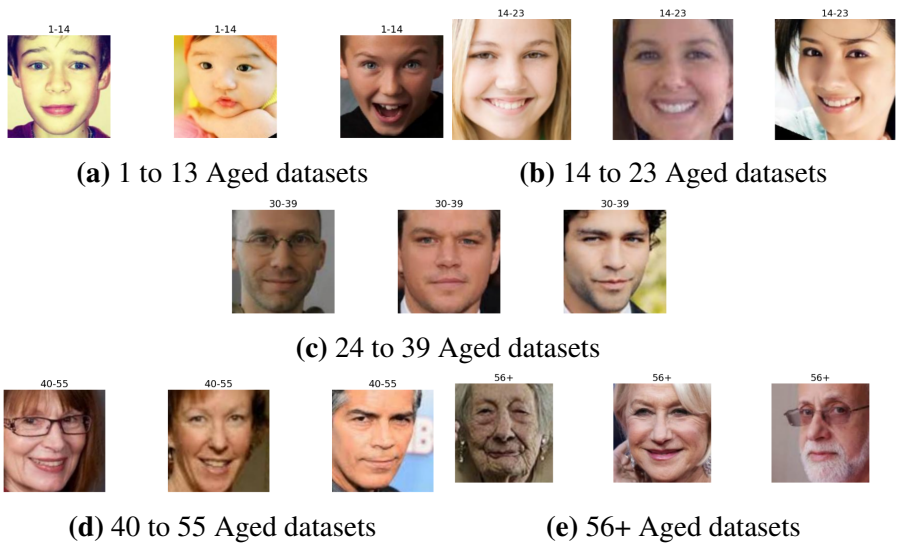


Fig. 3.2 Age distribution of UTKFace



(a) 1 to 13 Aged datasets **(b)** 14 to 23 Aged datasets

(c) 24 to 39 Aged datasets

(d) 40 to 55 Aged datasets **(e)** 56+ Aged datasets

Fig. 3.3 UTKFace datasets

3.2 Common Voice:

The Common Voice dataset (92) is a multilingual and openly available dataset of human voices that is created and maintained by Mozilla, the organization behind the Firefox

web browser. The purpose of this dataset is to train and develop speech recognition and voice-related machine learning models that can understand and generate human speech across various languages. The Common Voice dataset is unique because it is crowd-sourced, meaning it is created by volunteers who contribute their voices by reading sentences provided by the dataset creators. This collaborative effort allows for a diverse and extensive collection of voice samples from different regions and speakers with various accents and pronunciations. The dataset includes audio clips of individuals reading out short sentences in their native language. Additionally, it also provides accompanying text transcriptions of the spoken sentences. The Common Voice dataset is made available under a Creative Commons license, making it freely accessible for research and development purposes. For convenience, the corpus has been divided into different sections. The subgroups with ”valid” in their names are audio clips that have at least two listeners who, on average, agree that the audio matches the text.
However,we employ the English subset and valid-other-train sectio for our work. The recordings are valid (two upvotes), invalid (two downvotes), and other (upvotes = downvotes). Additionally, train, dev, and test subsets are created from the valid footage and other clips. For training, development, and testing, we made use of the split-up, legitimate footage. Each subset includes a.csv metadata file that lists details about the audio clips, including filename, age, gender, accent, and clip length. Furthermore, according to the dataset description, the ages are split into eight groups, each of which represents a decade. In addition, the fact that a speaker who occurs in the training subset is not featured in the other two subsets or in the text is a noteworthy aspect of this dataset. As can be seen in Fig 3.4, a total of 74885 audio clips in mp3 format were taken from this dataset.

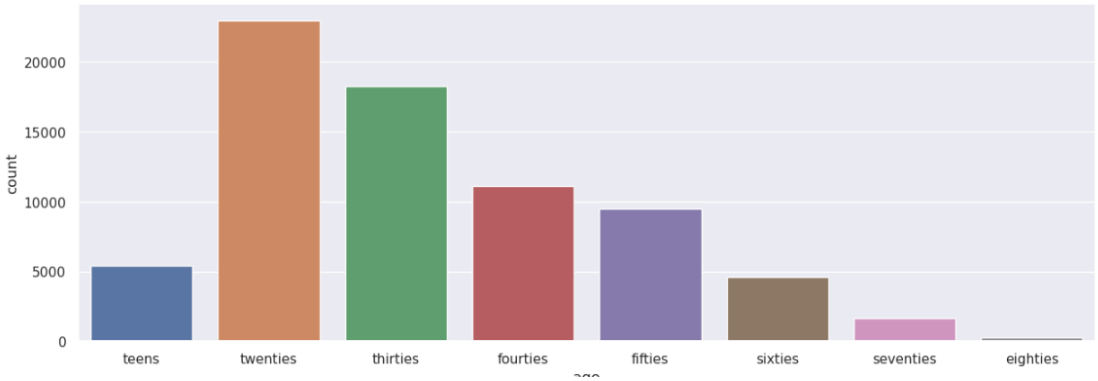


Fig. 3.4 Age distribution of Common Voice dataset

CHAPTER 4

Methodology

This chapter describes the methodology and description of the related component of our work.

4.1 Proposed Methodology:

The main goal of this work is to propose and evaluate a hybrid model for real age estimation. The proposed method consists of two stages: the first stage involves extracting features from pictures simultaneously utilizing Resnet50 and vision transformer. Then, age is determined using mlp. This section provides a detailed depiction of each phase of the suggested strategy indicated in Figure 4.1. which shows the overall pipeline of our proposed method Age estimation task is implemented by first detecting the face from

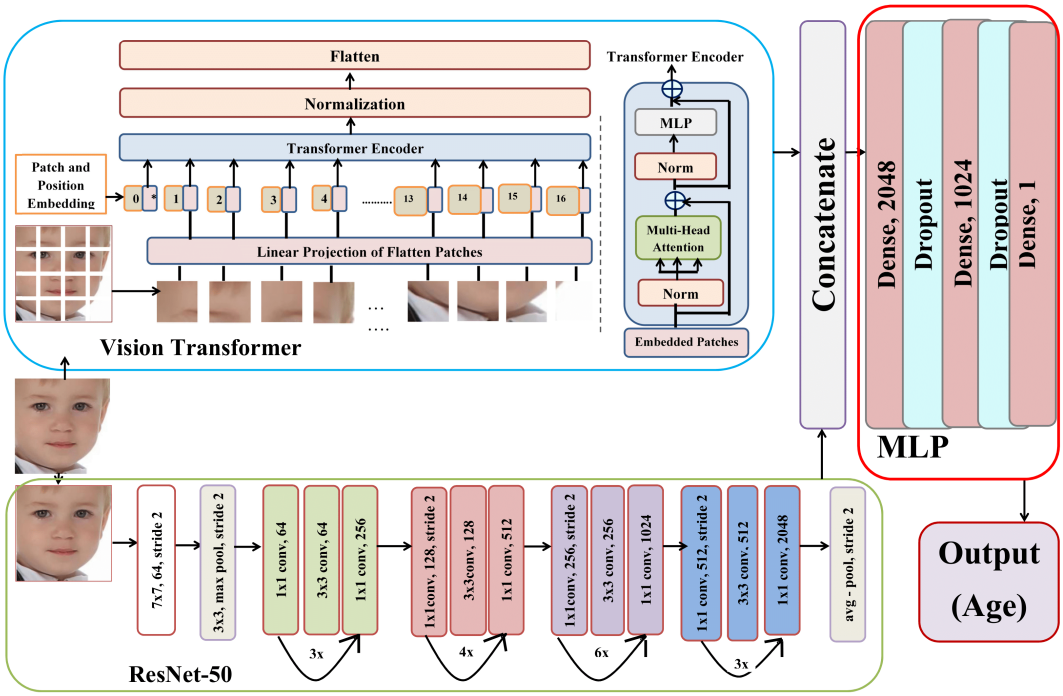


Fig. 4.1 Illustration of the proposed architecture

whole image. ResNet50 is a well-known convolutional neural network (CNN) architecture renowned for its superior feature extraction skills in image recognition applications

that can effectively captures low-level details including edges, textures, and forms. The Vision Transformer model, on the other hand, makes use of the attention mechanism to identify long-range dependencies and the overall context, allowing it to recognize high-level relationships and patterns in the image. Thus, we gain from both the local and global feature representations by integrating the two models, providing a more thorough knowledge of the face traits associated with aging. So, with increased accuracy, robustness, and flexibility to different vision tasks, the combination of ResNet50 with Vision Transformer offers a potent ensemble solution for age estimate from face pictures. Its success may be ascribed to how well the two models complement one another and how well they can efficiently capture both local and global elements. Besides these two, we also evaluate other features on age estimation task such as HOG features, manually feature concatenation of HOG-resnet50, HOG-resnet50-vision transformer. We also evaluate age estimation from voice using Mel-frequency cepstral coefficients (MFCCs) as features extracted from voice which presents a short-time power spectrum envelope and closely mimic the human voice tube. The final model will be a comparatively better approach with performance to estimate age from a facial image. Fig 4.2 shows the flow diagram of the proposed method to estimate age from real-times camera and images:

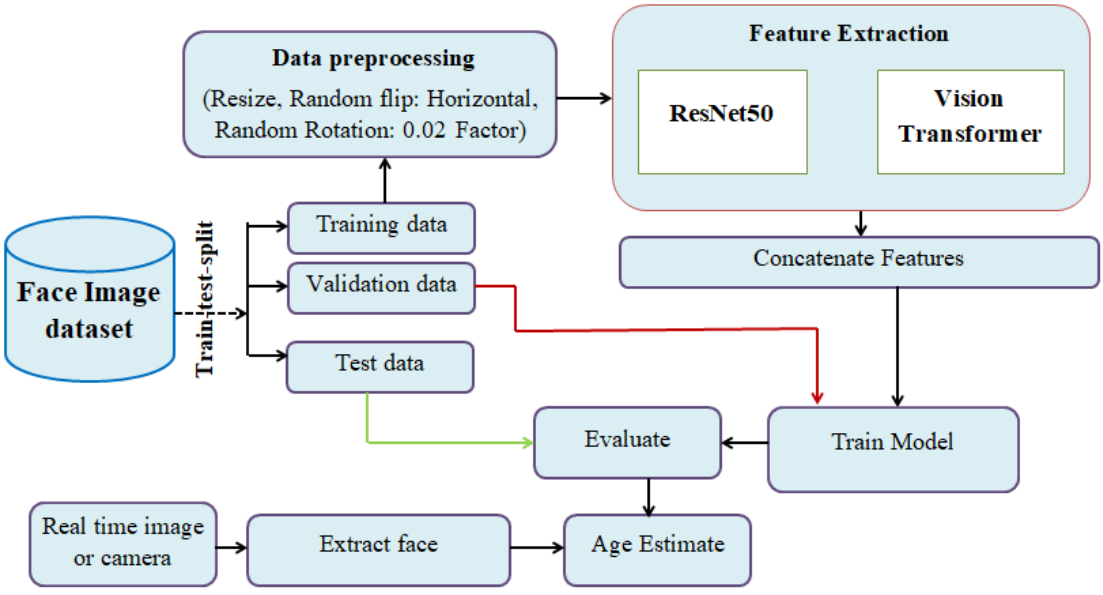


Fig. 4.2 Flow diagram of the proposed method

4.2 Facial Image Pre-processing:

Ageing is a natural process that is characterized by the appearance of wrinkles on the surface of the face and modifications to the facial tissues at their core brought on by gravity, sunlight exposure, and bone remodeling. Moreover, facial traits are essential for face identification, and the human brain has developed specific regions for studying

facial images. Although the accuracy of facial recognition has substantially improved, it is still not always reliable. Several factors, such as lighting, position, facial expressions, and image quality, influence how well an image may be classified. Besides, strong variations in stance and lighting pose the biggest threat to the effectiveness of facial recognition systems. In general, there is less variation between photographs of different faces than there is between images of the same face taken in various settings. For example, systems based on comparing photographs may incorrectly identify the input image because changes brought on by illumination may be greater than individual differences the variations between photographs of a single face taken under various lighting settings are bigger than the variations between photographs of other faces taken under the same lighting conditions.However, human face detection is complicated work as there is a vast variety of appearance and external and internal factors. Here are the utilized methods for pre-processing facial images of our work is given:

4.2.1 Face Detection and Alignment:

The first stage in many computer vision applications is face identification, which is especially important for facial analysis tasks like age estimate. Finding and identifying human faces in an image or video frame is the aim of face detection. It is a crucial stage in age estimate systems because it enables the extraction of interest face areas, which contain characteristics that may be used to predict age. However, dealing with different appearance variations and outside elements like position, lighting, occlusions, and complicated backdrops makes face identification difficult.In order to identify faces based on contrast and edge patterns, traditional face recognition algorithms like Viola-Jones and Haar cascades use characteristics such rectangular filters that are similar to those in Haar and a cascade classifier.This study makes use of a facial landmark detector (93). The eyes, brows, nose, mouth, and jawline are among the prominent facial areas that are localized and represented using facial landmarks. In order to predict the position of landmarks from a sparse set of pixel intensities, a fully discriminate model based on a cascade of boosted decision forests can be used, and it typically yields correct landmarks. Face detection allows for the realization of facial alignment. The CNNs can handle such a high degree of accuracy and can tolerate little alignment problems. The alignment is only a transition from one coordinate system to another that centers all the faces, places the eyes on a horizontal axis, and scales the faces to be almost similar in size. Compared to Haar cascades or HOG detectors, facial landmarks perform better for face alignment since the bounding box used to determine the location of the eyes was less precise. In our work, First, we used a facial identification method to identify the faces in each dataset by automatically locating them and localizing them by drawing a four-sided bounding box in the image. By utilizing the Facenet library to retrieve the

top, left, bottom, and right coordinates of the face, it can capture characteristics like the mouth, nose, and eyes. An image is cropped out if a face can be seen in it; else, the image is deleted. This technique was chosen because it can effectively manage occlusions, handle pictures with different orientations, and recognize faces at different sizes. After identifying the faces, we use the same Facenet to identify the landmarks and align the faces. The program chooses the faces' spatial positions for the important annotations. The facial characteristics are determined by measuring the distance between these two places after determining the face key point. The facial key point are used to take face photos from various angles to match the characteristics retrieved for face alignment. These whole works are done as an build in works of the datasets which we are used in our work. UTKFace have face alignment-cropped version by the developer of these datasets. From kaggle the preproceesed version of this dataset are used in our work. Then, we make three separate numpy array for training,testing,and validation dataset. We do this to avoid data inconsistency between training models.

4.2.2 Normalization and Standardization:

Image normalization is the process of bringing an image's pixel values into a uniform range or distribution. In order to improve visual comparability and consistency, deviations brought on by elements like lighting, contrast, and color channels have to be minimized. The particular normalization methods employed might change based on the job and needs. To place all values in the new range of 0 to 1 or -1 to 1, normalization entails rescaling the original range pictures. The Min-Max normalisation retains the link between the original pictures while offering linear transformation on the original range of images. However, the input pre-processing that each Keras Application anticipates varies.For example, preprocessing our inputs for Xception before providing them to the model need to be the input pixels will be scaled between -1 and 1. Again, EfficientNet includes input preparation as a Rescaling layer, preprocessing input method of EfficientNet is basically a pass-through function. The inputs for EfficientNet models are float tensors of pixels with values in the [0-255] range.In addition, Before providing our inputs to the model for ResNet, prior to zero-centering each color channel with regard to the ImageNet dataset without scaling, resnet preprocess-input method converts the input photos from RGB to BGR. In our work, we had to pre-process each dataset individually to meet the different requirement of each CNN model. Thus, we rescale the images while doing work on xception network, convertion to BGR are performed in time of Restnet Network.

4.2.3 Histogram Equalization (HE):

By modifying the intensity distribution of an image, the histogram equalization (HE) approach used in digital image processing can enhance the contrast of a picture. In order to make it easier to perceive details in a picture, it is important to increase an image's look and contrast. When estimating age from facial features, HE could be used to correct the effects of uneven lighting, shadows, and other differences in image quality that might make it challenging to determine age with accuracy. The appearance of wrinkles, textures, and other aspects on the face that are crucial for determining facial age can be improved with HE. While HE can be helpful in a variety of situations, depending on the particular dataset and task requirements, it may not always be the best option for face age estimate. In some circumstances, HE can intensify image noise and artifacts, thereby degrading performance. Additionally, it might change the relative intensities of certain face regions and features, which might not be ideal for age estimate. These effects could also have an impact on the overall aesthetic of the image. Fig 4.3 shows the effect Histogram Equalization(HE) on five images of UTKFace Dataset.

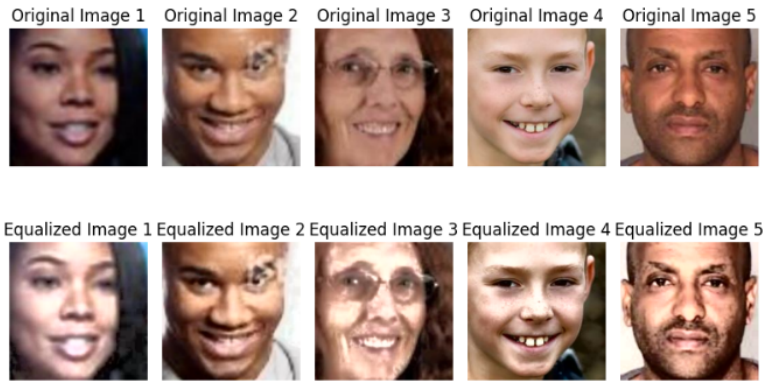


Fig. 4.3 Effect Histogram Equalization on UTKFace Dataset

4.2.4 Filtering:

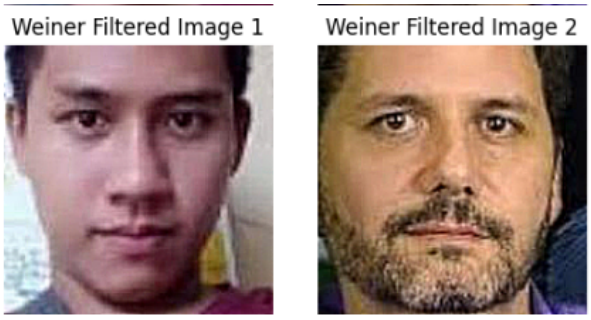
The digital photographs are constantly filled with noise. An undesirable aspect of the picture is noise. The accuracy of face recognition can be compromised by the presence of noise in a facial picture. Therefore, we want a suitable solution that can handle noise or improve image quality. Smoothing (filtering) is the greatest way to get rid of picture noise.In addition, numerous efforts have been made to address the issues with lighting, stance, and emotion, and they have produced some remarkable outcomes. However, the majority of techniques' identification accuracy would dramatically decline for noisy pictures. When being acquired, quantized, compressed, or transitioned, a face picture is susceptible to noise. Additionally, there are situations when a human may find it challenging to identify someone from an extremely loud face. Noise significantly impacts data usefulness and algorithm recognition, including Gaussian, uniform, and salt

and pepper noise. So, to denoise the image before the identification step, several ways have been suggested. When pictures are smoothed using the linear approach known as ”Gaussian filtering,” high-frequency noise is reduced but the structure of the image is preserved. By substituting median values for pixel values, median filtering effectively eliminates impulsive noise without distorting edges or fine features. While non-local means filtering employs patches to estimate the clean picture and reduce noise while keeping fine textures and features, Wiener filtering estimates the original image via deconvolution. Fig 4.4 shows the effect of filtering technique on some images of UTK-Face datset: We need to locate a typical image that has a particular size, location, and most crucially, background-free. Face detection is a common first step in age estimation methods. On the data set, we are working. UTK-Face includes the necessary, appropriately shaped human faces.As we are using transfer learning methods in our work, we resized the images of each dataset into 224*224 sized. As,we are using cropped and moderate proproccesd version in our work, we do not need to do alignment and cropping work. Because, experimental analysis showed that after pre-processing of these, the performance are significantly reduced. However, To extract just correct faces from a whole face picture, a variety of frameworks, algorithms, tools, and technological advancements are available. Examples include Haar Cascade, Deep Neural Networks DNN, and Histogram of Oriented Gradients (HOG). Some techniques also offer segmentation for detecting human faces. These utilities include (94), (95), and (96).

4.2.5 Data Augmentation:

Data augmentation is a strategy used to artificially expand the amount and variety of a training dataset by making multiple adjustments or alterations to the current data. It is commonly used in deep learning and machine learning tasks such as object recognition, image classification, and natural language processing. By injecting differences in the training data, data augmentation aims to increase the generalization and resilience of machine learning models. The model can become more invariant to multiple modifications and improve its capacity to handle real-world circumstances by being exposed to a greater variety of augmented instances. By creating additional data points from existing data, data augmentation includes growing the training dataset. Small adjustments or deep learning models are used to create fresh data points, decreasing overfitting, enhancing model performance, and cutting expenses and time-consuming processes. In computer vision applications including picture classification, object identification, and facial recognition, data augmentation is frequently used to overcome issues with minimal labeled data and enhance models' capacity to handle changes in real-world circumstances. In our work, we see from Fig 3.2 and 3.6 that we have a imbalanced datset for classification task. This can bias the performance of our model. To prevent

(a) Gaussian filtering



(b) Wiener Filter

(c) Non-local Means Denoising

(d) Median filtering

Fig. 4.4 Effect of different filtering technique on UTKFace.

for imbalance between classes while classification, we use data augmentation in our. Experimental analysis showed that without data augmentation, our model performance can not crossed over 50 Percent of accuracy. But Augmented data can results in higher accuracy. We looked at several techniques to see how overfitting affected classification accuracy and how augmentation may help. Because a picture may have been shot from the right or left, as opposed to a vertical flip, which could not be appropriate, we used a Random flipping set to the horizontal flip. The likelihood that the model will recognize an upside-down image is low. During training, we also used the Random rotation to adjust the object's angles in our dataset, setting it at 0.2.We added random enhancements to the photos and submitted the results into the training model. Fig 4.5 depicts some facial image after applying data augmentation:

Fig. 4.5 Data Augmentation on UTKFace Dataset

4.3 **Machine Learning:**

Humans pick up information from their experiences and the world around them, steadily expanding their knowledge over time. Similar to this learning process, machine learning (ML) is a system that uses past data to duplicate it. Based on the facts, it builds mathematical models and uses those models to forecast the future. Applications for machine learning may be found in a variety of domains, including voice recognition, recommendation engines, and picture identification, among others.

The three primary categories of machine learning are reinforcement learning, unsupervised learning, and supervised learning. Classification and regression are two forms of supervised learning. In classification, datasets include annotated samples that help the model understand the distinctive qualities of each type of object. On the other hand, regression concentrates on forecasting continuous numerical values based on input data.

While a supervised learning model is trained using data from the training set, its performance is evaluated using a different collection of data called the test set. A validation set is frequently used as well to adjust the model's parameters and enhance performance. Unsupervised learning, on the other hand, doesn't use labeled data. Instead, it seeks to elucidate the data's buried patterns and insights. This kind of learning simulates how people learn in ambiguous or unstructured environments. Unsupervised learning frequently involves the job of clustering, which requires assembling related data points. Another unsupervised learning method called association focuses on identifying connections and links between various dataset characteristics.There are various ways to put machine learning algorithms into practice. Some of the often employed algorithms are Random Forest, Linear Regression, Logistic Regression, Decision Trees, Support Vector Machines (SVM), Naive Bayes, K-Nearest Neighbors (KNN), K-Means Clustering, and Principal Component Analysis (PCA). These algorithms make it possible to create high-performance models that may be used to successfully address certain issues. An ensemble learning system called RandomForestClassifier mixes various decision trees to provide predictions. It provides accurate age forecasts by capturing the intricate correlations between face traits and age labels. A straightforward yet effective technique called DecisionTree Classifier builds a decision tree from training data, with each internal node representing a test on a particular characteristic. A learning technique called KNeighbors Classifier uses the k nearest neighbors in the feature space to predict the age of a given face. Gaussian distributions and the Bayes theorem are used by GaussianNB to estimate the likelihood that a face falls into each age group. A support vector machine, or SVC, divides facial characteristics into several age groups and forecasts the age of fresh faces. An artificial neural network called MLP Classifier learns patterns. From literature review we can see SVM is the most ML algorithm on facial age estimation. SVM is a well-liked classification technique that may also be used to solve regression issues like age estimation. The age of fresh face photos may be predicted using the hyperplane that SVM creates to divide data points into various age groups. In our work, we use these algorithm to evaluate machine learning performance on facial age estimation considering whole face at a whole.

4.3.1 Histogram of Oriented Gradients (HOG):

A feature descriptor called the Histogram of Oriented Gradients (HOG) is used in computer vision and image processing to identify objects.The extraction of face characteristics using HOG (Histogram of Oriented Gradients) features is a frequent method. The idea behind it is that a gradient orientation histogram may accurately depict the distinctive characteristics of the human face. HOG features are created by computing gradient orientations on discrete face spatial areas, which are then histogrammed to create HOG

feature descriptors. By taking the appropriate elements out of the face picture, HOG features may be utilized to estimate facial age. An artificial intelligence (AI) model that can estimate an image's subject's age may be trained using the retrieved attributes. The accuracy, efficiency, and robustness of HOG features may be compared against those of other feature extraction methods. On a shared dataset, comparative studies can be used to assess how well various feature extraction methods work. Other well-known feature extraction methods like LBP (Local Binary Patterns) and SIFT (Scale-Invariant Feature Transform) can be used to compare the performance of HOG features.Using a neural network, these various characteristics were utilized to estimate age and gender.
However, by segmenting the face picture into tiny, overlapping blocks, HOG characteristics are calculated in the context of estimating facial age. The pixel intensities for each block are determined as gradients, and these gradients are subsequently quantized into histogram bins according to their orientations. The local texture information contained in that block is represented by the resultant histogram of gradient orientations. In order to create the final HOG feature vector for the facial picture, these histograms are concatenated across all blocks. HOG characteristics capture the varying face texture patterns, including wrinkles, fine lines, and facial contours, which can be a sign of aging-related alterations. HOG features offer a concise description of the textural properties of the face picture by examining the distribution of gradients and orientations. It is a popular technique used in facial image processing for feature extraction and age estimation (46). The collected HOG characteristics are used as input for machine learning techniques like support vector machines (SVM) or random forests in order to estimate age (97), These algorithms provide the capacity to relate the patterns and distributions of HOG properties to different age groups. In our work,we use HOG features extracted from whole image. These features will be used to evaluate the machine learning performance properly.
Fig 4.4 shows histogram images of two images. Using the HOGDescriptor from 'opencv2', we extract the histogram feature vector from the pictures. Here, we utilized a winsize of 224 by 224, a block size of 8 by 8 by 8 cells, and a block stride of 4 by 4.We calculate the horizontal and vertical gradients of the pixels using an 8x8 pixel cell. compressed into 9 vectors (each from 0 to 360 degrees with 40-degree bins), in order to produce a histogram. The main idea behind the HOG descriptor is that the distribution of intensity gradients or edge directions (or orientations) may be used to define local entity appearance and form within an image. A histogram of gradient axes is produced for each cell's pixels, which are broken up into tiny linked areas called cells. All of these histograms come together to form the descriptor.

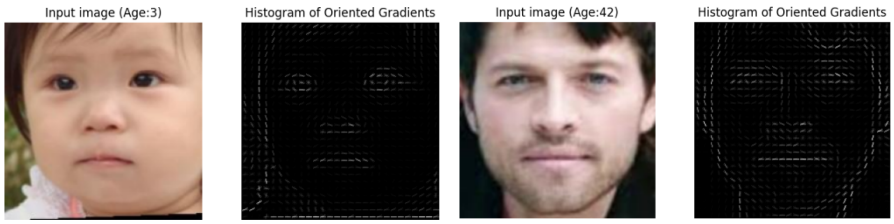


Fig. 4.6 HOG Images

4.4 **Deep neural networks:**

An artificial neural network that has numerous layers of linked nodes, or neurons, is called a deep neural network (DNN). Through various degrees of abstraction, these networks are intended to learn and extract hierarchical representations of data. The output of the layer before it is processed by the subsequent layer in a DNN, which enables the network to learn intricate patterns and connections in the data. Besides, the science of computer vision has undergone a revolution as a result of the considerable influence that deep neural networks (DNNs) have had on picture categorization tasks. It takes a lot of computing to process images at the pixel level in the RGB color space. For this, the majority of the study is done on a facial image in grayscale. However, due to the advancement of deep learning methodology, its many applications, and degree of face feature learning capabilities, it is now commonly employed for the age detection issue area. DNN may also pick up on elements that remain constant throughout changes to the face. Moreover, with the aid of cutting-edge methods like convolutional neural networks (CNNs), deep learning algorithms are frequently utilized for image categorization tasks. CNNs are particularly excellent in classifying pictures because they can scan images via convolutional layers to discover intricate patterns and characteristics. A CNN's intermediary layers may also be seen to better understand how the network is handling the picture. The deep neural network designs AlexNet, VGGNet, and ResNet are some of the frequently used ones for image categorization. To avoid overfitting in neural networks, regularization methods like dropout and weight decay can also be applied. On the whole, deep neural networks have shown outstanding results on a range of picture categorization challenges, sometimes outperforming human accuracy. DCNN is the model that is most frequently employed to estimate facial age, according to a review of the literature. the question of classification vs. regulation.
Convolutional Neural Networks (CNN) is the most significant deep learning approach for its extending capability and have been used for an image-based application like face detection, object detection, image classification, and so on.

4.4.1 Convolutional Neural Network:

Convolutional Neural Network is referred to as CNN. It is a kind of deep learning system made for processing and analyzing grid-like organized or unstructured data, including sequential data like audio or text, movies, and photos. Because they can automatically learn and extract useful characteristics from unprocessed pixel input, CNNs are particularly effective in computer vision applications. The design of a CNN is modeled after how neurons are grouped in layers in the visual brain of animals to detect more complex information. CNN is used mainly for image processing task. A color image contain RGB combination. Computer see these image as a three dimensional array corresponding its height, wide and last one for RGB colors in pixels. There are three layers in a CNN architecture:

1. **Convolutional array:** It is core building block of a CNN. Multiple convulutional layer follow each other involving a kernel or filter to check a image contain a feature or not. Image is converted to numerical values in this layer.
2. **Pooling layer:** It works like convolutional layer but takes less parameter, dimension of feature map to reduce complexity.
3. **Fully connected layer:** This layer classify a image based on extracted feature in previous layer.

These hierarchical layers learns features automatically and update through back propagation. CNN give high performance pr-defined model with more hidden layers with different capacity to learn features. Some of them are RestNet, VGG16, MobileNet and so on. We use these high performance model in our work to detect high performance among plain CNN and these models.

4.4.2 Transfer Learning:

A machine learning approach called transfer learning uses information gained from one job or domain to enhance performance in a related task or domain. A model is initially pre-trained in transfer learning utilizing a big dataset and a specific goal, frequently using a supervised learning methodology. The model can learn broad characteristics and recurring patterns from the data during this pre-training stage. The model is then fine-tuned on a smaller dataset relevant to the intended task or topic following pre-training. The notion is that by applying the information and representations acquired during pre-training to the new task, the model will generalize more effectively and perform better with less training data. Transfer learning may be done in a variety of ways, depending on the particular architecture and methodology employed. Freezing the previously learned layers and just training the final levels or new layers added especially for the current job is one typical strategy. This enables the model to adapt to the new job while

maintaining the learnt features. In fields where labeled training data is scarce or expensive to get, transfer learning has shown to be particularly successful. Transfer learning can minimize the quantity of labeled data needed to obtain good performance on the target task by beginning with a pre-trained model. The ability to transfer information and representations learnt from one activity or domain to another is made possible by the strong mechanism known as transfer learning.

4.4.2.1 RestNet50:

The ResNet (Residual Network) family of convolutional neural networks includes the ResNet-50 design. It was presented by Microsoft Research in 2015, and since then, the computer vision community has greatly benefited from its widespread use. The network's layer count is indicated by the "50" in ResNet-50. It is a pretty deep network with 50 convolutional layers. Because of the network's depth, it can extract complex and hierarchical information from pictures, which improves representation learning. ResNet-50's architecture is organized in a modular fashion. It is made up of blocks that are repeated and comprise several convolutional layers. Each block has a residual unit that consists of a collection of convolutional layers and short-cut connections. With the help of the residual units, the network is able to learn residual mappings, or the discrepancy between the intended output and the present forecast. These residual units are stacked, allowing the network to gradually improve its predictions and recognize more intricate patterns. Originally created for image classification tasks, ResNet-50 has displayed astounding performance on benchmark datasets like ImageNet. By adding further modules or layers on top of the fundamental ResNet-50 design, it has also been modified for various tasks including object identification and picture segmentation. Overall, ResNet-50 is a deep convolutional neural network architecture that employs residual connections to enable the training of deeper networks. It has proven to be highly effective in various computer vision tasks, offering excellent performance, strong generalization capabilities, and faster training convergence compared to earlier network architectures. ResNets (98) rely on the notion of skip connections to get around these problems and enable researchers to delve extremely deeply into neural networks without any concerns. Identity mapping may be used to make a path between the input and output layers while bypassing certain layers in between to build a ResNet module. Fig. 4.4's representation of the ResNet-34 design can help us to grasp the framework. In comparison to Inception and VGG-16, ResNet-34 is observed to have a top 5 error of 5.7%, which is much less. ResNet-50 achieves the best performance on the top 5 errors, with a value of 4.49%. The top 5 errors, as determined by results on 'imagenet'. Many computer vision applications, including object identification and picture segmentation, employ the deep neural network ResNet50. With 48 convolutional layers, 1 maximum

pool layer, and 1 average pool layer, it is a ResNet model version. There are 3.8 x 109 floating point operations available. Based on a residual network design, skip connections are used to get around the vanishing gradient issue. Additionally, it employs a bottleneck design that stacks three layers rather than two in each block to save training time and boost accuracy.

4.4.2.2 Xception:

Convolutional neural network (CNN) architecture called Xception, sometimes known as ”Extreme Inception,” was developed by François Chollet, the brains behind the Keras deep learning package. The Inception architecture, which was renowned for its efficiency in image recognition tasks, was extended and improved upon by this proposal. The usage of depthwise separable convolutions is the main concept of Xception. Traditional convolutions incur a high computational cost since each input channel is convolved with a unique set of filters. On the other hand, depthwise separable convolutions divide the convolution process into two distinct steps: a depthwise convolution in which each input channel is convolved independently, followed by a pointwise convolution in which 1x1 filters are used to combine the depthwise convolution's outputs. In order to preserve expressive strength while reducing computational cost, Xception uses depthwise separable convolutions. Because of this, Xception is appropriate for a variety of computer vision tasks and enables more effective and quick training and inference. The ”entry flow” and ”middle flow” modules make up the ”stacked modules” that make up Xception's design. Convolutions and pooling operations are carried out by the entry flow module in order to extract low-level features from the input picture. With a succession of repeated blocks, the center flow module enhances these traits even further. The usage of ”skip connections,” which is a feature of the ResNet design, in Xception is noteworthy. These skip connections give gradients shortcuts that make it easier for them to spread during training, improving optimization and resolving the disappearing gradient issue. On a number of picture recognition benchmarks, especially the ImageNet dataset, Xception has achieved state-of-the-art efficiency compared to older approaches. The Xception network design creates an effective and potent CNN model by fusing the advantages of depthwise separable convolutions, skip connections, and a multi-level hierarchical structure. It is a popular option for many computer vision applications because to its effectiveness in delivering high accuracy while minimizing processing complexity.

4.4.2.3 InceptionV3:

As a member of the Inception family of models, InceptionV3 is a convolutional neural network architecture created by Google researchers. It builds upon and enhances InceptionV1, also known as GoogleNet, its predecessor. To improve performance on picture identification tasks and fix some of InceptionV1's shortcomings, InceptionV3 was created. InceptionV3's usage of the "Inception module," which is made up of several parallel convolutional branches with various filter sizes, is its distinguishing characteristic. By efficiently managing both fine and coarse details in the input image, the model is able to capture features at various scales and resolutions. To extract features at various levels of abstraction, InceptionV3 combines 1x1, 3x3, and 5x5 convolutions inside each module. InceptionV3 uses a method known as "factorization into smaller convolutions" to lessen the network's computational complexity. It applies several smaller convolutions rather than one large one, such as 3x3 and 5x5, then 1x1 convolutions. By using this method, the computational complexity and number of parameters are decreased but the network's expressive capability is preserved. InceptionV3 also offers a method known as "dimensionality reduction," which use 1x1 convolutions to cut down on the quantity of input channels before using bigger convolutions. In doing so, the network is able to lessen the computational load while still collecting crucial information. The use of "auxiliary classifiers" during training is another crucial component of InceptionV3. The gradient flow during backpropagation is assisted by these auxiliary classifiers, which are positioned at the network's intermediary layers. As a result, generalization is enhanced and overfitting is decreased. They also encourage the network to acquire additional discriminative features and to offer regularization. InceptionV3 outperformed its predecessors and several other designs in a number of image classification benchmarks, including the ImageNet dataset. It has also been effectively used for other tasks, such object identification and picture segmentation, by modifying the underlying architecture with extra elements like region proposal networks (RPN) and region of interest (RoI) pooling. All things considered, InceptionV3 is a potent convolutional neural network design that makes use of the Inception module to collect multi-scale data and enhance computational effectiveness. It continues to be heavily utilized and important in the field of deep learning and has shown good performance on a variety of computer vision tasks.

4.4.2.4 VGG16:

The University of Oxford's Visual Geometry Group (VGG) created the convolutional neural network (CNN) architecture known as VGG16. It is called after the group and the 16 weight layers (13 convolutional layers and 3 fully linked layers) that make up the structure. Because of its simplicity and homogeneity in design, VGG16 is well regarded

for being simple to comprehend and use. Besides, the primary principle of VGG16 is to downsample the spatial dimensions by stacking a number of small-sized convolutional filters (3x3), followed by max pooling layers. The network can successfully catch local patterns thanks to the tiny filter size, and the model can learn more intricate and abstract characteristics thanks to the stacking of these filters. With a fixed filter size of 3x3 and a stride of 1, the sequential architecture of VGG16 stacks convolutional layers one after the other. Periodically, the spatial dimensions of the feature maps are condensed using max pooling layers with a stride of 2. As we delve farther into the network, the number of filters grows steadily, enabling more expressive capability. he feature maps are flattened and routed through numerous convolutional and pooling layers before going through three fully connected layers and a softmax activation for classification. As a classifier, the fully connected layers translate the learnt characteristics to the relevant class probabilities. VGG16, which has 16 layers altogether, is renowned for its depth. The network can capture a wide variety of characteristics thanks to this depth, however the model is computationally costly to train and infer due to this depth. VGG16 is more prone to overfitting because to its large number of parameters, especially when the amount of training data is constrained.In addition, VGG16 has been modified and used for additional computer vision tasks, including object identification and segmentation, even though it was first developed for image classification tasks. To further boost the model's capability, variants of VGG16, such as VGG19 with 19 layers, have been created. Including the ImageNet dataset, VGG16 has demonstrated good performance on a number of image classification benchmarks. Its efficiency and simplicity have made it a popular option for transfer learning, when a model that has already been trained on a large-scale dataset is adjusted for a particular job with less labeled data.

4.4.3 Vision Transformer (ViT)

With the help of the Vision Transformer (ViT), a deep learning architecture, the Transformer model, which was developed for natural language processing (NLP), is now applied to computer vision. By utilizing self-attention processes to identify spatial connections in pictures, it marks a shift from conventional convolutional neural networks (CNNs). The main concept is to split an input image into patches and consider each patch as a token. Then, after being linearly embedded, these patches are processed via many layers of Transformer encoders. The self-attention mechanism enables the model to pay attention to various patches and record the interactions between them, allowing the model to comprehend overall linkages within the image. ViT uses positional embeddings, which are like the token embeddings in NLP Transformers, to capture positional information. The relative locations of the patches in the image are known to the model thanks to these embeddings. As part of its pretraining process, ViT also

includes ”pretraining with large-scale datasets.” In this phase, the model is pretrained on sizable datasets like ImageNet using a proxy job termed ”image inpainting” or ”image masking.” With smaller labeled datasets, the pretrained model is then refined for use in downstream tasks like object identification or picture classification. The Vision Transformer has a drawback in that it relies on patching the picture together, which can restrict its capacity to capture fine-grained information in comparison to CNNs. Hybrid models have been proposed to alleviate this drawback, in which a CNN is employed as a feature extractor for low-level visual data before being sent into the Transformer for higher-level processing. The Vision Transformer has demonstrated promising results despite its drawbacks and has achieved competitive performance on a number of computer vision benchmarks, including picture classification tasks.

4.4.4 Regression and Classification:

Regression is used to define the relation between input and output variables. It mainly works on continuous data likes: Market trading, weather forecasting, and so on. Some regression algorithms are Linear regression, polynomial regression, bayesian regression. On the other hand, a classification algorithm is used when output data are explicitly classified. For example, Male-female, Yes-No, and so on. Some classification algorithms are random forest, decision tree, Logistics regression, and Support vector Machine(SVM).

Age detection problem is under regression problem because age shows a continuous value. Given that age is a continuous variable, ideally a collection of discrete classes, age prediction falls within the heading of a regression issue. The pre-trained commercial models used for ImageNet classification are made up of neurons for each type of object and are normalized using the S]softmax function. But (2) says that if age is considered as a regression problem, their proposed model would show uncertain prediction. Unfortunately, the instability while managing outliers causes a significant error rate when training a CNN only for age regression tasks. Consequently, due to steep gradients and unreliable prediction, the model’s convergence was challenging. They used the euclidean loss function in the output layer of their CNN model. They thought about the age regression job using a conventional classification method in which the ages have been divided into K groups. In addition, the network’s output layer employed the softmax function to convert the arbitrary values generated into probabilities for the projected age classes.

By specifying the classes as discrete values falling inside the range [0, 86], we try also enhance the estimation of the expected age. By increasing the number of classes to 87, considerable representation for each class from the training set is provided, which, together with balanced distribution by calculating class weights, results in proper class

training. By calculating the normalized exponential function for each class taken into account as an expected value based on probability (99), the accuracy of prediction is increased. Thus, we take age both regression and classification problem. We first estimate age linearly from face image by taking the linear output value from our hybrid model.The softmax function was then employed by the output part of the same network to convert the generated random values into probabilities for the expected age ranges of 0 to 86. Then from this probability results, age is estimated from the error between actual age and predicted age class.Finally, the best estimated age from linear regression and classification through regression will be our estimated age.

4.5 Experimental Setup and Parameter Settings:

TensorFlow, Keras was used to accomplish the suggested methodology. Google Collaboratory an Kaggle GPU with 12 GB ram was employed for network training. We utilized mean absolute error (MAE) for regression as loss function. The difference between the actual and expected value is shown by the loss function. In addition, it has previously been mentioned that human aging has some coherence and that little changes in facial structure may occur that are invisible to the human eye. For instance, the age patterns of individuals who are 31 and 33 years old may be comparable, making it difficult in some circumstances for a human to tell them apart. An RGB (Red, Green, and Blue) facial picture of 224*224 pixels serves as the network's input. The Adam optimizer is used by the network training optimizer, and its learning rate is 0.001 with weight decay is of 0.0001. Besides, we use patch size of 28*28. For vision transformer, number of head is set to 12, the dimension of the projected feature space for each patch in the transformer encoder is set to 32, the number of stacked transformer encoder layers is set to 40. For MultiHeadAttention layer, kernel initializer is set to 'glorot-uniform' with dropout of 0.1. Additionally, ViT used 'gelu' activation function while resnet50 used 'relu' as activation function. For resnet50, we freeze 25% layers from total layers, here we use the fact that the model's lowest layers pick up on fundamental visual patterns like edges, corners, and textures. The value of Epoch was set to 20 with a batch size of 32. The sample size of the available data was severely unbalanced, thus appropriate weights for each class were provided for data balancing. However, it is known that with a large number of parameters, the gradient disappears easily when training the network while the low number parameters of the network are difficult to train, thus, this work add Dropout and L2 regularizer with parameter is set to 0.03 to the network to overcome the overfitting problem. We also use ModelCheckpoint and EarlyStopping with monitor on validation loss with patience on 7 epoch.

4.6 Performance Metric:

Both positive and negative evaluations of our technique for estimating the age of masked faces were conducted. Mean absolute error (MAE), a standard metric, was employed as a measurable performance indicator. For voice,our experiment tend to use the f1 score.

4.6.1 Mean Absolute Error (MAE):

MAE computes the error between the estimated ages and the actual ages. The equation to calculate MAE is:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|A_i - P_i| \tag{4.1}$$

where N= number of test, A_i = Actual age, P_i = Estimate age.

4.6.2 Accuracy:

Accuracy metrics is used to show the performance of model. The equation to find accuracy is:

$$\frac{T_p + T_n}{T_p + F_p + F_n + T_n} \tag{4.2}$$

CHAPTER 5

Experiment

5.1 Experiment Result:

The main target of our work is age estimation using machine learning and deep learning model and giving the best performance result among them. In order to examine whether system operates more effectively and maintains greater coherence, the suggested age prediction is carried out using both classification and regression methods. The method chooses the final fully linked layer's N nodes to be added. The value of the N will depend on the number of age groups that can be categorised separately in the dataset. When using a regression method, least square regression is taken into account for this job as the final layer only comprises one unit with a linear activation function, or N = 1. On the other hand, for classification the N is 86 as we take ages of between 0 to 86 with each age depicts one class and estimate the age from the error between predicted class and actual class or age.On the UTKface benchmark datasets, the classification and regression results are examined in this section.

5.1.1 Experimental Result on HOG:

We extract Histogram of Oriented Gradients (HOG) features from photos and assess how well various machine learning models perform utilizing these features. First, the parameters used to create the HOG descriptor are determined, including the window size is set to 224*224, block size, block stride, cell size are set to 8*8, number of bins is 9, and the gradients signed are used.After that, a user defined function is developed to calculate the HOG features for each picture in the dataset. Using the previously established HOG descriptor, the function iterates through each picture in the dataset, computes its HOG features, and saves the results in a list. The list of HOG features is translated into a NumPy array and returned when all the photos have been processed. Next, mean absolute error as measure loss and a list of machine learning regression models are created to find MAE of regression result between the predicted labels and the true labels. Then for estimate age from classification output,a new set of machine learning and deep learning models, including the Random Forest Classifier, K-Nearest Neigh-

Table 5.1 Experimental result using HOG features

Regression		Classification through Regression	
Model Name	MAE	Model Name	MAE
Linear Regression	12.5767	SGD Classifier	19.4309
Ridge	12.4374	KNeighbors Classifier	21.9073
Lasso	15.3758	ExtraTrees Classifier	17.8818
		Random Forest Classifier	17.9018
Bayesian Ridge	9.4603	SVC	17.9012
		GaussianNB	22.5889
DNN	**7.0853**	MLP Classifier	14.9582
		DNN	**8.12555**

bors Classifier, Stochastic Gradient Descent Classifier, Extra Trees Classifier, Gaussian Naive Bayes Classifier, Support Vector Classifier, and Multi-Layer Perceptron Classifier, DNN is defined after all the models were evaluated using the HOG features. The following Table 5.1 shows the experimental result on HOG features.

5.1.2 Experimental Result on CNN:

A common procedure is used to train transfer learning models like ResNet50, Xception, Inception, VGG16 and others. The pre-trained version and dataset are utilized first. The model is then taken out, fine tune with required fully connected layers, and the output layer will contain the required number of neuron to estimate age. Then, Loss functions, learning rates, and optimizers are set according to the experiment. To prevent overfitting, the model is trained over a number of epochs while accuracy and loss are tracked. In addition,techniques for data augmentation can be used to small data sets. After training, fine-tuning is carried out by thawing out previously learned layers and carrying out further training at a slower learning rate and validation on a regular basis is used for wise decision-making. Fig 5.1 shows the experimental regression results on pre-trained transfer learning model:

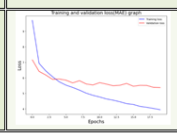
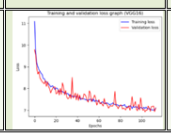
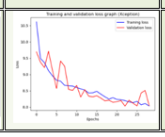
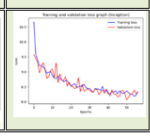
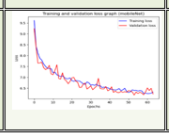
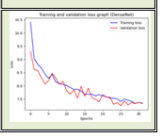
	RestNet50	VGG16	Xception	InceptionV3	MobileNet	DenseNet
MAE	5.0217	6.8215	7.7723	7.7388	6.504805	6.8760
Time(s)	2866	2,520	5,361	9,839	11,131	5,910
Loss(MAE)						

Fig. 5.1 Regression result on transfer learning model

5.1.3 Experimental Result on Features Concatenation of HOG,CNN,ViT:

We analyze the effect of explicitly feature concatenation. of HOG-CNN, HOG-CNN-ViT. To do this we use resnet50 as it gives lowest MAE comparatively lower than other CNN models. We also use the same parameter to extract hog features vector. The table 5.2,5.3 shows the experimental result of feature concatenation of HOG, CNN and ViT.

Table 5.2 Experimental result on HOG-ResNet50 Combined Features

Regression		**Classification through Regression**	
Model Name	MAE	Model Name	MAE
Linear Regression	10.94734	SGD Classifier	19.5348
Ridge	10.6915	KNeighbors Classifier	19.8690
Lasso	9.67540	ExtraTrees Classifier	17.6181
		Random Forest Classifier	17.9018
Bayesian Ridge	7.27642	SVC	19.140
		GaussianNB	22.28518
DNN	**6.0821**	MLP Classifier	19.5494
		DNN	**6.104**

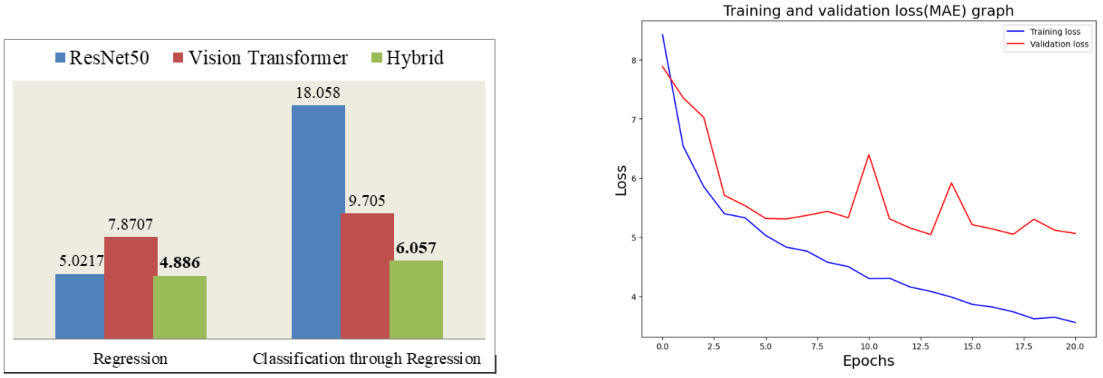
Table 5.3 Experimental result on HOG-ResNet50-ViT Combined Features

Regression		**Classification through Regression**	
Model Name	MAE	Model Name	MAE
Linear Regression	12.8466	SGD Classifier	20.1594
Ridge	12.1714	KNeighbors Classifier	19.4462
Lasso	9.5369	ExtraTrees Classifier	17.6488
		Random Forest Classifier	17.9855
Bayesian Ridge	7.1699	SVC	18.505
		GaussianNB	22.3712
DNN	**5.8121**	MLP Classifier	14.958
		DNN	**5.6152**

5.1.4 Experimental Result on Proposed Hybrid model:

Our proposed hybrid model shows lower MAE than ResNet50 and vision transformer(ViT) individually and use this feature vectors to estimate age. The following chart in Fig 5.2(a) shows the comparative MAE between Resnet50 nad ViT experimental result. The graph of training and validation of the hybrid Resnet50 and ViT is also shown in Fig 5.2(b):

Fig 5.3 reports a graphical description of the quantitative performance (MAE) over the



(a) Comparative result of ResNet50 and ViT

(b) Training and validation MAE graph of proposed hybrid model

Fig. 5.2 Experimental Result and Loss Graph of Proposed Model

real world test image using our proposed model. Finally, table 5.4 further illustrates

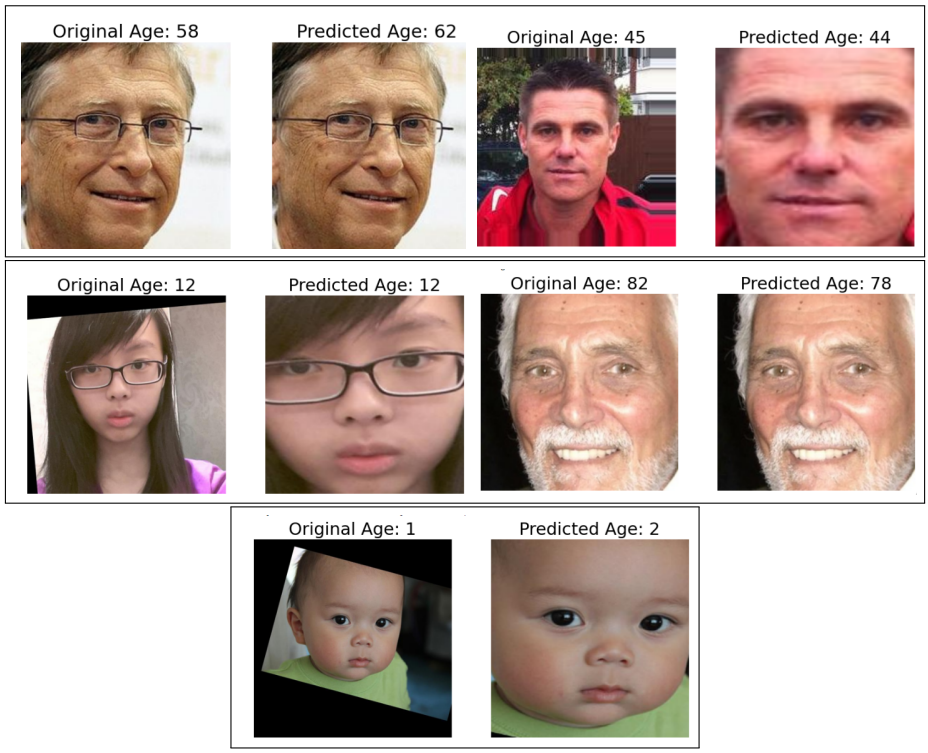


Fig. 5.3 Examples of photos with associated MAE assessed using the best model

the proposed method's superiority over the most state of arts efforts on UTKFace dataset. According to this table, our developed ResNet-50-ViT network performs noticeably better during testing than alternative network architectures.The models are trained for 20 epochs and use Mean Absolute Error(MAE) for the 'loss' parameter while training.

Table 5.4 Comparison with others' work

Method	**Age Estimation(MAE)**
Linear Regression (81)	11.73
ResNet50 (100)	9.66
ResNeXt-50 (32x4d) (81)	7.21
Inceptionv3 (100)	9.50
DenseNet (100)	9.19
Custom CNN (101)	5.67
AGES(47)	6.77
SVM (102)	7.25
CORAL(68)	5.39
EfficientNet-B3 (103)	6.71
Our proposed hybrid model	**4.88**

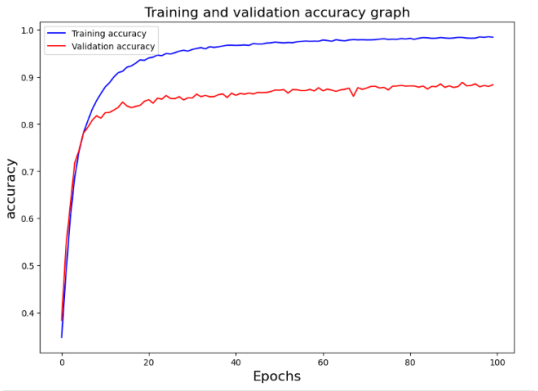
5.1.5 Experimental result on voice data:

In addition, to experiment the age estimation from voice,128 MFCC features are extracted from voice dataset using librosa library. Then, after scaling the features the above classification algorithm are used to classify age group from voice. Then, CNN with 5 convolution layer, 3 maxpool layer, 1 batchnormalization layer are experimented after reshaping features vectors. Besides, pre-trained model such as ResNet50, xception are also experimented. Here we use categorical crossentropy as loss function as adam as optimzer. The following table 5.5 shows the finding: The results from Table

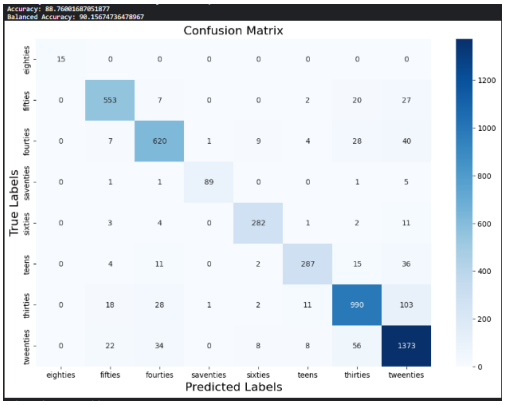
Table 5.5 Testing accuracy of age group classification from voice

Model	Accuracy
RandomForest Classifier	71.68000%
DecisionTreeClassifier	47.49000%
KNeighborsClassifier	85.51000%
AdaBoostClassifier	32.66000%
SGDClassifier	25.67000%
ExtraTreesClassifier	75.09000%
GaussianNB	32.65000%
SVC	71.03000%
CNN	**88.76001%**
Xception	84.905%
Restnet50	83.8490%

5.4 shows that training from scratch of CNN give better performance than pre-trained CNN modeland gives accuracy of 88.76%. The graph of training and validation accuracy of the CNN build from scratch is shown Fig 5.4(a) below: Fig 3.4(b) shows the



(a) Training and validation Accuracy graph



(b) Confusion matrix of age group classification from voice

Fig. 5.4 Experimental Result of Age Group Classification from Voice

confusion matrix of the trained model from voice features.Confusion matrix shows the synopsis of the prediction result on the classification problem. The number of right and unsuccessful prediction is totaled and split down by class using count values.

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion:

The main goal is to detect age from extracted features from human facial images. The facial image is a repository of age, gender, race, ethnicity, and other important information. This extracted age information is the basis of several governments and non-governmental for security, identity, usage, and business application. A simple change in the facial image such as removing wrinkles using photography tools, bright photo lights, and wearing glass or a beard can make the task of estimating the exact age the most difficult one. In today's world, such kind of complex task is solved by AI tools that need a vast variety of data. In our work, we take UTK-Face data-set consisting of 23,708 image samples with a diversity of age, gender, and race. In our work, we proposed an model to age detection using a hybrid model of vision transformer and resnet50 model. As a experiment, we also check the usage of different machine learning and deep learning algorithm along with explicitly extracting hog features, resnet50 features, ViT features and their combination are used to show their performance in age estimation. Here, we experiment the features to estimate linear regression age along with classification through regression age. Among all the experimental result our proposed hybrid model gives 4.88 MAE on UTKFace dataset which gives a lower MAE than others' work on this dataset. The approach has the potential to be extended for real-life age estimate by extracting face images from live recordings from webcams or security cameras because it performs well on benchmark datasets.

6.2 Future work:

Numerous system optimizations and enhancements may have been made, however they were not able to be done due to time restrictions. Here, our primary goal has been to estimate age with a tolerable degree of error between actual and predicted age (MAE). However, by taking use of the variety in data sets, this MAE might be reduced. Many various variables that may be changed, such as the number of layers, neurons, learning rate, etc., might have been tweaked to create a better model. The best model may be

selected with ease if several different ones are trained and tested. However, as we've already mentioned, training several networks takes a lot of time, and in this task, time was of the essence. Additionally, we did not use gender or ethnicity's influence on age detection. Future work on this will utilize other facial picture datasets. Besides, we intend to expand our work in the future by taking input from not only face images but also other biometric indicator such as voice, fingerprint, handwriting and so on.

REFERENCES

[1] S. Kim, H. Kim, E.-S. Lee, C. Lim, and J. Lee, "Risk score-embedded deep learning for biological age estimation: Development and validation," *Information Sciences*, vol. 586, pp. 628–643, 2022.

[2] M. M. Islam and J.-H. Baek, "A hierarchical approach toward prediction of human biological age from masked facial image leveraging deep learning techniques," *Applied Sciences*, 2022.

[3] P. K. Chandaliya and N. Nain, "Childgan: Face aging and rejuvenation to find missing children," *Pattern Recognition*, vol. 129, p. 108761, 2022.

[4] T. Grubl and H. S. Lallie, "Applying artificial intelligence for age estimation in digital forensic investigations," *arXiv preprint arXiv:2201.03045*, 2022.

[5] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2234–2240, 2007.

[6] M. R. Rajput and G. S. Sable, "Age group estimation from human iris," in *Soft Computing and Signal Processing: Proceedings of 2nd ICSCSP 2019 2*, pp. 519–529, Springer, 2020.

[7] M. Rajput and G. Sable, "Deep learning based gender and age estimation from human iris," in *Proceedings of the international conference on advances in electronics, electrical and computational intelligence (ICAEEC)*, 2019.

[8] M. Erbilek, M. Fairhurst, and M. C. D. C. Abreu, "Age prediction from iris biometrics," in *5th International Conference on Imaging for Crime Detection and Prevention (ICDP 2013)*, pp. 1–5, 2013.

[9] A. K. Saxena and V. K. Chaurasiya, "Fingerprint based human age group estimation," in *2014 Annual IEEE India Conference (INDICON)*, pp. 1–4, 2014.

[10] A. Abbes, R. Boukhris, and Y. B. Ayed, "Age estimation and gender recognition using biometric modality," in *Intelligent Systems Design and Applications* (A. Abraham, N. Gandhi, T. Hanne, T.-P. Hong, T. Nogueira Rios, and W. Ding, eds.), (Cham), pp. 1068–1077, Springer International Publishing, 2022.

[11] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait-based human age estimation using age group-dependent manifold learning and regression," *Multimedia tools and applications*, vol. 77, pp. 28333–28354, 2018.

[12] M. A. R. Ahad, T. T. Ngo, A. D. Antar, M. Ahmed, T. Hossain, D. Muramatsu,

Y. Makihara, S. Inoue, and Y. Yagi, "Wearable sensor-based gait analysis for age and gender estimation," *Sensors*, vol. 20, no. 8, p. 2424, 2020.

[13] H. Zhu, Y. Zhang, G. Li, J. Zhang, and H. Shan, "Ordinal distribution regression for gait-based age estimation," *Science China Information Sciences*, vol. 63, pp. 1–14, 2020.

[14] M. Azhar, S. Ullah, K. Ullah, H. Shah, A. Namoun, and K. U. Rahman, "A three-dimensional real-time gait-based age detection system using machine learning," *CMC-COMPUTERS MATERIALS and CONTINUA*, vol. 75, no. 1, pp. 165–182, 2023.

[15] A. Sakata, Y. Makihara, N. Takemura, D. Muramatsu, and Y. Yagi, "Gait-based age estimation using a densenet," in *Computer Vision–ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers 14*, pp. 55–63, Springer, 2019.

[16] A. Sakata, Y. Makihara, N. Takemura, D. Muramatsu, and Y. Yagi, "How confident are you in your estimate of a human age? uncertainty-aware gait-based age estimation by label distribution learning," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10, IEEE, 2020.

[17] Q. Riaz, M. Z. U. H. Hashmi, M. A. Hashmi, M. Shahzad, H. Errami, and A. Weber, "Move your body: Age estimation based on chest movement during normal walk," *IEEE Access*, vol. 7, pp. 28510–28524, 2019.

[18] C. Xu, A. Sakata, Y. Makihara, N. Takemura, D. Muramatsu, Y. Yagi, and J. Lu, "Uncertainty-aware gait-based age estimation and its applications," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 479–494, 2021.

[19] M. Ilyas, A. Othmani, and A. Naït-Ali, "Auditory perception based system for age classification and estimation using dynamic frequency sound," *Multimedia Tools and Applications*, pp. 1–24, 2020.

[20] I. Tsimperidis, S. Rostami, and V. Katos, "Age detection through keystroke dynamics from user authentication failures," *International Journal of Digital Crime and Forensics (IJDCF)*, vol. 9, no. 1, pp. 1–16, 2017.

[21] I. Tsimperidis, S. Rostami, K. Wilson, and V. Katos, "User attribution through keystroke dynamics-based author age estimation," in *Selected Papers from the 12th International Networking Conference: INC 2020 12*, pp. 47–61, Springer, 2021.

[22] A. Buriro, Z. Akhtar, B. Crispo, and F. Del Frari, "Age, gender and operating-hand estimation on smart mobile devices," in *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–5, IEEE, 2016.

[23] S. Roy, U. Roy, and D. Sinha, "Protection of kids from internet threats: a machine learning approach for classification of age-group based on typing pattern,"

in *Proceedings of the international multiconference of engineers and computer scientists*, vol. 1, 2018.

[24] I. Tsimperidis, C. Yucel, and V. Katos, "Age and gender as cyber attribution features in keystroke dynamic-based user classification processes," *Electronics*, vol. 10, no. 7, p. 835, 2021.

[25] G. Dobry, R. M. Hecht, M. Avigal, and Y. Zigel, "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1975–1985, 2011.

[26] M. H. Bahari, M. McLaren, D. A. van Leeuwen, *et al.*, "Speaker age estimation using i-vectors," *Engineering Applications of Artificial Intelligence*, vol. 34, pp. 99–108, 2014.

[27] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "Speaker age estimation on conversational telephone speech using senone posterior based i-vectors," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5040–5044, IEEE, 2016.

[28] A. A. Mallouh, Z. Qawaqneh, and B. D. Barkana, "New transformed features generated by deep bottleneck extractor and a gmm–ubm classifier for speaker age and gender classification," *Neural Computing and Applications*, vol. 30, pp. 2581–2593, 2018.

[29] Y. Pan, V. S. Nallanthighal, D. Blackburn, H. Christensen, and A. Härmä, "Multitask estimation of age and cognitive decline from speech," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7258–7262, IEEE, 2021.

[30] A. Almomani, M. Alweshah, W. Alomoush, M. Alauthman, A. Jabai, A. Abbass, G. Hamad, M. Abdalla, and B. B. Gupta, "Age and gender classification using backpropagation and bagging algorithms.," *Computers, Materials and Continua*, vol. 74, no. 2, 2023.

[31] B. Wu, H. Lu, Z. Chen, C. Zhu, and S. Xu, "Erdbf: Embedding-regularized double branches fusion for multi-modal age estimation," *IEEE Access*, 2023.

[32] A. Saraf and E. Khoury, "Distribution learning for age estimation from speech," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8552–8556, 2022.

[33] R. Chakroun and M. Frikha, "A deep learning approach for text-independent speaker recognition with short utterances," *Multimedia Tools and Applications*, pp. 1–23, 2023.

[34] R. Zazo, P. S. Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak, "Age estimation in short speech utterances based on lstm recurrent neural networks," *IEEE Access*, vol. 6, pp. 22524–22530, 2018.

[35] N. Tawara, A. Ogawa, Y. Kitagishi, H. Kamiyama, and Y. Ijima, "Robust speech-age estimation using local maximum mean discrepancy under mismatched recording conditions," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 114–121, IEEE, 2021.

[36] Z. Huang, P. Shivakumara, M. A. Kaljahi, A. Kumar, U. Pal, T. Lu, and M. Blumenstein, "Writer age estimation through handwriting," *Multimedia Tools and Applications*, vol. 82, no. 11, pp. 16033–16055, 2023.

[37] V. Basavaraja, P. Shivakumara, D. S. Guru, U. Pal, T. Lu, and M. Blumenstein, "Age estimation using disconnectedness features in handwriting," *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1131–1136, 2019.

[38] N. AL-Qawasmeh and C. Y. Suen, "Transfer learning to detect age from handwriting," *Adv. Artif. Intell. Mach. Learn.*, vol. 2, 2022.

[39] I. Rabaev, I. Alkoran, O. Wattad, and M. Litvak, "Automatic gender and age classification from offline handwriting with bilinear resnet," *Sensors (Basel, Switzerland)*, vol. 22, 2022.

[40] P. Pirozmand, M. F. Amiri, F. Kashanchi, and N. Y. Layne, "Age estimation, a gabor pca-lda approach," *The Journal of Mathematics and Computer Science*, vol. 2, no. 2, pp. 233–240, 2011.

[41] N. Mehrabi and S. P. H. Boroujeni, "Age estimation based on facial images using hybrid features and particle swarm optimization," in *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, pp. 412–418, IEEE, 2021.

[42] M. M. Sawant and K. Bhurchandi, "Hierarchical facial age estimation using gaussian process regression," *IEEE Access*, vol. 7, pp. 9142–9152, 2019.

[43] D. Lu, D. Wang, K. Zhang, and X. Zeng, "Age estimation from facial images based on gabor feature fusion and the ciaso-sa algorithm," *CAAI Transactions on Intelligence Technology*, 2022.

[44] N. Ramanathan and R. Chellappa, "Modeling age progression in young faces," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, pp. 387–394, 2006.

[45] P. Koruga, M. Bača, and J. Ševa, "Application of modified anthropometric model in facial age estimation," in *Proceedings ELMAR-2011*, pp. 17–20, 2011.

[46] S. Kumar, S. Ranjitha, and H. N. Suresh, "An active age estimation of facial image using anthropometric model and fast ica," *Journal of Engineering Science and Technology Review*, vol. 10, pp. 100–106, 2017.

[47] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[48] S. Feng, C. Lang, J. Feng, T. Wang, and J. Luo, "Human facial age estimation by cost-sensitive label ranking and trace norm regularization," *IEEE Transactions on Multimedia*, vol. 19, no. 1, pp. 136–148, 2016.

[49] K. Chen, S. Gong, T. Xiang, and C. Change Loy, "Cumulative attribute space for age and crowd density estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2467–2474, 2013.

[50] P. Yang, L. Zhong, and D. N. Metaxas, "Ranking model for facial age estimation," *2010 20th International Conference on Pattern Recognition*, pp. 3404–3407, 2010.

[51] M. Hajizadeh and H. Ebrahimnezhad, "Classification of age groups from facial image using histograms of oriented gradients," *2011 7th Iranian Conference on Machine Vision and Image Processing*, pp. 1–5, 2011.

[52] H. Han and A. K. Jain, "Age , gender and race estimation from unconstrained face images," 2014.

[53] O. F. W. Onifade and D. J. Akinyemi, "Gwageer - a groupwise age ranking framework for human age estimation," *International Journal of Image, Graphics and Signal Processing*, vol. 7, pp. 1–12, 2015.

[54] I. H. Casado, C. Fernández, C. Segura, J. Hernando, and A. Prati, "A deep analysis on age estimation," *Pattern Recognit. Lett.*, vol. 68, pp. 239–249, 2015.

[55] K. Liu, S. Yan, and C.-C. J. Kuo, "Age estimation via grouping and decision fusion," *IEEE Transactions on Information Forensics and Security*, vol. 10, pp. 2408–2423, 2015.

[56] Y. Dong, Y. Liu, and S. Lian, "Automatic age estimation based on deep learning algorithm," *Neurocomputing*, vol. 187, pp. 4–10, 2016.

[57] H. Liu, J. Lu, J. Feng, and J. Zhou, "Ordinal deep learning for facial age estimation," *IEEE transactions on circuits and systems for video technology*, vol. 29, no. 2, pp. 486–501, 2017.

[58] F. S. Abousaleh, T. Lim, W.-H. Cheng, N.-H. Yu, M. A. Hossain, and M. F. Alhamid, "A novel comparative deep learning framework for facial age estimation," *EURASIP Journal on Image and Video Processing*, vol. 2016, pp. 1–13, 2016.

[59] S. Zaghbani, N. Boujneh, and M. S. Bouhlel, "Age estimation using deep learning," *Computers and Electrical Engineering*, vol. 68, pp. 337–347, 2018.

[60] O. Sendik and Y. Keller, "Deepage: deep learning of face-based age estimation," *Signal Processing: Image Communication*, vol. 78, pp. 368–375, 2019.

[61] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling, "Diagnosing deep learning models for high accuracy age estimation from a single image," *Pattern Recognition*, vol. 66, pp. 106–116, 2017.

[62] Z. Hu, Y. Wen, J. Wang, M. Wang, R. Hong, and S. Yan, "Facial age estimation with age difference," *IEEE Transactions on Image Processing*, vol. 26, no. 7,

pp. 3087–3097, 2016.

[63] H. Liu, J. Lu, J. Feng, and J. Zhou, "Group-aware deep feature learning for facial age estimation," *Pattern Recognition*, vol. 66, pp. 82–94, 2017.

[64] Z. Tan, J. Wan, Z. Lei, R. Zhi, G. Guo, and S. Z. Li, "Efficient group-n encoding and decoding for facial age estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2610–2623, 2017.

[65] S. A. Rizwan, Y. Y. Ghadi, A. Jalal, and K. Kim, "Automated facial expression recognition and age estimation using deep learning.," *Computers, Materials and Continua*, vol. 71, no. 3, 2022.

[66] I. Aruleba and S. Viriri, "Deep learning for age estimation using efficientnet," in *Advances in Computational Intelligence: 16th International Work-Conference on Artificial Neural Networks, IWANN 2021, Virtual Event, June 16–18, 2021, Proceedings, Part I 16*, pp. 407–419, Springer, 2021.

[67] B. Yoo, Y. Kwak, Y. Kim, C. Choi, and J. Kim, "Deep facial age estimation using conditional multitask learning with weak label expansion," *IEEE Signal Processing Letters*, vol. 25, no. 6, pp. 808–812, 2018.

[68] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognition Letters*, vol. 140, pp. 325–331, 2020.

[69] Y. Shou, X. Cao, and D. Meng, "Masked contrastive graph representation learning for age estimation," *arXiv preprint arXiv:2306.17798*, 2023.

[70] G. Chen, J. Peng, L. Wang, H. Yuan, and Y. Huang, "Feature constraint reinforcement based age estimation," *Multimedia Tools and Applications*, vol. 82, no. 11, pp. 17033–17054, 2023.

[71] Z. Huang, J. Zhang, and H. Shan, "When age-invariant face recognition meets face age synthesis: A multi-task learning framework," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7282–7291, 2021.

[72] V. Raman, K. ELKarazle, and P. Then, "Gender-specific facial age group classification using deep learning," *Intell. Autom. Soft Comput*, vol. 34, pp. 105–118, 2022.

[73] S. K. Gupta and N. Nain, "Single attribute and multi attribute facial gender and age estimation," *Multimedia Tools and Applications*, vol. 82, no. 1, pp. 1289–1311, 2023.

[74] S. H. Nam, Y. H. Kim, N. Q. Truong, J. Choi, and K. R. Park, "Age estimation by super-resolution reconstruction based on adversarial networks," *IEEE Access*, vol. 8, pp. 17103–17120, 2020.

[75] A. Greco, A. Saggese, M. Vento, and V. Vigilante, "Effective training of convolutional neural networks for age estimation based on knowledge distillation,"

Neural Computing and Applications, pp. 1–16, 2021.

[76] S. Taheri and Ö. Toygar, "On the use of dag-cnn architecture for age estimation with multi-stage features fusion," *Neurocomputing*, vol. 329, pp. 300–310, 2019.

[77] Z. Jiang and C. Zhou, "Age detection by optimizing the structure of layers and neurons in the neural network," *Journal of Optics*, pp. 1–17, 2023.

[78] C. Miron, V. Manta, R. Timofte, A. Pasarica, and R.-I. Ciucu, "Efficient convolutional neural network for apparent age prediction," in *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 259–262, 2019.

[79] S. E. Bekhouche, F. Dornaika, A. Ouafi, and A. Taleb-Ahmed, "Human facial age estimation: Handcrafted features versus deep features," *Emerging Technologies in Biomedical Engineering and Sustainable TeleMedicine*, pp. 31–37, 2021.

[80] I. Dagher and D. Barbara, "Facial age estimation using pre-trained cnn and transfer learning," *Multimedia Tools and Applications*, vol. 80, pp. 20369–20380, 2021.

[81] A. Fariza, A. Z. Arifin, *et al.*, "Age estimation system using deep residual network classification method," in *2019 International Electronics Symposium (IES)*, pp. 607–611, IEEE, 2019.

[82] S. V. L. C. Wang, H., "Improving face-based age estimation with attention-based dynamic patch fusion," *IEEE Trans. on Image Process.*, vol. 31, pp. 1084–1096, 2022.

[83] X. Liu, Y. Zou, H. Kuang, and X. Ma, "Face image age estimation based on data augmentation and lightweight convolutional neural network," *Symmetry*, vol. 12, no. 1, p. 146, 2020.

[84] S. Hiba and Y. Keller, "Hierarchical attention-based age estimation and bias estimation," *arXiv preprint arXiv:2103.09882*, 2021.

[85] K. Zhang, N. Liu, X. Yuan, X. Guo, C. Gao, Z. Zhao, and Z. Ma, "Fine-grained age estimation in the wild with attention lstm networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 3140–3152, 2019.

[86] H. O. Ikromovich and B. B. Mamatkulovich, "Facial recognition using transfer learning in the deep cnn," *Open Access Repository*, vol. 4, no. 3, pp. 502–507, 2023.

[87] J. M. Sahan, E. I. Abbas, and Z. M. Abood, "A facial recognition using a combination of a novel one dimension deep cnn and lda," *Materials Today: Proceedings*, vol. 80, pp. 3594–3599, 2023.

[88] H. N. Vu, M. H. Nguyen, and C. Pham, "Masked face recognition with convolutional neural networks and local binary patterns," *Applied Intelligence*, vol. 52, no. 5, pp. 5497–5512, 2022.

[89] R. Hammouche, A. Attia, S. Akhrouf, and Z. Akhtar, "Gabor filter bank with

deep autoencoder based face recognition system," *Expert Systems with Applications*, vol. 197, p. 116743, 2022.

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

[91] S. E. X. B. I. G. R. R. E Agustsson, R Timofte, "Apparent and real age estimation in still images with deep residual regressors on appa-real database.," in *12th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2017*, IEEE, 2017.

[92] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[93] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 896–903, 2013.

[94] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[95] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, p. 137–154, may 2004.

[96] B. K. Savaş, S. İlkin, and Y. Becerikli, "The realization of face detection and fullness detection in medium by using haar cascade classifiers," in *2016 24th Signal Processing and Communication Application Conference (SIU)*, pp. 2217–2220, 2016.

[97] B. Bas Pujols and F. Riera Pol, "Facial image-based gender and age estimation," B.S. thesis, Universitat Politècnica de Catalunya, 2013.

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

[99] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 144–157, 2018.

[100] M. Akhand, M. Ijaj Sayim, S. Roy, and N. Siddique, "Human age prediction from facial image using transfer learning in deep convolutional neural networks," in *Proceedings of International Joint Conference on Computational Intelligence: IJCCI 2019*, pp. 217–229, Springer, 2020.

[101] V. Sheoran, S. Joshi, and T. R. Bhayani, "Age and gender prediction using deep cnns and transfer learning," in *Computer Vision and Image Processing: 5th In-*

ternational Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5, pp. 293–304, Springer, 2021.

[102] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Transactions on pattern Analysis and machine Intelligence*, vol. 24, no. 4, pp. 442–455, 2002.

[103] E. Barcic, P. Grd, I. Tomicic, and B. Okresa Duric, "Age estimation of occluded faces using efficientnet-b3 cnn model," in *Proceedings of the 2022 European Interdisciplinary Cybersecurity Conference*, pp. 95–96, 2022.

Appendix A

Code (Part 1):

The following python code generate our image dataset of train,test,validation image array from original UTKFace dataset which we used in our all experiment:

```
images = []
age = []
BASE_DIR =
    'path_to_utkface_cropped_aligned_image_dataset'
for img in os.listdir(BASE_DIR):
    label = img.split("_")[0]
    img = cv2.imread(str(BASE_DIR)+"/"+str(img))
    img = cv2.resize(img, (224, 224))
    img = cv2.cvtColor(img,cv2.COLOR_BGR2RGB)
    images.append(np.array(img))
    age.append(np.array(label))
age = np.array(age,dtype=np.int64)
images = np.array(images)
x_train_image, x_test_image, y_train_age, y_test_age =
    train_test_split(images, age,test_size=0.2,
    random_state=42)
x_train_image, x_valid_image, y_train_age, y_valid_age =
    train_test_split(x_train_image,
    y_train_age,test_size=0.2, random_state=42)
np.savez('UTKTrain.npz', age=y_train_age,
    images=x_train_image)
np.savez('UTKValidation.npz', age=y_valid_age,
    images=x_valid_image)
np.savez('UTKTtest.npz', age=y_test_age,
    images=x_test_image)
```

Appendix B

Code (Part2):

The following box shows the hyper parameter configuration of our proposed hybrid model:

```
learning_rate = 0.001
weight_decay = 0.0001
batch_size = 32
num_epochs = 10
image_size = 224
patch_size = 28
num_patches = (image_size // patch_size) ** 2
projection_dim = 32   #the dimension of the projected
   feature space for each patch in the transformer
   encoder.
num_heads = 12 #the number of attention heads in the
   multi-head attention mechanism in the transformer
   encoder.
transformer_units = [
    projection_dim * 2,
    projection_dim]
transformer_layers = 40
mlp_head_units = [2048, 1024]
```

The following code shows the resnet50's configuration of our proposed hybrid model:

```
resNet50 = ResNet50(weights='imagenet',
   include_top=False, input_tensor=augmented_image)
freeze_layers = math.ceil(len(resNet50.layers) * 0.25)
for layer in resNet50.layers[:freeze_layers]:
    layer.trainable = False
resNet50_features =
   layers.GlobalAveragePooling2D()(resNet50.output)
```

The following code shows the ViT's configuration and combine with resnet50 features of our proposed hybrid model:

```
inputs = layers.Input(shape=input_shape)
augmented_image = data_augmentation(inputs)
patches = Patches(patch_size)(augmented_image)

# Encode patches.
encoded_patches = PatchEncoder(num_patches,
   projection_dim)(patches)
# Create multiple layers of the Transformer block.
for _ in range(transformer_layers):
    # Layer normalization 1.
    x1 = LayerNormalization(epsilon=1e-6)
       (encoded_patches)
    # Create a multi-head attention layer.
    attention_output = MultiHeadAttention
       (num_heads=num_heads,
       kernel_initializer='glorot_uniform',
    key_dim=projection_dim, dropout=0.1)(x1, x1)
    # Skip connection 1.
    x2 = Add()([attention_output, encoded_patches])
    # Layer normalization 2.
    x3 = LayerNormalization(epsilon=1e-6)(x2)
    # MLP.
    x4 = mlp(x3, hidden_units=transformer_units,
       dropout_rate=0.1)
    # Skip connection 2.
    encoded_patches = layers.Add()([x4, x2])

representation =
   LayerNormalization(epsilon=1e-6)(encoded_patches)
vit_features = Flatten()(representation)

#Concatenation of ViT and Resnet50 features
concatenated_features =
   layers.Concatenate()([vit_features,resNet50_features])
```