

ENHANCING AGE ESTIMATION USING LIGHTWEIGHT MACHINE LEARNING MODELS FOR EDGE
DEVICES: CLOSING THE PERFORMANCE GAP

RAOOF RAZA KHAN
LJMU MASTERS IN DS, OCTOBER 2022

Final Thesis Report

OCTOBER 2023

ABSTRACT

The rapid proliferation of edge devices and the burgeoning need for real-time data processing accentuates the importance of lightweight machine learning models. This research voyages into the intricate domain of age estimation using such models, particularly focusing on their deployment feasibility on resource-constrained edge devices. Aiming to bridge the gap between model efficacy and computational efficiency, the study harnesses the potential of pre-trained architectures, exemplified by the MobileNet framework.

An exhaustive exploration and analysis of the Adience dataset, comprising 19,370 images across diverse age brackets, forms the backbone of this research. Through meticulous exploratory data analysis, the dataset's intricacies, including age distribution, gender variance, and potential labeling anomalies, were brought to the forefront. The study underscored the importance of robust preprocessing techniques and advanced feature engineering to optimize model inputs and enhance predictive capabilities.

In its pursuit of model refinement, the research embarked on an iterative journey, employing a range of techniques from augmenting data with the Augmentor to the inclusion of Stratified K-fold for better data representation during training. The ever-present challenge of overfitting was tackled through strategic interventions, such as the application of L2 regularization. This iterative approach bore fruit, with the research achieving an impressive binary classification accuracy of 88.88% in discerning 'Minors' from 'Adults', and an impressive 74.32% accuracy in classifying ages into six distinct groups.

Furthermore, the research, through its comprehensive evaluation metrics, provides a holistic view of model performance, emphasizing not just accuracy but also metrics like the ROC curve, underscoring the importance of a rounded evaluation paradigm. Additionally, it casts a critical eye on the Adience dataset, offering insights into its attributes and potential limitations, making it a beacon for future age estimation endeavors using this dataset.

In essence, this research stands as a testament to the potential of lightweight machine learning models in age estimation tasks, providing both a benchmark and a blueprint for future academic and practical pursuits. By intertwining advanced methodologies, dataset insights, and rigorous evaluations, it offers a panoramic view of age estimation on edge devices, marking a significant stride in this domain.

TABLE OF CONTENTS

TABLE OF CONTENTS.....	3
LIST OF TABLES.....	6
LIST OF ABBREVIATIONS.....	9
Chapter 1.....	10
INTRODUCTION	10
1.1 Background	10
1.1.1 The Multidisciplinary Nature of Age Estimation	10
1.1.2 Traditional Paradigms in Age Estimation.....	11
1.1.3 Lightweight Models	11
1.1.4 Real-World Impact.....	12
1.1.5 Insights from the Literature	13
1.1.6 The Imperative for Harmonization	13
1.1.7 Summary	13
1.2 Problem Statement	14
1.3 Aims and Objectives	15
1.4 Significance of the Study	15
1.5 Scope of Study	16
1.5.1 In Scope:	16
1.5.2 Out Scope:.....	17
1.6 Required Resources.....	17
1.6.1 Data Resource:	17
1.6.2 Computational Resources:	18
1.7 Structure of the Study.....	18
Chapter 2.....	19
LITERATURE REVIEW	19
2.1 Overview of the Research Landscape	19
2.1.1 Analytical Representation of Keywords and Abstracts	19
2.1.2 Typology of Architectures Employed.....	22
2.1.3 Temporal Distribution of Selected Research Papers	24
2.1.4 Distribution Plot of Selected Research Papers.....	25
2.2 Table of Comparison.....	26
2.2.1 Comparative Analysis.....	26
2.3 Descriptive Content for Selected Works.....	35
2.4 Summary	37

Chapter 3.....	39
RESEARCH METHODOLOGY.....	39
3.1 End-to-End Framework.....	39
3.1.1 Data Acquisition and Preprocessing.....	39
3.1.2 Model Architecture and Training.....	39
3.1.3 Hyperparameter Tuning.....	40
3.1.4 Model Evaluation.....	40
3.2 Dataset Details.....	40
3.3 Data Processing.....	46
3.3.1 Normalization and Resizing.....	47
3.3.2 Data Augmentation.....	48
3.3.3 Label Encoding and Categorization.....	52
3.3.4 Data Generators.....	53
3.4 Model Architecture.....	55
3.4.1 Pre-Trained Architecture Selection.....	55
3.4.2 Proposed CNN Architecture with Pretrained MobileNetV2.....	62
3.5 Validation Strategy.....	65
3.6 Model Evaluation.....	66
3.7 Summary.....	67
Chapter 4.....	69
ANALYSIS.....	69
4.1 Introduction.....	69
4.2 General Data Characteristics.....	70
4.2.1 Number of RGB Images.....	70
4.2.2 Distribution of Images Across Folds.....	71
4.2.3 Missing Values Analysis.....	72
4.2.4 Outlier Analysis.....	74
4.2.5 Descriptive Statistics Analysis.....	76
4.3 Distribution Analysis.....	78
4.3.1 Distribution of Age Groups.....	78
4.3.2 Distribution of Gender.....	80
4.4 Facial Feature Analysis.....	82
4.4.1 Distribution of Face Sizes.....	82
4.4.2 Distribution of Tilt and Yaw Angles.....	83
4.5 Orientation and Angles.....	87

4.5.1	Correlation between Angles.....	87
4.5.2	Orientation	89
4.6	Summary	91
Chapter 5.....		93
EXPERIMENT AND RESULTS		93
5.1	Deep Model.....	94
5.2	Wide Model.....	101
5.3	Deep Model with Fine-Tuned Architecture	109
5.4	Wide Model with Fine-Tuned Architecture	114
5.5	Wide Model with Fine-Tuned Architecture and Augmentor Library	119
5.6	Wide Model with Fine-Tuned Architecture and Stratified Cross Validation	126
5.7	Wide Model with Stratified Cross Validation and L2 Regularization	131
5.8	Binary Classification Adaptation to Distinguishing Minors from Adults.....	138
5.9	Summary	143
Chapter 6.....		146
FUTURE SCOPE.....		146
6.1	Conclusion.....	146
6.2	Solution to Objective 1.....	147
6.3	Solution to Objective 2.....	148
6.4	Contribution to Knowledge.....	150
6.5	Future Work	152
REFERENCES.....		154
APPENDIX A.....		158
GANTT CHART		158
APPENDIX B		159
Research Proposal.....		159

LIST OF TABLES

Table 1 Comparative Overview of Selected Works in Age Estimation Research.....	26
Table 2 Distribution of images in each age group across the five folds.	42
Table 3 Fold-wise distribution of age categories after grouping.	53
Table 4 Summarizing All Models.....	144
Table 5 Gantt Chart of Research Plan.....	158

LIST OF FIGURES

Figure 1 Word cloud visualization of predominant keywords.....	19
Figure 2 Word cloud visualization of Principal themes within Abstracts	21
Figure 3 Pie chart visualization of the distribution of architectures	22
Figure 4 Pie chart of architecture distribution, grouped by CNN-based types and others	23
Figure 5 Line graph of annual publication count of selected papers	24
Figure 6 Distribution plot with KDE of selected papers by year.....	25
Figure 7 Sample Images taken from Adience Dataset.....	44
Figure 8 VGG16 Architecture	56
Figure 9 InceptionV3 Architecture	58
Figure 10 MobileNetV2 Architecture.....	59
Figure 11 EffecientNet-B0 Architecture.....	60
Figure 12 AlexNet Architecture.....	60
Figure 13 Descriptive Summary	77
Figure 14 Combined Distribution of Age Groups Across All Folds	79
Figure 15 Combined Dtribution of Gender Across All Folds.....	81
Figure 16 Combined Dtribution of Face Area Across All Folds	83
Figure 17 Combined Dtribution of Tilt and Yaw Angles Across All Folds	84
Figure 18 Dtribution of Tilt Angle	85
Figure 19 Dtribution of Yaw Angle	86
Figure 20 Correlation between Angles	88
Figure 21 Orientation.....	90
Figure 22 Confusion Matrix Model 1	96
Figure 23 Loss and Accuracy vs Epochs Model 1.....	98
Figure 24 ROC curve Model 1.....	100
Figure 25 Confusion Matrix Model 2	104
Figure 26 Loss and Accuracy vs Epochs Model 2.....	105
Figure 27 ROC curve Model 2.....	107
Figure 28 Confusion Matrix Model 3.....	110
Figure 29 Loss and Accuracy vs Epochs Model 3.....	111
Figure 30 ROC Curve Model 3.....	113
Figure 31 Confusion Matrix Model 4.....	116
Figure 32 Loss and Accuracy vs Epochs Model 4.....	116
Figure 33 ROC Curve Model 4.....	118
Figure 34 Confusion Matrix Model 5	121
Figure 35 Loss and Accuracy vs Epochs Model 5.....	122
Figure 36 ROC Curve Model 5.....	124
Figure 37 Confusion Matrix Model 6.....	127
Figure 38 Loss and Accuracy vs Epochs Model 6.....	128
Figure 39 ROC Curve Model 6.....	130
Figure 40 Confusion Matrix Model 7	134
Figure 41 Loss and Accuracy vs Epochs Model 7.....	135
Figure 42 ROC Curve Model 7.....	138
Figure 43 Confusion Matrix Model 8.....	140
Figure 44 Loss and Accuracy vs Epochs Model 8.....	141

Figure 45 ROC Curve Model 8.....	142
----------------------------------	-----

LIST OF ABBREVIATIONS

AUC.....	Area Under the Curve
CNN.....	Convolutional Neural Network
CSV.....	Comma Separated Values
DNN.....	Deep Neural Network
EDA.....	Exploratory Data Analysis
FPR.....	False Positive Rate
GAN.....	Generative Adversarial Network
IQR.....	Inter Quantile Range
KDE.....	kernel density estimate
LSTM.....	Long Short-Term Memory
MAE.....	Mean Absolute Error
OOD.....	Out-of-Distribution
RGB.....	Red Green Blue
ROC.....	Receiver Operating Characteristic
SVM.....	Support Vector Machine
TPR.....	True Positive Rate
UUID.....	Universally Unique Identifier

CHAPTER 1

INTRODUCTION

As technological advancements continue to permeate various aspects of modern life, there is a growing imperative for intelligent systems that can effectively interface with human attributes. Age estimation, a seemingly simple yet profoundly complex task, sits at the intersection of multiple disciplines and holds significant societal implications, especially when deployed on resource-constrained edge devices.

1.1 Background

The background section endeavours to provide a comprehensive understanding of the various dimensions influencing age estimation, especially its deployment on edge devices. The subsequent subsections will delve into the interdisciplinary nature of the field, the challenges and limitations of traditional paradigms, and the emergence and significance of lightweight models. The intent is to establish the context within which this study operates, highlighting both the technological and societal implications

1.1.1 The Multidisciplinary Nature of Age Estimation

Age estimation serves as a striking illustration of interdisciplinary collaboration, seamlessly integrating the computational capabilities inherent in computer vision, the meticulous accuracy emphasized in biometrics, and the nuanced intricacies present in behavioural science. This amalgamation of varied disciplines gives rise to a unique, yet complex, set of challenges and opportunities. Each facet necessitates specialized solutions, particularly in striking an optimal balance between predictive accuracy and computational efficiency. This equilibrium gains heightened significance when considering edge devices as the intended deployment platforms. Such devices, which span a range from ubiquitous smartphones and wearables to more specialized Internet of Things (IoT) devices, inherently operate under a set of resource limitations. These constraints, which include but are not limited to, limited processing power, constrained memory capacity, and stringent energy efficiency requirements, mandate a reevaluation and potential overhaul of conventional modelling paradigms. Consequently, devising age estimation solutions suitable for edge devices necessitates a judicious blend of methodologies derived from each of its foundational disciplines. The overarching objective

remains the development of a model that not only excels in predictive accuracy but also adheres to the computational efficiency prerequisites of edge-based deployment scenarios.

1.1.2 Traditional Paradigms in Age Estimation

Traditional approaches to age estimation have predominantly relied on the computational strengths of CNNs along with other intricate machine learning architectures. These architectures, celebrated for their superior performance in image recognition and feature extraction, often demand extensive computational resources, not only during the training phase but also for real-time inference. This high computational requirement often renders them impractical for deployment on edge devices, which are constrained by limited processing capabilities, finite storage options, and strict energy consumption guidelines. Consequently, the extensive array of conventional methodologies in age estimation often finds itself misaligned with the practical necessities of real-world applications on edge devices. This incongruence poses a significant challenge and accentuates the imperative for innovative research. The burgeoning demand for real-time age estimation in various contexts, from content filtering to emergency response, amplifies the need to adapt these computationally intensive models to be more resource-efficient. In doing so, the aim is to reconcile their inherent complexities with the resource constraints of edge devices, thereby extending their utility in diverse, real-world scenarios.

1.1.3 Lightweight Models

In an attempt to circumvent the computational bottlenecks associated with traditional models, the research community has witnessed a surge in the development of lightweight models. These models are engineered with a focus on reducing the computational complexity, often by employing fewer parameters, simplified architectures, and optimized algorithms. However, the quest for computational frugality has frequently led to a compromise on predictive accuracy, resulting in a notable performance gap when compared to state-of-the-art models. This dichotomy between computational efficiency and predictive precision stands as a formidable challenge, impeding the broader applicability of age estimation technologies in everyday scenarios, particularly those involving edge devices.

1.1.4 Real-World Impact

The significance of efficient and accurate age estimation transcends academic inquiry, exerting a profound influence on various societal sectors and daily life activities. In the digital sphere—where user-generated content is omnipresent—the need for robust, real-time age verification mechanisms is heightened, fulfilling both legal obligations and ethical standards. Public venues with high footfall, such as shopping centres, transport stations, and entertainment facilities, could deploy automated systems that utilize lightweight age estimation models to prioritize assistance to vulnerable demographics, thereby potentially averting crises and enhancing user experience.

Targeted advertising stands as another realm where accurate age estimation holds pivotal importance. It allows marketers to formulate campaigns that are not only more tailored but also ethically compliant, negating the need for intrusive data gathering methods. The healthcare sector could also benefit from lightweight age estimation models, with automated pharmacy kiosks using such technology to recommend age-appropriate medications or treatments.

In educational ecosystems, these models facilitate the delivery of age-suited educational content, ensuring that learning materials are cognitively appropriate. Furthermore, age-restricted venues like casinos and clubs could employ these models for streamlined, non-invasive age verification. During emergency evacuations, lightweight age estimation models could expedite the identification and prioritization of individuals needing special assistance, such as the elderly and children.

One of the most critical and ethically compelling applications lies in the domain of digital forensics and child safety. Lightweight age estimation models could serve as an initial safeguard against the dissemination of illegal content featuring minors, by identifying underage individuals in digital images and thereby interrupting the distribution at its source.

In terms of privacy and cost-efficiency, cloud-based solutions often pose risks and are not always viable for real-time, remote applications. Lightweight models, in contrast, offer the advantage of on-the-fly processing directly on edge devices. This not only preserves user privacy by reducing data transmission to centralized servers but also makes the technology more accessible and cost-effective, particularly in remote locations with limited connectivity.

By leveraging the computational efficiency and reduced hardware requirements of lightweight models, these wide-ranging applications can be realized directly on edge devices, providing real-time, privacy-preserving solutions that are both ethically and economically advantageous.

1.1.5 Insights from the Literature

The arena of age estimation is enriched by a plethora of methodological approaches, each contributing a unique perspective to the overarching narrative. Initial forays into this domain were often restricted to addressing computational inefficiencies. However, as the field matured, researchers began to grapple with more nuanced challenges, such as data imbalances, Out-of-Distribution (OOD) generalization, and even the incorporation of unconventional data types like handwritten signatures. These progressive shifts not only underscore the continual evolution of age estimation methodologies but also highlight the ever-expanding thematic scope of this research area. While these contributions have individually advanced the state of the art in either predictive accuracy or computational efficiency, a harmonized solution that effectively balances these two key aspects remains elusive.

1.1.6 The Imperative for Harmonization

The current landscape of age estimation research is marked by a pressing need for models that excel not just in isolated dimensions of predictive accuracy or computational efficiency, but in a harmonized blend of both. The diverging paths of computational and predictive optimization must be reconciled to pave the way for a new generation of age estimation models. These models should not only be adept at handling the resource constraints of edge devices but also be capable of delivering reliable and accurate age estimates, thereby fulfilling a spectrum of real-world applications ranging from content filtering and emergency response to targeted advertising and user experience personalization.

1.1.7 Summary

This study contributes to the existing body of work in the field of age estimation, specifically targeting the optimization of lightweight models suitable for edge device deployment. Through rigorous empirical evaluations and comparative analyses, this research has sought to narrow, the prevailing performance gap between state-of-the-art models and lightweight alternatives. The ultimate objective of this research has been to broaden the real-world applicability of age estimation technologies, rendering them both computationally efficient and predictively reliable across a diverse array of edge-based scenarios.

1.2 Problem Statement

The landscape of age estimation has seen substantial advancements over the years, propelled by a combination of sophisticated machine learning algorithms, rich datasets, and increasingly intricate architectural designs. These developments have largely focused on optimizing predictive accuracy, often culminating in models that are computationally intense and resource-demanding. While these state-of-the-art methods offer impressive accuracy rates, they typically require robust computing environments to function efficiently. This computational burden limits their applicability in resource-constrained settings, notably edge devices, where limitations in processing power, memory capacity, and energy efficiency are significant concerns.

Current lightweight models, designed with a focus on computational efficiency, often make trade-offs that adversely impact their predictive performance. This results in a performance gap when compared to their state-of-the-art counterparts, which, although accurate, are unsuitable for real-time, edge-based applications. This dichotomy between computational efficiency and predictive accuracy forms the crux of the problem this study aims to address.

The core issue framing this research is multifaceted: Firstly, how can one construct lightweight models that are in alignment with the computational constraint's endemic to edge devices? Secondly, how can these computationally frugal models maintain a predictive accuracy that is on par with their more resource-intensive counterparts? This dual inquiry sits at the intersection of computational efficiency and predictive performance, representing a significant lacuna in the existing scholarly landscape. It raises imperative questions about the adaptability of existing machine learning paradigms to resource-constrained environments and the resultant trade-offs that must be navigated. This research aims to address these pressing questions, thereby contributing to the ongoing discourse in age estimation methodologies suitable for real-world, edge-based applications.

In summary, the problem statement of this study centres on the development of optimized lightweight models capable of high-performance age estimation on edge devices. It addresses the need for a nuanced approach that reconciles the often-conflicting objectives of computational efficiency and predictive accuracy, thereby enabling the broader applicability of age estimation technologies in various real-world, edge-based scenarios.

1.3 Aims and Objectives

The primary aim of this research is to utilize pre-trained architectures for the development of lightweight machine learning models. These models, designed to be effective in age estimation tasks, are intended to be suitable for deployment on resource-constrained edge devices. Rigorous evaluation is to be conducted using the Adience dataset to ascertain the models' applicability in real-world scenarios.

Arising from the principal aim, the specific objectives of this research are delineated as follows:

- To formulate and implement lightweight CNN models, built upon pre-trained architectures, that are optimized for age estimation tasks.
- To devise and implement effective data preprocessing techniques aimed at enhancing the input quality from the Adience dataset, thereby contributing to improved model performance.
- To explore and integrate feature engineering techniques to augment predictive capabilities.
- To iteratively adapt the model architectures by adding specific layers, employing regularization methods, or making other adjustments with the objective of optimizing predictive performance without compromising computational efficiency.
- To conduct a robust evaluation of the developed lightweight models using a range of performance metrics such as accuracy, ROC curve, and confusion matrix.
- To assess the attributes and limitations of the Adience dataset in the context of training lightweight models and evaluating their performance.

1.4 Significance of the Study

The significance of this research is manifold and extends across both academic and practical realms. At its core, this study aims to mitigate the performance gap between lightweight models and their state-of-the-art counterparts in the domain of age estimation. By focusing on the deployment of such models on resource-constrained edge devices, this research elevates the discourse surrounding facial attribute identification to a more practically relevant context.

Methodologically, this study brings innovation by leveraging pre-trained architectures as foundational blocks and further refining them through various fine-tuned architectures, data augmentations, and regularization techniques. This approach, which can be termed as a

"coarse-to-fine" strategy, ensures that the models capture general facial patterns before zeroing in on age-specific features. Rigorous evaluations are conducted using the Adience dataset, serving as an empirical basis for assessing the models' predictive capabilities and computational efficiency.

From a societal perspective, the enhanced lightweight models developed in this study hold the potential to be more widely and easily deployed in real-world, edge-based applications. This includes age-based content filtering, emergency response prioritization, and targeted advertising, among others. Consequently, these models could lead to improved user experiences and more effective public service provisions, all while maintaining privacy and reducing computational costs.

The study also addresses the critical challenge of balancing predictive accuracy with computational frugality, a topic of increasing relevance given the proliferation of edge devices in daily life. The research outcomes are expected to serve as an asset for future work aiming to reconcile the often-conflicting objectives of computational efficiency and predictive accuracy.

In summary, this research not only contributes to the existing body of knowledge in age estimation but also paves the way for more efficient, real-world deployment of facial attribute recognition technologies on edge devices.

1.5 Scope of Study

The scope of this research is carefully defined to focus on key areas that contribute to the study's primary aim and objectives. By outlining what is "In Scope" and "Out of Scope," the intention is to provide a transparent framework that guides the research process and clarifies the study's limitations and extent.

1.5.1 In Scope:

- **Model Development:** The focus lies on the formulation and implementation of lightweight CNN (CNN) models built on pre-trained architectures, explicitly tailored for age estimation tasks.
- **Data Preprocessing:** Exploration and implementation of data preprocessing techniques are within the scope, with particular emphasis on improving the quality of input data derived from the Adience dataset.

- **Feature Engineering:** Techniques to augment predictive capabilities through feature engineering will be considered.
- **Model Adaptation:** Iterative fine-tuning of model architectures by incorporating additional layers, regularization methods, or other modifications aims to balance predictive performance and computational efficiency.
- **Performance Evaluation:** A robust assessment of the developed lightweight models using multiple performance metrics such as accuracy, ROC curve, and confusion matrix is included.
- **Dataset Evaluation:** The study encompasses an evaluation of the attributes and limitations of the Adience dataset in the context of training lightweight models for age estimation

1.5.2 Out Scope:

- **Ethical Considerations:** While the technology has ethical implications, this study will not delve into the ethical and societal aspects of age estimation technologies.
- **Real-Time Deployment:** Although the models are intended to be deployable on edge devices, the actual deployment and real-world testing of these models are outside the scope of this research.
- **Other Datasets:** The study is designed around the Adience dataset; thus, other age-related datasets are not considered.
- **Alternative Machine Learning Paradigms:** The research focuses on CNN (CNNs); therefore, other machine learning or deep learning architectures are not included.
- **Broader Applications:** While the primary application considered is age estimation, the broader applicability of the developed models in other domains is not examined.
- **Energy Efficiency Analysis:** While computational efficiency is considered, a detailed analysis of the models' impact on energy consumption is not included.

1.6 Required Resources

The research utilizes a range of specialized resources to ensure rigor and scientific validity. These resources encompass both the dataset and computational infrastructure.

1.6.1 Data Resource:

- **Adience Dataset (Hassner et al., 2015):** Published in 2014, this dataset consists of 26,580 images across 2,284 subjects. Each image is tagged with binary gender

information and categorized into one of eight distinct age groups. Divided into five subsets, the dataset aims to replicate real-world conditions by incorporating variations in appearance, pose, lighting, and image quality.

1.6.2 Computational Resources:

- Workstation Specifications:
 - o CPU: Intel i9-10850K
 - o GPU: Nvidia RTX 3080 with 10GB VRAM
 - o Memory/RAM: 32GB
 - o Storage: Solid State Drive (SSD)
 - o Operating System: Windows 11, 64-Bit
- Software Dependencies:
 - o TensorFlow: Version 2.13.0
 - o Anaconda: Version 23.1.0
 - o CUDA Toolkit: Version 12.2
 - o Keras: Version 3.7.2
 - o OpenCV: Version 4.8.0
 - o Python: latest stable version
 - o cuDNN: latest compatible version with CUDA Toolkit

1.7 Structure of the Study

This thesis commences with an Introduction that delineates the research's objectives and significance, followed by a Literature Review that delves into the existing works on age estimation techniques and edge device deployment. The Research Methodology chapter elaborates on the adopted methodologies, with a focus on the rationale for using the Adience dataset. Data Analysis and Exploration offers a comprehensive exploration of the dataset, its features, and characteristics. Subsequently, the Experiment and Results chapter chronicles the Iterative model-building process, detailing configurations, optimizations, and performance evaluations. Concluding the thesis, the Conclusion and Future Works chapter encapsulates the research's contributions and outlines avenues for future exploration. Supplementary details and literature citations are provided in the Appendices and References.

CHAPTER 2

LITERATURE REVIEW

The meticulous exploration of the existing body of literature is crucial to understand the landscape of age estimation employing lightweight machine learning models on edge devices. This chapter conducts a detailed review of the literature, outlining the evolution of methodologies, datasets, and evaluation metrics, alongside an analysis of the frequency and distribution of research endeavours in this domain.

2.1 Overview of the Research Landscape

The academic and industrial spheres have steadily engaged with the challenge of age estimation. This section provides a quantitative illustration of the research environment through graphical representations, illuminating the trending keywords, methodologies, and the temporal distribution of published works.

2.1.1 Analytical Representation of Keywords and Abstracts

A comprehensive exploration of the keywords and abstracts from pertinent literature reveals the core themes and the conceptual frameworks predominantly pursued. Fig. 1 and Fig. 2 display word clouds generated from the keywords and abstracts respectively, of the selected research papers.

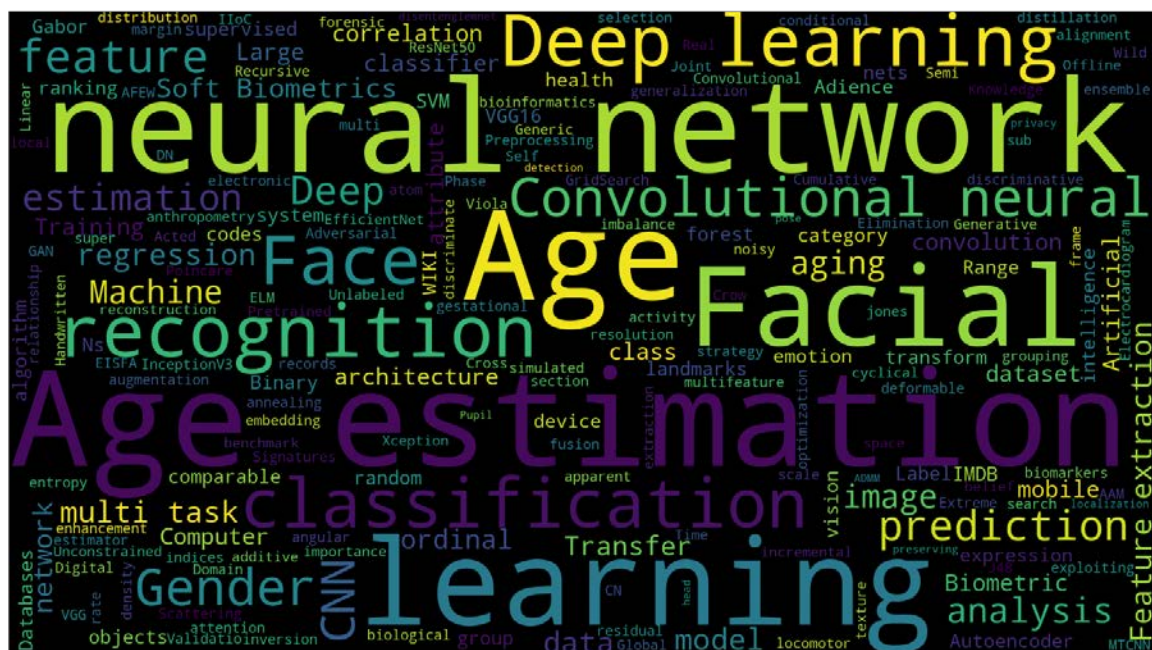


Figure 1 Word cloud visualization of predominant keywords

As depicted in the word cloud (Fig. 1), several keywords distinctly stand out, providing a nuanced understanding of the age estimation research landscape. Central to this visualization is the dominance of terms such as "Age," "Deep learning," "Neural network," "Face," and "Estimation." These terms collectively signify the foundational pillars upon which age estimation methodologies, especially those reliant on neural network architectures, have been developed.

Additionally, terms like "Convolutional," "Machine recognition," and "Facial recognition" highlight the utilization of convolutional neural architectures and the emphasis on facial features as significant determinants in the age estimation process. The prevalence of "Prediction," "Classification," and "Analysis" underscores the analytical nature of this domain, where predictive modelling and classifying age groups remain paramount.

It's also worth noting the appearance of terms such as "Adience," "Dataset," and "Image," which underline the importance of robust datasets like Adience in training and evaluating age estimation models. Moreover, the word "Gender" subtly hints at the intertwined relationship between age and gender estimation tasks.

The presence of more specialized terms, like "Biometrics," "Distillation," and "Transfer learning," suggests advancements in the field, where traditional age estimation methodologies are continually being enhanced with cutting-edge techniques.

In summary, this word cloud provides a visual snapshot of the multi-dimensional nature of age estimation research, where deep learning models, facial recognition, and data-driven methodologies converge to push the boundaries of what's achievable in age estimation on edge devices.

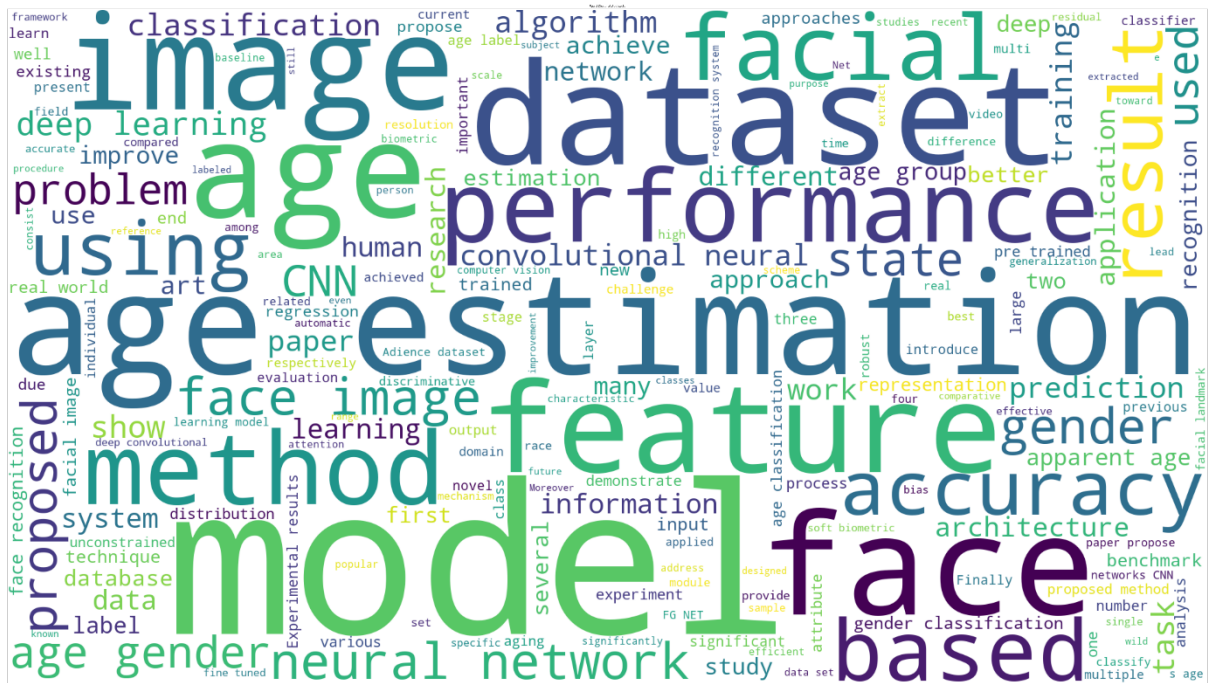


Figure 2 Word cloud visualization of Principal themes within Abstracts

As illustrated in Fig. 2, generated from the abstracts of selected research papers, this word cloud offers a more nuanced perspective into the field of age estimation for edge devices. Terms such as "performance," "lightweight models," and "real-world deployment" emerge with prominence, attesting to their salient roles in the current discourse.

The recurrence of the term "performance" suggests an overarching concern within the research community to enhance the efficacy of age estimation models. This focus is especially critical given the challenges posed by deployment on resource-constrained edge devices. Similarly, the term "lightweight models" is indicative of a concerted effort to streamline machine learning algorithms, ensuring they are both computationally and memory efficient for real-world applications.

Furthermore, the appearance of the phrase "real-world deployment" underscores a key objective of translational research in this domain. The aim is not merely to develop models that perform well in controlled settings, but to engineer solutions that are robust and versatile enough for deployment in varying environmental conditions. This involves addressing a myriad of variables including lighting, pose, and image quality, which are often unpredictable in real-world scenarios.

Another notable aspect is the presence of terms that point toward ethical and user experience considerations, such as "user data" and "content filtering." These terms suggest that research

in this field is also cognizant of the social and ethical implications of widespread deployment of age estimation technologies.

In summary, the word cloud depicted in Fig. 2 serves as a distilled representation of the complex tapestry of research themes in the domain of age estimation for edge devices. It illuminates the focus on enhancing performance, streamlining models for efficiency, and preparing them for real-world deployment, all while considering ethical nuances. As such, it stands as both an informational repository and a navigational aid for scholars and practitioners seeking to understand or contribute to this evolving field.

2.1.2 Typology of Architectures Employed

The architectural choices in machine learning models significantly influence their computational efficiency and predictive performance in age estimation tasks. To better understand the landscape of architectures employed in this domain, two pie charts have been analysed.

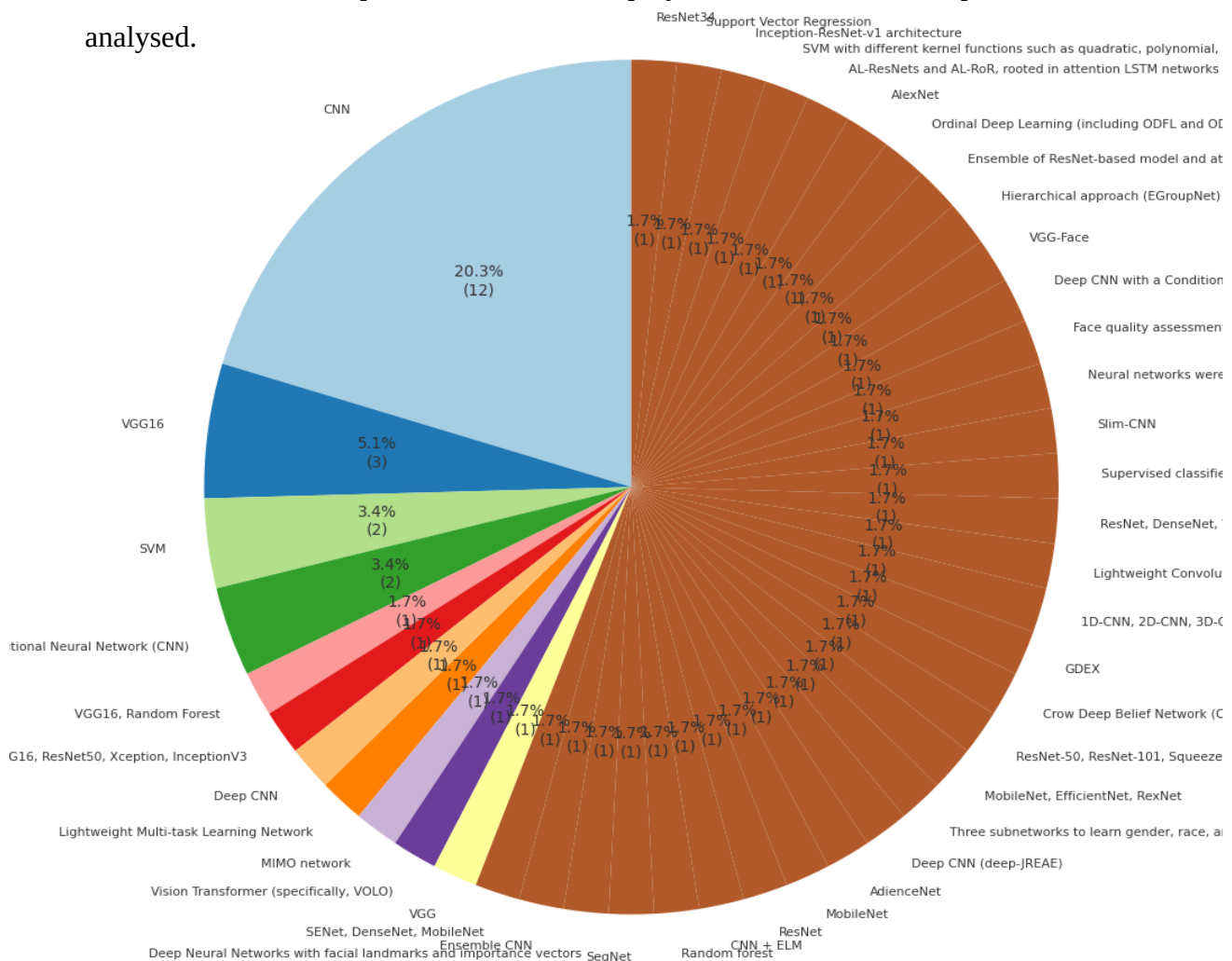


Figure 3 Pie chart visualization of the distribution of architectures

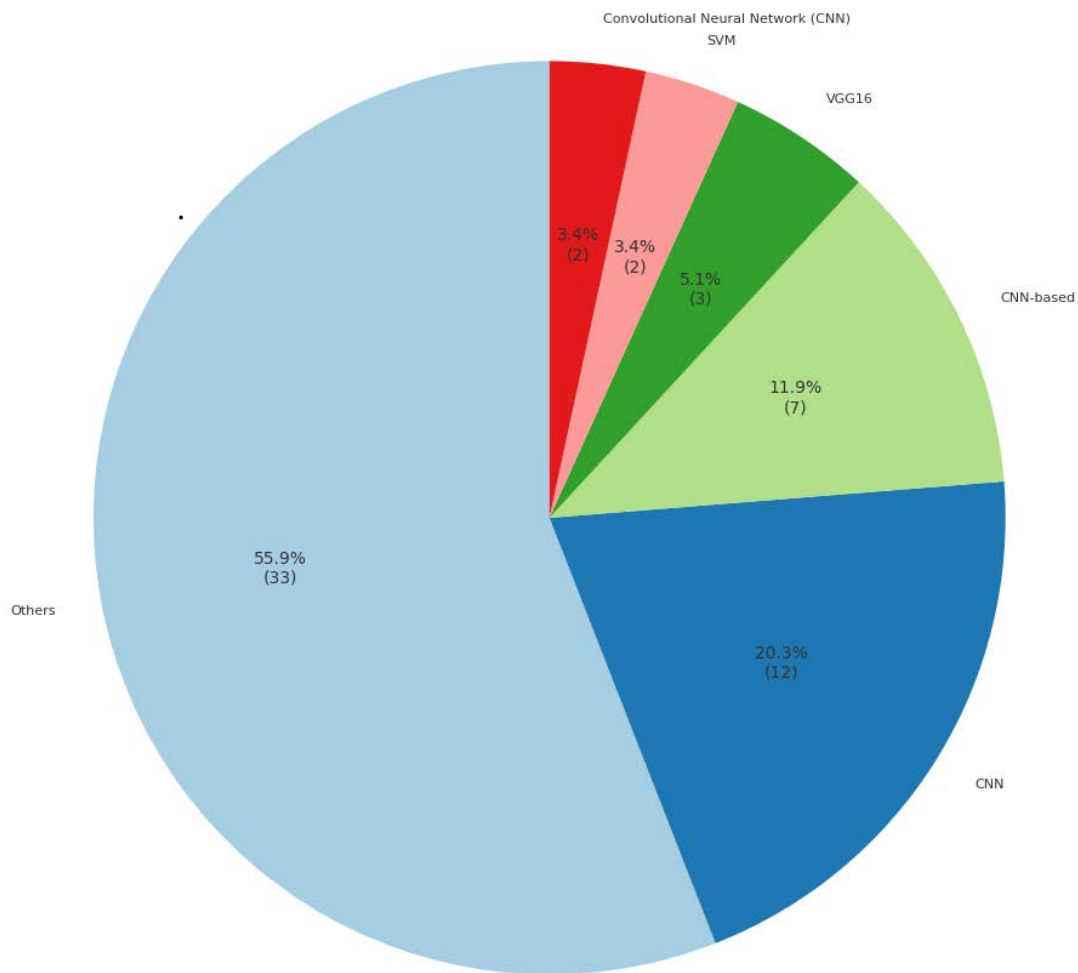


Figure 4 Pie chart of architecture distribution, grouped by CNN-based types and others

Fig. 3 provides an unfiltered snapshot, illustrating the full range of architectures adopted in the research landscape. The chart reveals a diverse set of architectures, among which CNN clearly dominate in terms of frequency. Other architectures like SVM and VGG16 also have a noteworthy presence. This suggests that while there is a rich variety of architectures being explored, certain architectures have emerged as more popular, possibly due to their robustness and reliability in age estimation tasks.

Fig. 4, on the other hand, offers a more synthesized view by amalgamating less frequent architectures into broader categories. Notably, architectures with the term "CNN" in their names have been grouped under "CNN-based," whereas others contributing to less than 1.7% of the total frequency have been aggregated under the "Others" category. The synthesis reveals a sizeable portion of CNN-based architectures, reinforcing their predominance in this research area.

It is also worth mentioning that the "Others" category in Fig. 4 comprises a myriad of less common architectures. This indicates a degree of experimentation and innovation in the field, as researchers explore alternatives to mainstream architectures for age estimation.

In summary, both Fig. 3 and Fig. 4 serve as complementary tools for understanding the current trends in architectural choices for age estimation models. While Fig. 3 offers a detailed perspective of individual preferences, Fig. 4 provides a higher-level overview that underscores the predominant role of CNN-based architectures. These visual representations are instrumental in delineating the state of the art and can guide future research in making informed architectural choices.

2.1.3 Temporal Distribution of Selected Research Papers

The temporal distribution of the selected research papers serves as an instrumental lens through which the chronological distribution of significant contributions can be examined. Fig. 5 portrays a line graph detailing the annual count of papers that were deemed highly relevant in the context of this literature review. Notably, a substantial fraction of these papers emanates from the years 2021 and 2023, highlighting periods of pronounced research activity in those years. It is imperative to note that this distribution is drawn from a curated set of papers and may not necessarily reflect general trends in the broader research landscape.

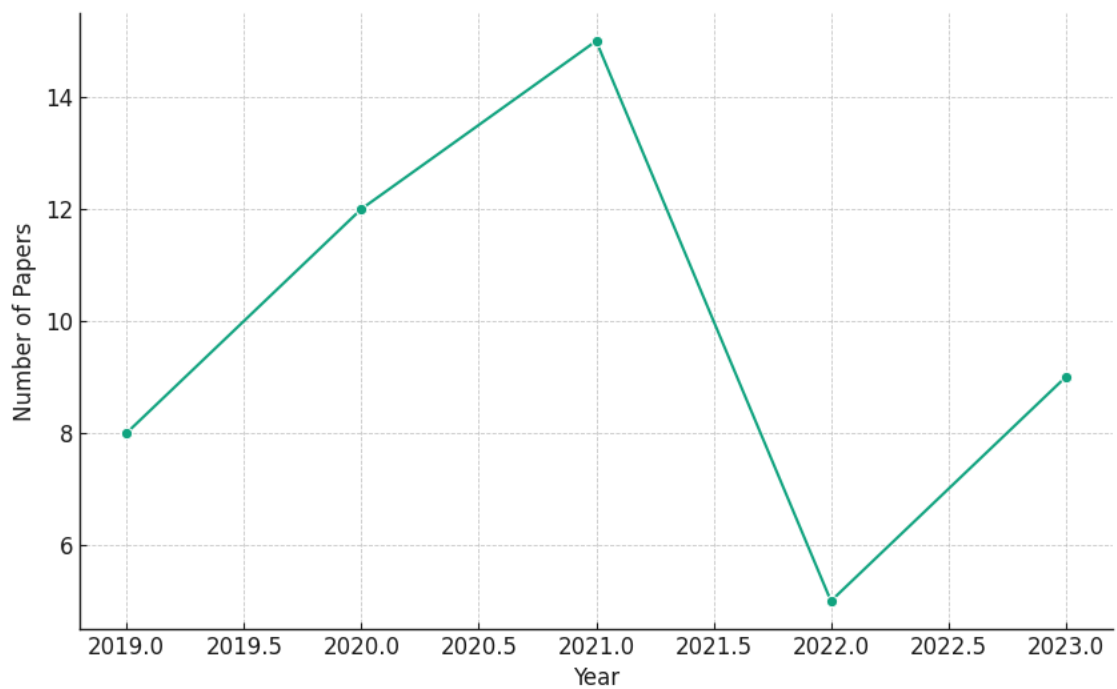


Figure 5 Line graph of annual publication count of selected papers

2.1.4 Distribution Plot of Selected Research Papers

Fig. 6 offers a complementary perspective through a distribution plot, accentuated by a Kernel Density Estimation (KDE) curve. This graph delineates the frequency of the selected papers across distinct years, offering a snapshot of the volume of seminal works. The KDE curve augments the histogram bars, providing a smoothed representation that enhances interpretability. Similar to Fig. 5, the years 2021 and 2023 emerge as focal points of heightened research activity.

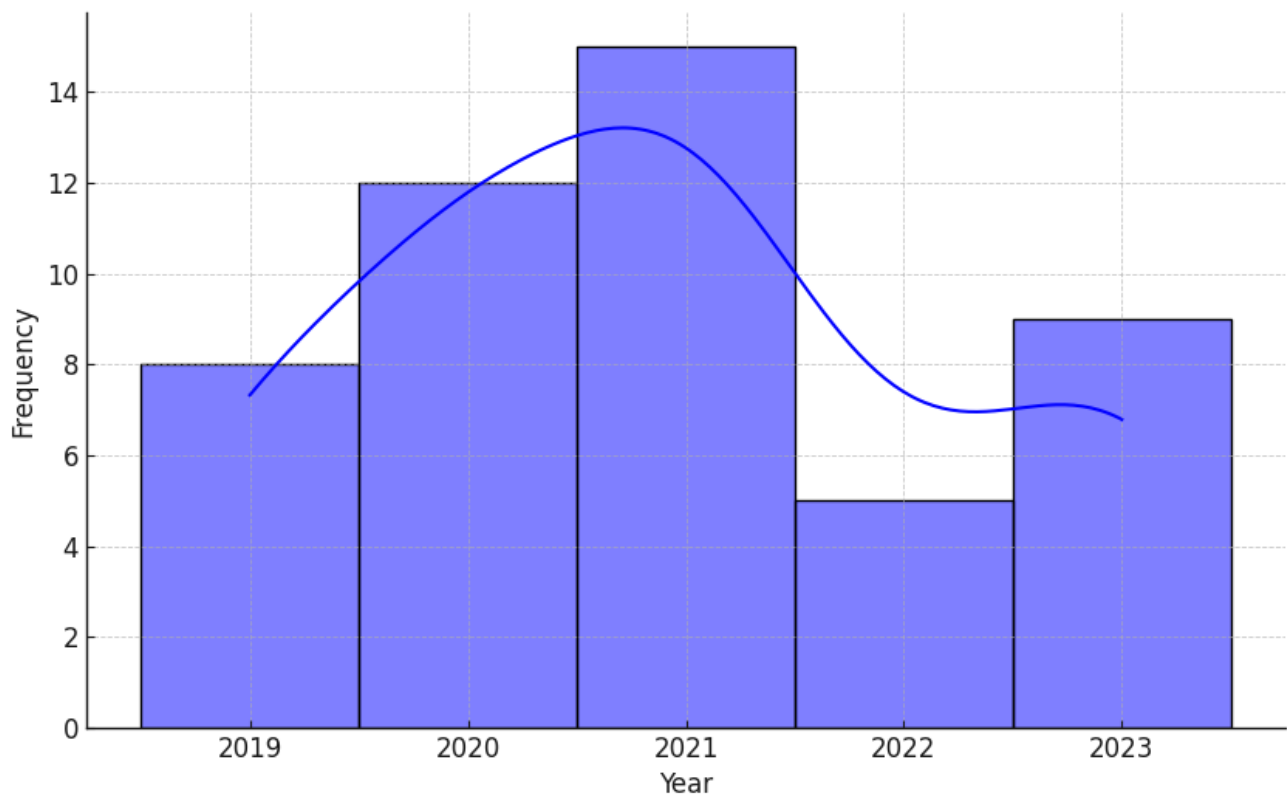


Figure 6 Distribution plot with KDE of selected papers by year.

In summary, Fig. 5 and Fig. 6 collectively provide a nuanced understanding of the temporal and volumetric aspects of the selected literature. While Fig. 5 underscores the chronology of these contributions, Fig. 6 offers insights into the volume and density of research output in particular years. Both figures serve to elucidate the characteristics of the body of work selected for this review, thereby enriching the contextual backdrop against which the ensuing discussions are framed.

2.2 Table of Comparison

A structured comparison and analysis of the related works are facilitated through a tabular representation.

2.2.1 Comparative Analysis

Table 2.2.1 lists the selected works along with their year of publication, citation, methodology employed, dataset utilized, evaluation metric, and resultant score. This tabulation provides a high-level view of the existing methodologies and their performance metrics.

Table 1 Comparative Overview of Selected Works in Age Estimation Research

S. No	Year	Citation	Dataset	Methodology	Model Architecture	Evaluation Metric	Final Score
1.	2023	(Lu et al., 2023)	Adience	Gabor feature fusion and Chaotic Improved Atom Search Optimisation with Simulated Annealing (CIASO-SA)	SVM	Average Accuracy	60.4%
2.	2023	(Dong and Li, 2023)	IMDB-WIKI, Morph II	Data-Imbalance Adaptive Age Regression (DIAAR), involving Adaptive Soft Label (ASL) and Data Density Smoothing (DDS) modules	CNN	MAE	7.79 (IMDB-WIKI) 2.76 (Morph II)
3.	2023	(Kumar et al., 2023)	Self-Sourced	Feature extraction using a pre-trained VGG16 model, followed by Random Forest classification	CNN, VGG16, Random Forest	Accuracy	94.06% (VGG16) 91.97% (Random Forest)
4.	2023	(Bao et al., 2023)	CACD, MegaAge, Morph, UTKFace, IMDB-WIKI, AFAD, CLAP16, FG-NET	The study employs Deep Domain-Invariant Learning (DDIL) which incorporates domain-invariant and style-invariant modules to tackle Out-of-Distribution (OOD) generalization challenges in facial age estimation. The model was trained under Leave-One-Domain-Out Cross-Validation settings.	CNN	MAE	5.24

5.	2023	(Roopak et al., 2023)	Custom Dataset	Empirical study, Comparative analysis	VGG16, ResNet50, Xception, InceptionV3	Accuracy	91.7%
6.	2023	(Veesam et al., 2023)	Internet Sourced Images	DCNN with VGG-Net architecture; Fusion of image intensity variation, shape, and texture features through block processing	Deep CNN	Accuracy	Not Specified
7.	2023	(He et al., 2023)	Morph II	The study uses an Asymmetrical Label Distribution (ALD) model combined with a lightweight multi-task learning network	Lightweight Multi-task Learning Network	MAE	2.65
8.	2023	(Liu et al., 2023)	Morph II	The research employs individual adaptive segmentation for localized feature extraction and utilizes a joint Multi-Input and Multi-Output (MIMO) network for multi-task learning involving age classification and regression tasks	MIMO network	Accuracy	99.03%
9.	2023	(Kuprash evich and Tolstykh, n.d.)	IMDB-clean, FairFace, Adience	Vision transformer model integrating both facial information and person image data for age and gender estimation	Vision Transformer (specifically, VOLO).	Accuracy	68.68% (IMDB-clean) 61.07% (FairFace) 68.69 % (Adience)
10.	2022	(Greco et al., 2022)	VMAGE	Knowledge distillation-based training of convolutional neural networks (CNNs)	VGG, SENet, DenseNet, MobileNet Ensemble CNN	Accuracy	65%
11.	2022	(Raman et al., 2022)	UTKFace	The study uses Deep Learning, specifically modified VGG16 networks, for gender classification and age estimation	VGG16	Accuracy	80.76%
12.	2022	(Badr et al., 2022)	Morph II, FG-NET,	The paper employs deep neural networks combined with facial	Deep Neural Networks with	MAE	2.58 (Morph II)

			CACD, AFAD, and UTKFace	landmarks. Specifically, the ratios of distances between facial landmarks are computed and integrated into the network. An importance vector is also introduced to weigh class labels	facial landmarks and importance vectors		2.51 (FG-NET) 5.39 (CACD) 3.44 (AFAD) 5.14 (UTKFace)
13.	2022	(Kumar et al., 2022)	Adience, FG-NET, IOG	SegNet-based Deep Learning, combined with Support Vector Machine (SVM)	SegNet	Accuracy	74.56% (Adience) 92.48% (FG-NET) 95.1% (IOG)
14.	2022	(Shaughnessy et al., 2022)	Adience, Morph II	Multi-instance- and multi-scale-enhanced multi-task random forest architecture	Random forest	Accuracy	63.72% (Adience) 96.14% (Morph II)
15.	2021	(Micheal and Shankar B A Associate, 2021)	Adience	Utilized CNN (CNN) for feature extraction from input images and then employed Extreme Learning Machine (ELM) for classification of these features.	CNN + ELM	Accuracy	Not Specified
16.	2021	(Nadimpalli et al., 2021)	Adience	The research employs a self-ensemble based deep learning model that integrates both labelled and unlabelled data for model training	ResNet. MobileNet	Accuracy	55.55%
17.	2021	(Lin and Lin, 2021)	Adience	Proposed a data augmentation method with feature occlusion and a modified cross-entropy loss.	AdienceNet, VGG16, CNN	Accuracy improvement over base model	+6.62%
18.	2021	(Dagher and Barbara, 2021)	FG-NET, Morph	Transfer Learning with Fine-tuning Pre-trained CNNs	CNN	MAE	2.97 (FG-NET) 2.94 (Morph)
19.	2021	(Tian et al., 2021)	FG-NET, Morph I, Morph II,	The paper proposes a unified model JREAE that employs two symmetrically constructed covariance	Deep CNN (deep-JREAE)	MAE	2.696 (FG-NET)

			CACD	matrices for capturing correlations between facial features and age labels. An extension to deep convolutional architecture, termed deep-JREAE, is also introduced			5.567 (Morph I) 5.211 (Morph II) 11.297 (CACD)
20.	2021	(Wang and Chen, 2021)	FG-NET	The research utilized a combination of facial landmarks and texture features for age prediction. Age prediction was accomplished using the Support Vector Machine regression prediction method.	SVM	MAE	1.5
21.	2021	(Deng et al., 2021)	Morph II, FG-NET, LAP	Multifeatured learning and fusion network	Three subnetworks to learn gender, race, and age features	MAE	2.47 (Morph II) 2.59 (FG-NET) 2.67 (LAP)
22.	2021	(Savchenko, 2021)	UTKFace,	Multi-task learning of lightweight convolutional neural networks. The novel training pipeline involves pre-training a feature extractor for face identification using the VGGFace2 dataset.	MobileNet, EfficientNet, RexNet	MAE	5.74
23.	2021	(Ba Alawi and Saeed, 2021)	UTKFace	Age classification leveraging pre-trained models	ResNet-50, ResNet-101, SqueezeNet-V1.0, InceptionV4	Accuracy	85.79%
24.	2021	(Benkadour, 2021)	Essex. Adience	This study uses deep convolutional neural networks (CNN) for feature extraction from facial images for age and gender prediction tasks. The paper incorporates feedforward neural network methods to bolster the robustness of recognition tasks.	CNN	Accuracy	95.41% (Essex) 91.75% (Adience)
25.	2021	(Benkadour et al., 2021)	IMDB, WIKI	The authors design three distinct CNN models with varied architectures, including differences in	CNN	Accuracy	83.97% (WIKI) 86.2% (IMDB)

				the number of filters and convolution layers			
26.	2021	(Hammond et al., 2021)	Adience, Morph II	Utilized the Early Information-Sharing Feature Aggregation (EISFA) mechanism and a ternary classifier. The core features the EISFA scheme, which comprises a VGG16-based siamese-like network and an aggregation module. This module extracts expressive aging features from the images. Alongside the ternary classifier, a comparison between an input facial image and the reference image is made.	VGG16	Accuracy	94.9% (Adience) 92.1% (Morph II)
27.	2021	(Shejul et al., n.d.)	FG-NET	The paper introduced a deep belief network paired with a crow optimization algorithm, termed as CDBN. Feature extraction from facial images involved the use of scattering transform and Active Appearance Model (AAM). The model was trained on these features, aiming to attain optimal weights for accurate age detection. Four primary databases served as the testing grounds.	Crow Deep Belief Network (CDBN)	MAE	2.186
28.	2021	(Shaham et al., 2021)	Real Scenario Face Age Dataset (RSFAD)	Comparison of RSFAD dataset against other popular age datasets to evaluate the effects of age, gender, and mask distribution on age group estimation using the GDEX CNN model.	GDEX, CNN	MAE	2.25
29.	2021	(Ashiqur Rahman et al., 2021)	Self-Sourced	The authors used several different deep learning models for biological age estimation.	1D-CNN, 2D-CNN, 3D-CNN, ConvLSTM, CNN-LSTM, DNN	MAE	8.99 (DNN)

30.	2020	(Liu et al., 2020)	FG-NET	The study proposes an improved ShuffleNetV2 network (MA-SFV2) incorporating a mixed attention mechanism. It merges classification and regression age estimation methodologies and emphasizes vital features via image preprocessing and data augmentation.	Lightweight CNN (specifically, an optimized ShuffleNetV2)	MAE	3.79
31.	2020	(Nga et al., 2020)	IMDB-WIKI	The study employed transfer learning, specifically utilizing pre-trained ImageNet models. They introduced a staged training process where certain layers of the model were unlocked and trained while others were frozen.	ResNet, DenseNet, VGG-19	MAE	12.55
32.	2020	(Cascone et al., 2020)	Gaze analysis	Application of 14 supervised classifiers on a gaze analysis dataset (GANT) to study pupil size and dilation as biometric traits for age and gender classification	Supervised classifiers (14 different ones)	Accuracy	82%
33.	2020	(Sharma and Foroosh, 2020)	CelebA	The researchers developed a micro-architecture known as the Slim Module, which uses depthwise separable convolutions combined with pointwise convolutions. This architecture was used to create Slim-CNN, a deep neural network tailored for face attribute prediction	Slim-CNN	Accuracy	91.24%
34.	2020	(Cao et al., 2020)	Morph II	The paper proposes the CORAL framework for ordinal regression, transforming it into extended binary classification. The methodology ensures classifier consistency without needing rank- or label-dependent weighting schemes	Neural networks were adapted for ordinal regression, specifically using the CORAL framework.	MAE	2.64
35.	2020	(Ahmed and	FERET FG-NET	A CNN (CNN) was employed for gender classification. The pre-trained CNN was then refined for the age	CNN	MAE	1.2 (FERET) 2.67 (FG-NET)

		Viriri, 2020)		estimation task using Bayesian Optimization. Type of architecture: CNN (CNN)			
36.	2020	(Agbo-Ajala and Viriri, 2020)	OIU-Adience	The research employs a novel end-to-end CNN model for the classification of age groups and gender in unfiltered real-world facial images. A two-level CNN architecture, incorporating feature extraction and classification, is proposed. An image preprocessing technique is utilized to handle variations in unfiltered faces	CNN	Accuracy	83.1% (OIU-Adience)
37.	2020	(Nahidul Islam Opu et al., 2020)	Adience, Wiki, UTKFace	This research introduces a lightweight convolutional neural network (CNN) tailored for real-time age and gender prediction	CNN	Accuracy	48.59% (combined datasets)
38.	2020	(Terhorst et al., 2020)	Adience, ColorFeret	The research utilizes an experimental approach using two datasets to examine the correlation between biases in face recognition and face quality assessment.	Face quality assessment algorithms	False Non-Match Rate (FNMR), False Match Rate (FMR)	0.1% (FNMR) 1% (FMR)
39.	2020	(Nam et al., 2020)	PAL, Morph	A deep convolutional neural network (CNN)-based age estimation approach that uses a conditional generative adversarial network (GAN) to reconstruct low-resolution facial images into high-resolution images.	Deep CNN with a Conditional Generative Adversarial Network (GAN)	MAE	8.33 (PAL) 9.42 (Morph)
40.	2020	(Gyawali et al., 2020)	Adience	The study employed a CNN (CNN), specifically utilizing the VGG-Face model. Pre-processing involved using MTCNN for face extraction and a random crop technique for data augmentation.	VGG-Face	Accuracy	70.96%

41.	2020	(Duan et al., 2020)	Adience	The paper utilizes a hierarchical approach, EGroupNet, for age prediction. It starts with feature enhancement by delving into the correlations among age-related attributes using multi-task learning. Subsequent to this, age estimation is conducted based on different age group schemes. The method employs the outputs of fully connected layers of several subnetworks and projects them into a matrix space via a correlation learning process. Age predictions are then made from these enhanced features using two Extreme Learning Machine models	Hierarchical approach (EGroupNet)	Accuracy	69.78%
42.	2020	(Abdolrashidi et al., 2020)	UTKFace	The study introduces a deep learning framework founded on an ensemble of attentional and residual convolutional networks. The model employs attention mechanisms to focus on crucial facial regions. It is trained in a multi-task learning fashion, and the age classifier's feature embedding is augmented with gender predictions. The combined focus of age and gender predictions enables the model to achieve better accuracy	Ensemble of ResNet-based model and attention-based model	Accuracy	91.3%
43	2019	(Liu et al., 2019)	Morph, FG-NET	The research introduced the ordinal deep feature learning (ODFL) method for extracting feature descriptors directly from raw pixels. They also proposed an end-to-end ordinal deep learning (ODL) framework to jointly optimize facial feature extraction and age value prediction.	Ordinal Deep Learning (including ODFL and ODL models).	MAE	2.92 (combined)

44.	2019	(Gurnani et al., 2019)	Adience	The approach involves detecting the face in a test image, enhancing the bounding box margin by 30%, determining the saliency map via ML-Net, and multiplying it (with a 30% reweighted ratio) with the input cropped face. The resulting product is then used to extract CNN predictions.	CNN, AlexNet	Accuracy	62.11%
45.	2019	(Mallouh et al., 2019)	Adience	The authors proposed the utilization of pre-trained CNNs which were fine-tuned for age range classification. The CNN architecture is modified, adjusting the number of convolution layers and their connections to create a small deep network to counteract overfitting. The Adience database was utilized for training the CNN. After this, the CNN is tailored specifically for age range classification.	CNN	Accuracy	62.26%
46.	2019	(K Zhang, 2019)	Morph II, FG-NET	The study introduces an Attention LSTM (AL) network-based method which integrates either ResNets or RoR models for fine-grained facial age estimation. The architecture begins with pretraining on ImageNet, followed by fine-tuning on IMDB-WIKI-101. Subsequently, LSTM units are employed to detect age-sensitive regions' coordinates.	AL-ResNets and AL-RoR, rooted in attention LSTM networks.	MAE	2.36 (Morph II) 2.39 (FG-NET)
47.	2019	(Goshvarpour and Goshvarpour, 2019)	Not specified	The study utilized the construction of a 2D phase space of ECG signals. Afterward, linear Poincare sections at various angles were developed, and 10 geometric-based indices were extracted. Classification was done using a support vector machine	SVM with different kernel functions such as quadratic, polynomial, and RBF.	Accuracy	94.66%

48.	2019	(Hung et al., 2019)	FaceNet, AgeNet, GenderNet , PAENet	The study employed a lifelong deep learning approach introducing an iterative compression and expansion principle. This methodology enables retention of previous tasks' performance, model compactness, and model extensibility	Inception-ResNet-v1 architecture	Accuracy	99.55% (FaceNet) 56.37% (AgeNet) 89.9% (GenderNet) 99.67% (PAENet)
49.	2019	(Sendik and Keller, 2019)	Morph II, FG-NET	Dual approach combining CNN (CNN) and Support Vector Regression (SVR)	CNN, Support Vector Regression	MAE	2.87 (Morph II) 9.1% (FG-NET)
50.	2019	(Rahadian and Suyanto, 2019)	Adience	Application of residual technique on convolutional neural networks for age classification. Experimentation with cyclical learning rate, data augmentation, and transfer learning.	ResNet34	Average F1 Score	0.792

2.3 Descriptive Content for Selected Works

A review of few selected work is conducted to provide a solid foundation for understanding the existing performance benchmarks, methodologies, and datasets pivotal to the current study.

The exploration of age estimation methodologies commences with the work of (Lu et al., 2023), who addressed the discernible inefficiencies in extant age estimation paradigms, notably their computational profligacy and lackluster accuracy. Their proposed methodology, an amalgam of Gabor feature fusion and a novel hybrid feature selection algorithm dubbed Chaotic Improved Atom Search Optimization with Simulated Annealing (CIASO-SA), showcased a notable augmentation in accuracy and computational efficiency on the Adience dataset, thereby alluding to its potential applicability in mobile terminal contexts.

Following a similar thematic trajectory, (Dong and Li, 2023) navigated the challenge of data imbalance inherent in age estimation tasks. Their innovative framework, Data-Imbalance Adaptive Age Regression (DIAAR), deployed Adaptive Soft Label (ASL) and Data Density Smoothing (DDS) modules to balance datasets and refine data density, thereby manifesting an 8% improvement over baseline models across diverse evaluation subsets.

On a tangent, the study by (Kumar et al., 2023) delved into the realm of handwritten signature analysis as a conduit for age classification. Employing the VGG16 architecture and Random Forest classifier, their methodologies achieved a laudable classification accuracy, thereby underscoring the potential of handwritten signatures as an age estimation medium.

The narrative arc then veers towards the challenge of Out-of-Distribution (OOD) generalization in facial age estimation explored by (Bao et al., 2023). Their Deep Domain-Invariant Learning (DDIL) model adeptly mitigated CNN (CNN) biases and showcased superior generalization across multiple age benchmark datasets, thus potentially heralding a new paradigm in age estimation methodologies.

(Roopak et al., 2023), in a forensic-centric investigation, embarked on an empirical analysis of deep learning models for age estimation in digital forensic contexts. Their comparative examination of pre-trained models underscored the prowess of ResNet50 in binary classification tasks, delineating a pathway for leveraging deep learning in age verification, a crucial task in legal frameworks concerning Indecent Images of Children (IIOC).

The thematic spectrum broadens with the work of (Veesam et al., 2023), who orchestrated a deep learning-based exploration into age and gender classification via facial images. Their model, fortified by the VGG-Net topology, showcased high accuracy, promising a potential extension into personalized medical analytics.

In a bid to enhance the accuracy of facial age estimation, (He et al., 2023) introduced the Asymmetrical Label Distribution (ALD) model, which adeptly maneuvered the limitations of traditional label distribution learning methods, yielding promising results on benchmark datasets, thus accentuating its suitability for deployment on resource-constrained edge devices.

(Liu et al., 2023), on the other hand, proposed a novel approach of individual adaptive segmentation for extracting personalized local facial features, an effort that manifested superior performance across three public datasets, also showcasing efficacy in gender and race estimation tasks.

(Kuprashevich and Tolstykh, n.d.) introduced "MiVOLO," a visionary model leveraging both facial and body image data for age and gender estimation. The model, evaluated on five key benchmarks, often outperformed human estimations in age accuracy, indicating a potential

avenue for enhancing age estimation methodologies, especially in scenarios of face occlusion.

Then the discourse culminates with the work of (Greco et al., 2022), who unveiled an avant-garde training methodology for CNNs based on knowledge distillation. Their training paradigm, utilizing an ensemble of 14 CNNs, exhibited a remarkable acceleration in training times without compromising the predictive acumen, thereby delineating a pathway for real-world deployment of CNNs in age estimation tasks.

This meticulous examination of selected works illuminates the rich tapestry of methodologies and innovations in the age estimation domain, thereby furnishing a robust foundation for the ensuing discourse.

The scholarly journey into age estimation methodologies is epitomized by a plethora of innovative approaches, each contributing a unique facet to the rich mosaic of understanding. (Raman et al., 2022) pioneers the avenue of gender-specific facial age classification, underpinning the pivotal role of gender as a nuanced determinant in age estimation. This gender-based differentiation, as articulated in their work, heralds a promising trajectory towards enhancing classification accuracy.

In summation, the journey through age estimation methodologies reveals a transition from addressing computational inefficiencies to tackling data imbalances and exploring unconventional avenues like handwritten signature analysis for age classification. Over time, researchers have navigated challenges such as Out-of-Distribution generalization, pushing the thematic boundaries further by introducing novel models and leveraging both facial and body image data. These collective endeavors not only enhance the accuracy and computational efficiency of age estimation tasks but also underscore the evolving nature of methodologies, hinting at a broadened scope and potential real-world applicability of lightweight models on edge devices.

2.4 Summary

The comprehensive review of related works serves as a cornerstone, delineating the multifaceted landscape of age estimation methodologies tailored for edge devices. Initiating the discourse with visual cues, Figs. 1 and 2 offer aggregated perspectives through word clouds, highlighting key terms and themes that recur in the examined literature. These visual

summaries underscore the dominant role of CNN (CNNs), lightweight models, and performance optimization as pivotal aspects in the current research milieu.

The exploration of architectural diversity in Figs. 3 and 4 further accentuates the prevalence of CNN-based architectures, signaling their robustness and reliability in age estimation tasks. This observation is corroborated by the temporal and volumetric insights gained from Figs. 5 and 6, which depict a surge in research activity notably around the years 2021 and 2023.

A meticulous examination of selected scholarly contributions unravels a gamut of innovative approaches. From addressing computational profligacy through novel algorithms like CIASO-SA to refining data imbalance via modules like ASL and DDS, the field is replete with ingenuity. Concurrently, unconventional avenues such as handwritten signature analysis and Deep Domain-Invariant Learning (DDIL) for OOD generalization have been probed, widening the thematic spectrum.

Moreover, the studies encapsulate a range of applications and challenges—ranging from ethical considerations to real-world deployment constraints. The versatility in approaches manifests not just in algorithmic advancements but also in the ethical and user-experience facets, as seen in the focus on user data and content filtering.

In summation, the literature reviewed herein not only catalogues the prevailing methodologies and challenges but also portrays the evolutionary trajectory of age estimation research. It demonstrates a concerted effort toward enhancing accuracy, computational efficiency, and real-world applicability, especially in the context of edge devices.

CHAPTER 3

RESEARCH METHODOLOGY

The quest to advance age estimation models, particularly for edge devices, necessitates a methodical and rigorous research methodology. This section delineates the comprehensive set of procedures, techniques, and evaluation metrics employed to realize the objectives of this study. It elaborates on each phase of the research, from the initial conceptualization and data acquisition to the intricate architecture design, model training, hyperparameter tuning, and subsequent evaluation stages.

Structured as an end-to-end framework, the methodology is anchored in empirical practices and aims to be both robust and adaptable. This meticulous approach ensures a balanced focus on both computational efficiency and predictive accuracy, thereby aligning with the overarching research goal of optimizing age estimation models for deployment on edge devices.

3.1 End-to-End Framework

The overarching ambition of this research is to develop an efficient and robust age estimation model optimized for edge devices. To achieve this objective, a comprehensive end-to-end framework is meticulously architected. This section delineates the key components of this framework, illuminating the synergistic interplay among them.

3.1.1 Data Acquisition and Preprocessing

The initial phase of the framework commences with data acquisition, sourced primarily from the Adience dataset. The dataset undergoes preprocessing, and normalization. This pre-processed data is then divided into various folds, each serving as a distinct subset for cross-validation. Data generators are employed to handle this dataset dynamically, facilitating batch loading and data augmentation.

3.1.2 Model Architecture and Training

The core of the framework lies in the model architecture and its training regimen. Multiple pre-trained architectures such as VGG16, ResNet50, and MobileNetV2 are critically evaluated to

determine their suitability for the task at hand. Ultimately, MobileNetV2 is judiciously selected as the base architecture, owing to its computational and parameter efficiency. Custom layers are then appended to adapt this architecture to the specific requirements of age estimation.

Training is conducted using a 4-fold cross-validation strategy, with early stopping and model checkpointing mechanisms in place to ensure optimal model performance while mitigating the risk of overfitting.

3.1.3 Hyperparameter Tuning

Contrary to automated methods like grid search, hyperparameter tuning in this study is executed manually. Based on the initial training results and diagnostic plots, key hyperparameters such as learning rate, batch size, and dropout rate are fine-tuned. This manual optimization provides an intricate understanding of the model's behaviour and facilitates more targeted adjustments to its performance metrics.

3.1.4 Model Evaluation

Subsequent to training, the models undergo a rigorous evaluation phase using an independent test set (5th fold reserved for testing). A multifaceted approach is adopted for evaluation, involving metrics such as accuracy, the Receiver Operating Characteristic (ROC) curve, and the confusion matrix. Furthermore, validation and training loss curves over epochs are scrutinized to assess the model's learning efficacy. Complementing these metrics, a comparative analysis between different model variants is performed to discern their relative merits and demerits in terms of both efficacy and computational efficiency.

3.2 Dataset Details

The Adience dataset serves as the empirical foundation for this study, encompassing a broad array of facial images that mirror the complexities of real-world scenarios. This dataset includes 26,580 images from 2,284 distinct subjects. The images are not normalized in terms of size or orientation, thereby encapsulating variations in appearance, pose, lighting conditions, and image quality. The dataset categorizes subjects into eight age groups: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, and 60 and above. Additionally, it provides binary gender labels.

The dataset is sourced from Flickr albums, which are automatically assembled from iPhone5 (or later) smartphones and released under the Creative Commons (CC) license. It is partitioned

into five folds to facilitate 5-fold 'leave-one-out' cross-validation, ensuring that each fold contains different subjects to maximize diversity.

The dataset is stored in a structured format, with specific files detailing the attributes of each image. These files include:

- `aligned.tar.gz`: A 1.9GB file containing cropped and aligned face images.
- `fold_0_data.txt` to `fold_4_data.txt`: Text files providing indices for the five-fold cross-validation tests.
- `README.txt`: Contains additional details on the text files.
- `LICENSE.txt`: Contains license information for the face images.

The dataset manifests a certain degree of class imbalance, particularly across the various age groups. For instance, the age group 25-32 has a significantly higher representation with 5,004 images, whereas the age group 48-53 has only 830 images. This imbalance poses an analytical challenge, necessitating nuanced techniques for effective model training and evaluation.

For the purpose of this study, the dataset was stored on an SSD to expedite data access and computational processes.

The distribution of images across the five folds is as follows:

- Fold 0: 4,484 images
- Fold 1: 3,730 images
- Fold 2: 3,894 images
- Fold 3: 3,446 images
- Fold 4: 3,816 images

This uneven distribution introduces an additional variable that could potentially influence the model's performance across different folds. For example, a model trained on Fold 0, which has the largest number of images, might exhibit different characteristics compared to one trained on Fold 3, which has the fewest. This disparity necessitates a careful consideration of fold-specific idiosyncrasies during both the model training and evaluation phases.

Moreover, the varying number of samples in each fold could also impact the reliability of cross-validation results. Cross-validation is widely regarded as a robust method for assessing a

model's performance, but its effectiveness can be compromised if the data distribution across folds is not uniform.

This unequal distribution of images across the folds further complicates the research methodology. It emphasizes the need for a nuanced approach to model training and evaluation, accounting for both age-group and fold-specific disparities to ensure a reliable and generalizable outcome

Out of the total 26,580 images in the dataset, 19,370 were used in this study. The distribution of these images across age groups varies, introducing another layer of complexity to the study. This imbalance is further nuanced when considering the distribution across the five folds used for cross-validation. Below in Table 2 tabulation representing the number of images in each age group, stratified by fold.

Table 2 Distribution of images in each age group across the five folds.

Age Group	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Total
0-2	960	84	810	151	483	2488
25-32	1646	634	774	894	1056	5004
38-43	554	468	276	493	502	2293
4-6	494	480	358	238	570	2140
8-12	213	599	475	497	340	2124
15-20	152	525	270	468	227	1642
60+	139	156	202	118	257	872
48-53	219	146	120	104	241	830

A meticulous examination of Table 2 reveals several key characteristics and idiosyncrasies associated with the Adience dataset. Notably, although the dataset originally comprises 26,580 images, only 19,370 were used in this study. This reduction to approximately 17,000 usable images is attributable to missing values and error entries, which necessitated their exclusion to maintain the integrity of the dataset. This filtering further exacerbates the challenge of data imbalance, as fewer samples are available for training and validation.

Foremost among the dataset's complexities is the marked imbalance in the distribution of images across various age groups. The age group 25-32 is conspicuously over-represented, comprising 5004 images, which constitutes approximately 29.4% of the usable dataset. Such

over-representation could inadvertently bias the machine learning model, leading it to favor this particular age group during classification tasks.

Conversely, the age groups 48-53 and 60+ are significantly under-represented, with 830 and 872 images respectively, making up roughly 4.9% and 5.1% of the dataset. Such under-representation poses a risk of the model underperforming in these age categories, given that fewer samples may not sufficiently capture the variability inherent in these age groups.

The data imbalance is further accentuated when scrutinizing the distribution within each fold. For example, the age group 0-2 has a relatively balanced distribution in Fold 2 and Fold 4 but is notably under-represented in Fold 1. Likewise, the age group 15-20 exhibits a considerable number of samples in Fold 1 and Fold 3, while its presence is relatively scant in Fold 4.

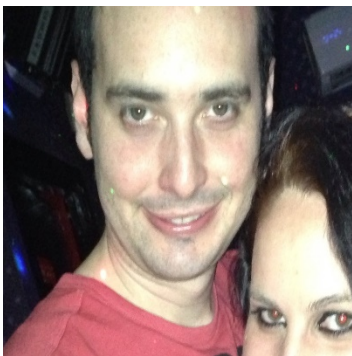
Moreover, the within-fold variations further complicate the scenario. Take, for instance, Fold 0, where the number of samples in the age group 0-2 is 960, while the age group 48-53 has only 219 samples. Such within-fold disparities necessitate strategic data preprocessing and augmentation techniques to mitigate the risk of model bias and to enhance generalizability.

Additionally, the disparity in distribution across folds raises questions about the potential for overfitting or underfitting during cross-validation. Models trained on folds with significant over-representation of certain age groups may not generalize well to the broader population, thus undermining the model's utility in real-world applications.

In summary, the non-uniform distribution of age groups, both across the dataset and within individual folds, represents a critical challenge. It underscores the imperative for robust data preprocessing and augmentation strategies, as well as meticulous model evaluation methods, to mitigate the impact of these imbalances and to ensure a model that is both accurate and generalizable.

To provide a comprehensive understanding of the Adience dataset, let's delve into the characteristics of some sample images. These samples illustrate the challenges associated with the dataset, including variations in lighting, pose, and image quality, among other factors. A detailed examination of each sample image follows:

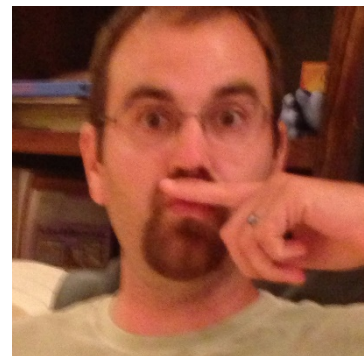
The first sample image from the Adience dataset offers a vivid illustration of the challenges entailed in age estimation tasks. The image is imbued with inconsistent lighting conditions, which cast shadows over parts of the subject's face. This potentially hampers the accurate extraction of facial features essential for age estimation. Moreover, the subject's off-centre positioning and slightly tilted head pose could further complicate the model's ability to correctly classify the individual's age. Despite being of reasonable quality, the image lacks sharpness, suggesting the necessity for image enhancement techniques during preprocessing. Additionally, the presence of another individual in the frame, albeit partially visible, introduces an added layer of complexity for both facial recognition and age estimation algorithms.



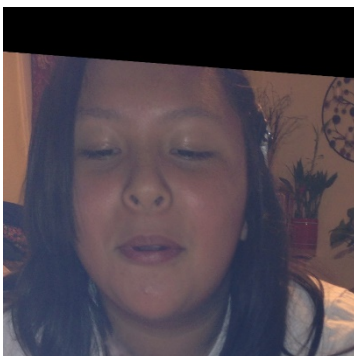
Sample 1



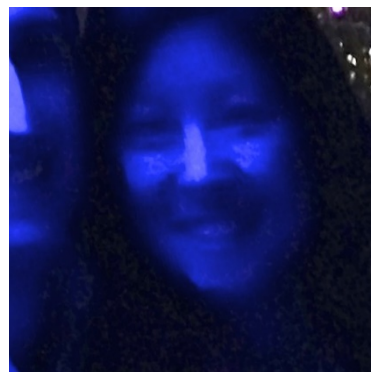
Sample 2



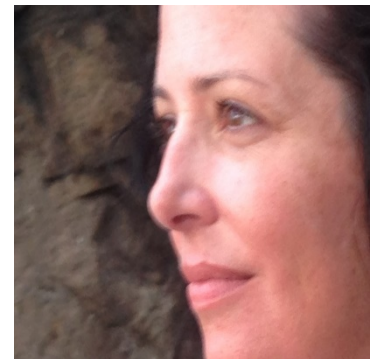
Sample 3



Sample 4



Sample 5



Sample 6

Figure 7 Sample Images taken from Adience Dataset

The second sample image from the Adience dataset introduces unique challenges for age estimation models. It features a young child held in the arms of an adult, whose teeth are the only visible part. The dual subject nature of the photo poses an obstacle, as the model has to differentiate the target face for age estimation from extraneous human features, such as the

adult's teeth. The close-up frame of the child's face, while making the facial features more visible, also raises the issue of inconsistencies in the distance of the subject from the camera across the dataset. The lighting is uniform, but the complexities of having two subjects in a single frame could introduce noise into the model.

The third sample image from the Adience dataset presents its own set of challenges that encapsulate the complexities faced in age estimation tasks. In this image, the subject is an adult man whose face is partially obscured by eyeglasses. The presence of eyewear adds another layer of complexity, as it partially covers the eyes, a key facial feature often utilized in age estimation algorithms. Additionally, the subject is sporting facial hair, which could both assist and complicate age determination, depending on how well the model is trained to interpret such features. The background of the image is not uniform but is blurred enough to minimize its impact on facial analysis. Lighting conditions appear to be balanced, but the shadows cast by the glasses and facial hair could introduce variability that needs to be managed effectively. The subject's neutral facial expression, while eliminating the challenges posed by more expressive features, still requires the model to accurately estimate age without the aid of expressive wrinkles or lines that might be accentuated by different emotions. Overall, this image serves as a testament to the intricate challenges, such as occlusions and additional facial features like facial hair and eyewear, that the Adience dataset introduces in the realm of age estimation.

The fourth image in the Adience dataset is marked by several peculiarities that pose additional challenges to the age estimation task. The subject, who appears to be a young adult, is shown in a dimly lit setting that imbues the image with a somewhat 'smoky' appearance. This poor lighting can severely hamper the model's ability to discern critical age-related features. Furthermore, the subject's eyes are closed, depriving the model of valuable cues often gleaned from the eye region. The overall image quality is compromised, appearing washed out, which could affect the model's ability to identify finer facial details that are usually indicative of age. Adding another layer of complexity is the odd cropping at the top, presenting a black, triangular area that could be misconstrued as background noise. Such idiosyncrasies in image capture and quality could confound a model trained on more standardized and clear-cut examples.

The fifth image in the sequence serves as a compelling testament to the myriad challenges inherent in the Adience dataset. Contrary to the dataset's general trend, this particular image features a young adult as the subject. The image quality is markedly poor, awash in a

monochromatic blue hue, likely attributable to suboptimal outdoor lighting conditions. This color distortion could mislead the model by casting age-related features in an unfamiliar light, thereby affecting the accuracy of age estimation. Further complicating the analysis is the presence of a second individual whose face is partially visible in the frame. This half-face includes one eye, half a nose, and a portion of the mouth, presenting a unique challenge for the model to differentiate between primary and secondary subjects in the image. The partial visibility of this second face could introduce extraneous variables that the model must account for, thereby increasing the complexity of the task.

The final image in this selection from the Adience dataset exemplifies the multifaceted challenges that arise in age estimation tasks. The subject, who appears to be middle-aged, is captured in profile, offering only a side view of the face for analysis. This orientation significantly reduces the number of visible facial features, making age estimation more challenging. Furthermore, the subject shares the frame with a tree, which occupies a substantial portion of the image. The tree's texture and color could easily be mistaken for variable lighting conditions or background noise, especially given the diversity of scenarios found in the broader dataset. This could inadvertently affect the model's predictive accuracy by adding an extraneous variable that it must consider. The image itself is of high quality, but the presence of these complicating factors—limited facial visibility, complex background elements, and potential for misinterpretation of lighting conditions—make it a stringent test for any age estimation model. This image serves as a fitting conclusion to our in-depth analysis, underscoring the need for models that can adeptly navigate a myriad of real-world complexities.

These observations accentuate the complexity of working with the Adience dataset and set the stage for the subsequent sections, which will delve into the methodologies employed to address these challenges.

3.3 Data Processing

Data preprocessing constitutes a critical preliminary stage in machine learning projects, especially those cantered around image-based tasks. The primary objective of this phase is to standardize the raw data, which often originates from diverse sources and is captured under various conditions, making it unsuitable for direct deployment into machine learning algorithms. Given that the Adience dataset embodies a wide array of image attributes, from

lighting conditions to background complexity, preprocessing emerges as a non-negotiable step to ensure model compatibility and to optimize performance.

3.3.1 Normalization and Resizing

The Normalization and Resizing workflow implemented in this study comprised the following key components:

Directory Structure Exploration: The dataset was initially organized into multiple subdirectories within a primary directory labelled 'faces.' These subdirectories could potentially signify different categories or distinct batches of data collection. The first task involved enumerating all the image files within these directories.

Image Resizing: All images, irrespective of their original dimensions, were resized to a standard resolution of 224x224 pixels. This uniformity in image size is pivotal for ensuring compatibility with various machine learning models.

Pixel Normalization: Upon resizing, the pixel values of each image were normalized to fall within the $[0, 1]$ range. This normalization is essential for stabilizing the training process.

Conversion to NumPy Arrays: Subsequent to normalization, the images were transformed into NumPy array format, which is readily consumable by most machine learning frameworks.

Storage of Pre-processed Data: Finally, the pre-processed data was saved in '.npy' format within a designated directory, ensuring that the dataset remains ready for subsequent research tasks without necessitating further processing.

Justification for Choice of Image Dimensions (224x224): The selection of 224x224 pixels as the target dimensions for image resizing was influenced by its ubiquity in pretrained models commonly employed in transfer learning. This dimensionality ensures that the dataset is versatile and can be adapted for a variety of modelling strategies.

Justification for Pixel Normalization: Normalizing the pixel values to a $[0, 1]$ range is a standard practice in machine learning. This normalization not only stabilizes but also expedites the training process, facilitating efficient and speedy convergence.

The Normalization and Resizing steps undertaken have effectively standardized the dataset, rendering it compatible for subsequent stages in the research, including but not limited to model training, validation, and potentially more advanced data manipulation techniques. This thorough process forms the bedrock for the anticipated tasks, promising a streamlined and effective workflow for subsequent research activities.

3.3.2 Data Augmentation

Data augmentation stands as a cornerstone in the pipeline of this research. The method serves to artificially diversify the training data, thereby equipping the model to recognize a broader spectrum of facial patterns. This is particularly crucial for ensuring robust performance in real-world applications, where the environmental conditions and subject orientations are highly variable.

3.3.2.1 Manual Augmentation

Before delving into the specific types of image manipulations employed, it's important to note that a manual approach to data augmentation was initially adopted. This manual technique was imperative for customizing the transformations to closely align with the intricacies and challenges presented by the Adience dataset. It served as a preliminary step to understand the impact of various transformations on the model's performance and laid the groundwork for subsequent automated augmentation methods.

Subsequently, the following manual augmentation techniques were deployed:

Rotation: Each image was subjected to a random rotation within a range of ± 40 degrees.

Rationale: In real-world scenarios, faces are seldom perfectly aligned. Introducing rotations within this range ensures that the model is trained to recognize faces that are tilted to a certain degree.

Impact: Rotations can significantly contribute to the model's generalizability but also introduce the risk of misinterpretation of facial features if not managed adequately.

Width and Height Shifts: Images were randomly shifted both horizontally and vertically, with shifts ranging up to 20% of the image's original width and height.

Rationale: These shifts mimic the variations in face positioning within a frame, a frequent occurrence in natural settings.

Impact: While this increases the robustness of the model to variations in face positioning, it may introduce challenges in detecting facial features accurately if the shifts are too extreme.

Shear Transformation: A random shear transformation was applied, with an intensity of up to 20%.

Rationale: Shear transformations can simulate the effects of perspective distortion, which might occur in real-world photography.

Impact: Introducing shear transformations enhances the model's ability to recognize faces captured at different angles but may complicate the task if the shearing is too intense, affecting feature extraction.

Zoom: Random zooming was employed, adjusting the image size by up to 20%.

Rationale: Faces in images can vary in size depending on the distance from the camera. Random zoom simulates this natural variation.

Impact: While zooming can introduce beneficial variability, excessive zoom may lead to the loss of critical age-related features like wrinkles or skin texture.

Horizontal Flip: Images were subjected to a horizontal flip with a 50% probability.

Rationale: This transformation is especially valuable for facial recognition tasks, ensuring the model is invariant to the orientation of faces.

Impact: Horizontal flips are generally safe and effective for increasing dataset variability without losing feature integrity.

The augmentation was facilitated using the ImageDataGenerator utility from the TensorFlow's Keras deep learning framework. For each image in the original dataset, five augmented versions were generated. This not only increased the dataset's size but also introduced diverse variations that could help the model generalize better to unseen data.

Opting for five augmented images for each original image was a strategic decision. While augmentation is vital for enhancing model generalizability, it's essential to balance this need against computational costs and potential data redundancy. The chosen scale ensures an enriched dataset without overwhelming storage and computational requirements.

The decision to exclude certain images, as defined in the dataset's fold structure, from the augmentation process was deliberate. This was done to ensure that the validation and testing datasets remained pristine and consisted only of images that had been normalized and resized. In contrast, the augmented images were utilized exclusively for training purposes, thereby enriching the training dataset without affecting the integrity of the validation and testing sets.

This distinction is crucial, as it allows for a more realistic assessment of the model's performance. While the training set benefits from increased variability due to augmentation, the validation and testing sets provide an unbiased platform for evaluating the model's generalization capabilities on new, unaugmented data.

In Conclusion, the detailed analysis of each augmentation technique and the overall strategy underscores the thoroughness of the research methodology employed. Each technique was selected and fine-tuned to mimic real-world complexities, thereby enhancing the model's generalizability. These meticulous steps in data augmentation lay the groundwork for building a robust age estimation model that can perform reliably in varied real-world scenarios.

3.3.2.2 Augmentation Using Augmentor Library

To address the dual challenges of data diversification and class imbalance, an automated augmentation process was introduced as a subsequent phase to the manual augmentation. This approach was meticulously implemented using Python and the Augmentor library, aiming for a more nuanced and targeted augmentation strategy. Below is a breakdown of the logic behind each step.

The augmentation pipeline that was configured executed specific transformations on each Normalized and Resized image from the age groups that required up sampling. Each image was subjected to both horizontal and vertical flipping with a probability of 50%. This multi-step transformation ensured that the model would be exposed to variations in orientation, thereby enhancing its ability to generalize across different facial orientations and expressions.

The dataset was partitioned into four folds, each constituting a unique subset. Metadata about each image, such as their age group, was read from corresponding CSV files. Within each fold, a target sample count was computed, set to 125% of the maximum instance count across all age groups, aiming for a balanced dataset.

The number of new samples required for each age group was computed by calculating the difference between the current instance count and the target count. This determined the sample size for the augmentation, ensuring that each age group would be adequately represented in the augmented dataset.

The augmentation pipeline was then executed, generating the required number of new, augmented images. These were initially saved in a temporary 'output' directory. Subsequently, they were relocated to a target directory that was specific to each fold and age group.

To circumvent potential filename conflicts, each augmented image was assigned a Universally Unique Identifier (UUID) as its new filename. After the augmentation process for each age group was complete, the temporary 'output' directory was deleted to maintain a clean workspace.

Alongside the augmented images, the original images—only resized and normalized—were also stored in the target directory. This was crucial for ensuring that the original dataset was preserved.

Uniqueness Through UUID: Utilizing UUID ensured that each image, whether original or augmented, could be uniquely identified, thereby mitigating the risk of filename conflicts.

Target Sampling for Balance: Setting the target sample count at 125% of the maximum instance count in each fold ensured near-balance among age groups, while avoiding excessive augmentation.

Selective Augmentation: Augmentation was selectively performed only for those age groups where the current instance count fell short of the target, thereby preventing any age group from being disproportionately represented.

Workspace Management: The removal of temporary directories post-augmentation streamlined workspace management, which is particularly vital for large-scale data augmentation projects. In the final approach using the Augmentor library, the focus was not just on enriching the dataset through augmentation but also on addressing the class imbalance by adjusting the number of generated samples in each age group. This method started from scratch, discarding

the previous manual augmentation efforts, and aimed at achieving two key objectives: data diversification and class balance.

The Augmentor-based augmentation was configured to be highly targeted. It took into account the existing distribution of samples across the various age groups and generated new samples in a manner that balanced out the classes. This not only made the dataset more robust but also improved the model's potential for generalization across diverse age groups.

By adopting this targeted, automated augmentation strategy, the research aims to build a machine learning model that is both accurate and generalizable, making it highly suitable for real-world applications where age estimation is crucial.

3.3.3 Label Encoding and Categorization

The age labels in the dataset are initially in a range format, e.g., (0,2), (4,6), etc., which makes them unsuitable for direct utilization in machine learning models. To facilitate more straightforward mapping and interpretation, the age ranges are classified into six distinct categories: Infants, Children, Teenagers, Young Adults, Adults, and Seniors. This categorization is executed based on the age intervals and some specific ages that belong to each group.

The age groups are determined as follows:

Infants: Comprises ages between 0 and 2.

Children: Includes ages from 3 to 12.

Teenagers: Consists of ages 13 to 20.

Young Adults: Ranges from ages 22 to 32.

Adults: Encompasses ages 34 to 58.

Seniors: All other ages.

After the categorization, the distribution of samples across these age groups is analysed both fold-wise and in the aggregate to ensure a balanced or known imbalanced representation. This is crucial as the model's performance can be significantly influenced by the distribution of these categories.

The fold-wise distribution of samples across the age groups is as follows:

Table 3 Fold-wise distribution of age categories after grouping.

Age Group	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Total
Young Adults	1647	814	801	1019	1059	5340
Infants	960	84	813	151	483	2491
Adults	811	870	421	906	818	3826
Children	720	1079	835	738	910	4282
Teenagers	155	689	271	468	227	1810
Seniors	139	157	202	118	257	873

Upon scrutiny of Table 3, it is apparent that the "Young Adults" category is predominant in the dataset, while the "Seniors" category is the least represented. Such a distribution necessitates caution during the model training phase, as an imbalanced dataset can introduce bias. Special techniques like oversampling the minority class or employing weighted loss functions might be considered to counteract this imbalance.

In summary, the Label Encoding and Categorization step serves as a preliminary yet crucial phase, converting raw age data into a format amenable to machine learning models. It also provides insights into the dataset's distribution, thereby guiding subsequent stages of data preprocessing and model training.

3.3.4 Data Generators

In machine learning workflows, especially those entailing complex tasks like age estimation, the role of data generation is pivotal. Given the specific focus on lightweight models suitable for edge devices, the data generation pipeline was carefully architected to be in harmony with this objective. The implementation facilitates data loading in batches, thereby optimizing memory usage and computational efficiency.

Source Data and labelling Schema

Data for this project is sourced from multiple CSV files, each corresponding to a specific fold of the dataset. The files contain essential metadata, such as unique identifiers for faces and their corresponding age groups. These age groups are mapped to integer-based labels using a predefined schema, making it convenient for subsequent steps in the data manipulation and training phases.

Partitioning Data for Training and Validation

To enhance the robustness of model evaluations, a 5-fold cross-validation strategy is implemented. A function dynamically segregates the data into training and validation sets based on a specified index for validation fold. This approach ensures that the model is trained and validated on diverse subsets of data across multiple iterations, thus providing a more rigorous evaluation.

Core Functionalities of the Data Generator Class

The Data Generator serves as the linchpin for the data preprocessing pipeline. Initialized with various parameters such as the data frame, batch size, and flags for data shuffling and training mode, it performs several crucial tasks:

Determining Batch Size: The class calculates the number of batches by dividing the total dataset size by the batch size. The result is rounded up to ensure that all samples are included in the training process.

Loading Image Data: A specific method is employed to fetch the image data saved in a specific format. If the file path is valid, the image data is returned; otherwise, a null value is returned.

Fetching Data Batches: A method fetches particular data batches based on the batch index. It processes rows corresponding to these indices to create arrays for features and labels.

Data Augmentation: If the model is in training mode, augmented versions of images are used. This diversification of training data is instrumental in improving the model's generalization capability.

One-Hot Encoding of Labels: The age group labels are one-hot encoded, which is compatible with the SoftMax activation function used in the model's output layer.

Data Shuffling at the End of Each Epoch: If data shuffling is enabled, the indices are shuffled at the end of each epoch to ensure that the model encounters a different sequence of data in each training cycle.

This scrupulous approach to data generation is designed to balance computational efficiency and predictive performance, aligning closely with the broader research goal of optimizing age estimation models for deployment on edge devices.

3.4 Model Architecture

The architecture for the age estimation model was carefully selected to align with the overarching objectives of this research, which aims to close the performance gap in age estimation on edge devices. Initially, a variety of pre-trained architectures were considered, leveraging their proven performance in image-based tasks for a quick, effective model prototyping. However, to customize the architecture more specifically to the unique requirements of age estimation, subsequent iterations involved developing a custom CNN (CNN). This dual-phase approach—starting with pre-trained models and transitioning to a custom architecture—was designed to ensure a comprehensive evaluation of model performance while taking computational efficiency into account.

3.4.1 Pre-Trained Architecture Selection

Several pre-trained architectures were evaluated to identify the most suitable one for this research:

VGG16: Known for its effectiveness in image classification tasks.

The VGG16 architecture as shown by Fig 8, developed by Simonyan and Zisserman in 2014, is a seminal CNN (CNN) comprising 16 layers designed for intricate image pattern recognition. Organized into feature extraction and classification segments, the network employs 3x3 convolutional filters and Rectified Linear Unit (ReLU) activation functions in the former, and fully connected layers in the latter. A softmax function finalizes the network's output. Notably, the architecture increases the number of filters progressively, starting from 64 and doubling after each max-pooling stage. Despite its influential design and depth, VGG16 is resource-intensive, limiting its suitability for edge device deployment.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224×224 RGB image)					
conv3-64	conv3-64 LRN	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 conv1-256	conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-256
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 8 VGG16 Architecture (Simonyan and Zisserman, 2015)

ResNet50: The ResNet architecture, developed by Kaiming He et al. in 2015, has been a pivotal advancement in the domain of deep learning, especially in tasks involving image analysis. The architecture is renowned for its depth, with variants ranging from ResNet-18 to ResNet-152, where the number indicates the depth of the network in terms of layers. ResNet is fundamentally divided into two core segments: the feature extraction part, dominated by convolutional layers, and the classification portion, consisting of fully connected layers.

In the feature extraction segment, ResNet introduces a groundbreaking concept known as "residual learning," which enables the network to learn identity mappings via shortcut connections that skip one or more layers. This innovation helps to alleviate the vanishing gradient problem, thereby facilitating the training of much deeper networks. Typically, 3x3 convolutional filters are used, akin to VGG16, but with batch normalization preceding the activation function, usually a Rectified Linear Unit (ReLU).

Upon completion of the convolutional sequence, the network transitions to a set of fully connected layers aimed at classification. These layers culminate in a softmax activation function, converting the network's raw output into probability scores, effectively classifying the images into their respective categories.

One notable feature of the ResNet model is the meticulous arrangement of its layers into blocks, often referred to as "residual blocks," each containing a specific number of filters that may or may not vary between blocks depending on the variant. The number of filters generally starts from 64 and can increase, usually doubling at each stage.

Despite its impressive depth and the advantages conferred by residual learning, ResNet is surprisingly efficient in terms of memory and computational requirements, compared to its predecessors like VGG16. This makes it a viable choice for applications even on resource-limited devices. However, its complexity can still pose challenges for extremely constrained environments. Overall, the ResNet architecture has had a monumental impact on the field, setting new performance benchmarks and inspiring a myriad of subsequent architectures.

InceptionV3: This complex architecture is potent for complex image recognition tasks due to its multiple filters and varying sizes.

The InceptionV3 architecture as shown by Fig 9, an evolution of the original Inception model, was developed to improve performance and computational efficiency. Introduced by Google researchers, this architecture incorporates "Factorized Convolutions" and "Asymmetric Convolutions" to reduce computational load while maintaining high performance. The model consists of 48 layers and employs multiple "Inception modules," which use a variety of convolutional filter shapes to capture different features in parallel. It also integrates batch normalization and auxiliary classifiers to accelerate training and mitigate the vanishing gradient problem. Despite its depth and complexity, InceptionV3 is designed for computational efficiency, making it adaptable to a range of applications, including those on resource-constrained edge devices. Nonetheless, the architectural complexity may still be a limiting factor in extremely resource-limited settings. Overall, InceptionV3 has significantly influenced the landscape of deep learning architectures, serving as a benchmark for subsequent designs

type	patch size/stride or remarks	input size
conv	$3 \times 3 / 2$	$299 \times 299 \times 3$
conv	$3 \times 3 / 1$	$149 \times 149 \times 32$
conv padded	$3 \times 3 / 1$	$147 \times 147 \times 32$
pool	$3 \times 3 / 2$	$147 \times 147 \times 64$
conv	$3 \times 3 / 1$	$73 \times 73 \times 64$
conv	$3 \times 3 / 2$	$71 \times 71 \times 80$
conv	$3 \times 3 / 1$	$35 \times 35 \times 192$
$3 \times$ Inception	As in figure 5	$35 \times 35 \times 288$
$5 \times$ Inception	As in figure 6	$17 \times 17 \times 768$
$2 \times$ Inception	As in figure 7	$8 \times 8 \times 1280$
pool	8×8	$8 \times 8 \times 2048$
linear	logits	$1 \times 1 \times 2048$
softmax	classifier	$1 \times 1 \times 1000$

Figure 9 InceptionV3 Architecture (Szegedy et al., 2016)

MobileNetV2: Designed for mobile and edge devices, MobileNetV2 optimizes both performance and computational efficiency.

The MobileNetV2 architecture as shown by Fig 10, developed by Google researchers, is designed for efficiency and is particularly well-suited for mobile and edge computing devices. Building on its predecessor, MobileNetV1, this architecture incorporates "Inverted Residual Blocks" with "Linear Bottlenecks," enhancing both computational efficiency and model performance. The architecture consists of 53 layers, each optimized for low computational complexity. A key feature is the use of depthwise separable convolutions, which reduces the number of parameters and operations compared to standard convolutions, making the network faster and smaller. Despite its lean structure, MobileNetV2 achieves competitive performance on a variety of tasks, including image classification. Its design philosophy prioritizes a balance between computational resources and predictive performance, thereby making it highly adaptable for real-world applications on resource-constrained edge devices.

Input	Operator	t	c	n	s
$224^2 \times 3$	conv2d	-	32	1	2
$112^2 \times 32$	bottleneck	1	16	1	1
$112^2 \times 16$	bottleneck	6	24	2	2
$56^2 \times 24$	bottleneck	6	32	3	2
$28^2 \times 32$	bottleneck	6	64	4	2
$14^2 \times 64$	bottleneck	6	96	3	1
$14^2 \times 96$	bottleneck	6	160	3	2
$7^2 \times 160$	bottleneck	6	320	1	1
$7^2 \times 320$	conv2d 1x1	-	1280	1	1
$7^2 \times 1280$	avgpool 7x7	-	-	1	-
$1 \times 1 \times 1280$	conv2d 1x1	-	k	-	-

Figure 10 MobileNetV2 Architecture (Sandler et al., 2018)

EfficientNet-B0: Introduced by Tan et al., this architecture scales all dimensions of the network width, depth, and resolution. It offers a balance between accuracy and computational complexity. (Paper)

The EfficientNetB0 model, part of the EfficientNet family, is engineered for both efficiency and accuracy, leveraging a compound scaling method that uniformly scales all dimensions of the network. Introduced by researchers at Google, this architecture employs depthwise separable convolutions and employs a neural architecture search algorithm to find the most efficient network for a given task. With just 5.3 million parameters, EfficientNetB0 is designed to be compact, yet it achieves performance metrics comparable to much larger networks. Its unique combination of low computational cost and high performance makes it particularly well-suited for deployment in resource-constrained environments like edge devices

AlexNet: AlexNet, developed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, marked a significant milestone in the field of deep learning when it won the ImageNet competition in 2012.

The architecture consists of five convolutional layers, followed by three fully connected layers, culminating in a 1000-way softmax function for classification.

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x5	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	14×14	112	3
7	MBConv6, k5x5	14×14	192	4
8	MBConv6, k3x3	7×7	320	1
9	Conv1x1 & Pooling & FC	7×7	1280	1

Figure 11 EffecientNet-B0 Architecture (Tan and Le, 2019)

It employs ReLU activation functions, which helped mitigate the vanishing gradient problem, making it possible to train deeper networks. The use of dropout layers and data augmentation techniques further contributed to its robustness. While revolutionary at the time, the AlexNet model's computational and memory demands have since been surpassed by more efficient architectures, making it less ideal for resource-constrained environments

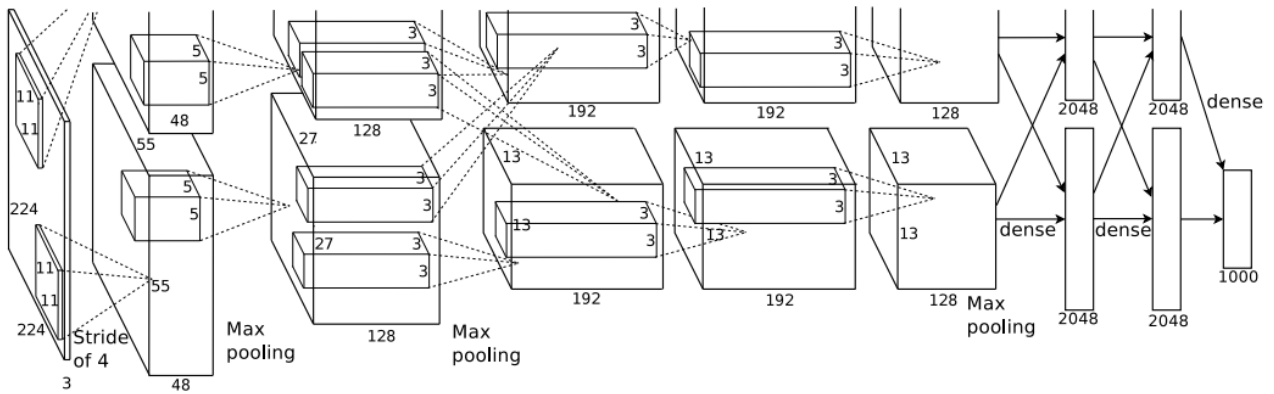


Figure 12 AlexNet Architecture (Alex Krizhevsky, Ilya Sutskever, 2007)

The overarching goal of this research is to devise an age estimation model that is not only precise but also computationally frugal, thereby making it amenable for deployment on edge devices. Against this backdrop, MobileNetV2 emerges as a compelling choice for several reasons:

Computational Efficiency:

One of the paramount factors is computational efficiency, a metric where MobileNetV2 excels. Designed explicitly for mobile and edge devices, MobileNetV2 is engineered to provide high performance with low computational overhead. This aligns seamlessly with the research objective of developing a lightweight yet effective age estimation model.

Parameter Efficiency:

MobileNetV2 boasts a reduced number of parameters compared to other heavyweight architectures like VGG16 and ResNet50. This not only accelerates the training process but also minimizes the storage requirements, which is crucial for edge devices with limited storage capabilities.

Inverted Residual Blocks:

The architecture employs inverted residual blocks with bottleneck features, which contribute to reducing the computational burden while retaining representational power. This enables the model to learn complex features efficiently, a vital attribute for age estimation tasks that necessitate capturing intricate facial features.

Versatility:

MobileNetV2 is designed to be a versatile architecture that is capable of serving multiple purposes, from object detection to image segmentation. Its versatility makes it a fitting candidate for age estimation, a task that involves nuanced features and varying conditions.

Prior Success:

MobileNetV2 has demonstrated remarkable success in a plethora of applications, thereby proving its robustness and adaptability. Its proven track record in real-world scenarios further bolsters the rationale for its selection.

Resource Constraints:

Considering the focus on edge device deployment, resource constraints such as memory, power, and computational capacity are of significant concern. MobileNetV2's design philosophy caters to these constraints, making it a pragmatic choice for this research.

In summary, MobileNetV2's computational and parameter efficiency, along with its versatile and robust nature, make it an ideal candidate for developing an age estimation model aimed at edge device deployment. Therefore, after a thorough review of various pre-trained architectures, MobileNetV2 was judiciously selected as the base architecture for this study.

3.4.2 Proposed CNN Architecture with Pretrained MobileNetV2

3.4.2.1 First Variant with 512 Neurons Single Layer

The initial stage involved loading the MobileNetV2 architecture without its top layers, essentially excluding the classifier. This variant of MobileNetV2 was pre-trained on the ImageNet dataset. The exclusion of the top layers was deliberate, paving the way for the subsequent addition of custom layers tailored to the age estimation task. Furthermore, the pre-trained layers were frozen to retain their already learned features, thus reducing the computational burden during the fine-tuning process.

Initial Configuration and Model Adaptation

Post feature extraction, a Global Average Pooling (GAP) layer was added. This layer serves to reduce the spatial dimensions of the input volume, effectively summarizing the essential features in each depth slice. Following the GAP layer, a fully connected Dense layer with 512 neurons was integrated, employing the Rectified Linear Unit (ReLU) activation function. This layer is designed to learn high-level features that are instrumental in age estimation.

To combat overfitting—a common issue in machine learning tasks—a Dropout layer with a rate of 0.5 was introduced immediately after the Dense layer. Finally, a Dense layer serving as the classifier was added. This layer comprises six neurons corresponding to the six age groups in the dataset and employs a SoftMax activation function to output probabilities.

Model Compilation and Parameters

The model was compiled using the Adam optimizer with a learning rate of 0.0001. The loss function employed was categorical cross-entropy, which is suitable for multi-class classification problems. Two metrics, specifically the model's loss and accuracy, were monitored during training.

The finalized architecture comprises approximately 2.91 million parameters, occupying an estimated 11.13 MB of memory. Notably, only around 2.51 MB (or 658,950 parameters) are trainable, while the remainder are non-trainable parameters inherited from the pre-trained MobileNetV2 model.

Training Strategy and Callbacks

Training was conducted using a 4-fold cross-validation approach. In each fold, a different subset of the data was used for validation, while the remaining data was used for training. Two callbacks were employed during training:

Early Stopping: This halts the training process if the validation loss ceases to improve for ten consecutive epochs, preventing overfitting.

Model Checkpoint: This saves the model achieving the highest validation accuracy, ensuring that the best-performing model is retained for evaluation.

The batch size for training was set at 128, and the model was trained for an initial 5 epochs, with the provision for further epochs depending on the early stopping criteria.

Model Evaluation

Following the completion of the training process across all folds, the model was subjected to a final evaluation on the test set. Both the loss and accuracy metrics were recorded to assess the model's performance.

This multi-faceted approach to customizing and training MobileNetV2 aims to construct a reliable, efficient, and robust age estimation model suitable for deployment on edge devices.

3.4.2.2 Second Variant with Multiple Layers

Similar to the first variant, the base model for this architecture is MobileNetV2, pre-trained on the ImageNet dataset and devoid of its top layers. The pre-trained layers were also frozen, preventing their weights from being updated during the training process.

Augmented Convolutional Architecture

In contrast to the first model, this variant adds an array of additional layers to enhance feature extraction capabilities. The added layers include four Convolutional 2D layers with kernel sizes of 3×3 , each followed by a Batch Normalization layer. The activation function for these convolutional layers is ReLU (Rectified Linear Unit).

Between the added convolutional layers, two Max Pooling layers with pool sizes of 2×2 are integrated to perform dimensionality reduction. This operation serves to emphasize the more salient features and reduce computational requirements.

Post-Convolutional Layers

The convolutional layers are followed by a Flatten layer, which reshapes the output into a one-dimensional array. Subsequently, a Dense layer with 128 neurons is added, employing the ReLU activation function for non-linearity. This layer is followed by a Batch Normalization and a Dropout layer with a rate of 0.5 to mitigate overfitting.

Output Layer and Compilation

The final layer is a Dense output layer with six neurons, which correspond to the six age groups under consideration. A SoftMax activation function is employed to convert the raw output into probabilities.

The model is compiled using the Adam optimizer with a learning rate of 0.0001 and employs categorical cross-entropy as the loss function.

Model Characteristics

The modified architecture has a total of approximately 2.51 million parameters, taking up an estimated 9.59 MB of memory space. Notably, only around 992.59 KB (or 254,102 parameters) are trainable, keeping the model lightweight and suitable for edge devices.

Training Procedure and Callbacks

The training protocol remains consistent with the first variant, employing a 4-fold cross-validation strategy and the same set of callbacks for early stopping and model checkpointing.

Model Evaluation

After the training process is completed across the four folds, the model undergoes a final evaluation on a separate test set to assess its loss and accuracy metrics.

Summary

The second variant of the MobileNetV2 adaptation was designed to explore the performance impact of augmenting the convolutional architecture. With its enhanced feature extraction capabilities and lightweight design, it serves as a competitive alternative to the first model, offering a balanced trade-off between computational efficiency and predictive performance.

This concludes the detailed discussion of the second variant of the MobileNetV2 adaptation for age estimation. Both models will be compared in the subsequent sections to evaluate their relative performances and suitability for deployment on edge devices.

3.5 Validation Strategy

Initially, the research was conducted utilizing the 5-fold cross-validation scheme inherent to the Adience dataset. The dataset was partitioned into five distinct folds, each containing a varying number of images across age groups. This approach was initially adopted to abide by the pre-defined structure of the dataset.

Despite the structured partitioning, the model's performance was considerably suboptimal. A deeper examination revealed a pronounced class imbalance within the folds, an issue that was extensively elaborated upon in earlier sections of this thesis. Data augmentation and sampling techniques, implemented using the Augmentor library, were employed to mitigate this issue but yielded minimal improvement in model performance.

Given the persistent issue of class imbalance and unsatisfactory model performance, a transition was made to a stratified k-fold cross-validation scheme. This approach ensures that each fold is made by preserving the percentage of samples for each class, thereby addressing the class imbalance issue effectively.

To implement the stratified k-fold cross-validation, the Scikit-Learn library's `train_test_split` function was used. The dataset was first split into training, validation, and test sets in a stratified manner, ensuring that each set had a similar distribution of age groups. The distribution was subsequently confirmed using Python's `Counter` class from the `collections` library.

The stratified fold validation strategy serves several significant purposes:

Balanced Representation: The distribution of age groups in the training, validation, and test sets is more balanced, thus mitigating the class imbalance issue.

Enhanced Generalizability: The strategy improves the model's ability to generalize well to new, unseen data by ensuring that each class is adequately represented in the training data.

Upon implementing the stratified k-fold cross-validation, the distribution of age groups in the training set, validation set, and test set was found to be considerably balanced.

Training Distribution: The training set exhibited a balanced distribution across age groups, with counts ranging from 698 for Seniors to 4,272 for Young Adults.

Validation Distribution: The validation set mirrored this balanced distribution, with counts ranging from 87 for Seniors to 534 for Young Adults.

Test Distribution: Similarly, the test set maintained this balance, with counts ranging from 88 for Seniors to 534 for Young Adults.

Consistency in Future Work: It is recommended to consistently use stratified fold cross-validation in future works involving this dataset to ensure fair evaluation and comparison.

In summary, the transition to a stratified fold cross-validation significantly improved the class distribution in each set, thereby aiding in effective model training and evaluation. This strategic change is essential for the robustness and generalizability of the age estimation models developed in this research.

3.6 Model Evaluation

In the final reckoning of any machine learning model, its performance metrics serve as the ultimate arbiter. Given the pivotal role of age estimation across an array of real-world applications, particularly those reliant on edge devices, the evaluation process is multi-

dimensional. The subsequent subsections expound upon the evaluation metrics and graphical representations employed to rigorously assess the model's efficacy and efficiency.

Accuracy: The cornerstone of this evaluation is accuracy, a holistic measure of the model's performance across the various age groups. However, the complexity of age estimation necessitates a multifaceted evaluation. To this end, additional metrics and visualizations are integrated into the evaluation methodology

Confusion Matrix: This matrix offers a granular perspective on the model's performance. It facilitates an in-depth analysis, highlighting specific areas where the model excels or requires improvement.

Loss and Accuracy Across Epochs: The trajectory of training and validation loss and accuracy over epochs serves as an invaluable diagnostic tool. It aids in discerning the model's learning behavior and identifying potential overfitting or underfitting scenarios.

Receiver Operating Characteristic (ROC) Curve: The ROC curve is instrumental in assessing the model's discriminative power between different age groups. It plots the true positive rate against the false positive rate, providing insights into the model's classification capabilities.

Summary

The model undergoes a comprehensive evaluation involving multiple metrics and visual techniques. This rigorous assessment affirms its predictive accuracy and lends credence to its suitability for deployment on edge devices. Such an exhaustive evaluation aligns seamlessly with the overarching objectives of this research, thereby validating the model's practical utility.

3.7 Summary

This chapter delineated the comprehensive methodology underpinning the research, encompassing every facet from data preparation to model evaluation. The meticulous architecture of an end-to-end framework served as the research's structural foundation, guiding each phase with a robust yet flexible strategy tailored to meet the unique demands of age estimation on edge devices.

The data generation pipeline was crafted to be in synergy with the research's primary objective of lightweight and efficient modelling. The methodology adopted a 5-fold cross-validation strategy, ensuring a rigorous and robust evaluation process. Multiple pre-trained architectures were assessed for their potential contributions to age estimation, culminating in the judicious selection of MobileNetV2 as the base architecture for the study.

Two distinct variants of the MobileNetV2 architecture were examined, each designed to probe the performance trade-offs inherent in optimizing for both computational efficiency and predictive accuracy. The first variant employed a single Dense layer with 512 neurons, while the second introduced additional convolutional layers to augment feature extraction capabilities. Both models were meticulously evaluated using multiple metrics and visualizations, including accuracy, confusion matrix, loss and accuracy over epochs, and the Receiver Operating Characteristic (ROC) curve.

The evaluation process was multi-dimensional, providing a holistic view of the model's performance across various fronts. This rigorous assessment not only affirmed the model's predictive prowess but also validated its suitability for deployment on resource-constrained edge devices.

In summation, the chapter offered a coherent and exhaustive exposition of the research methodology and model evaluation procedures, substantiating the model's efficacy and its alignment with the broader research objectives. This meticulous approach ensures that the resultant age estimation model is both accurate and computationally efficient, catering to the specific requirements of edge device deployment.

CHAPTER 4

ANALYSIS

4.1 Introduction

Exploratory Data Analysis (EDA) serves as an indispensable pillar in the life cycle of data-driven research, particularly in the domain of machine learning and age estimation models. The complexity of age estimation tasks, compounded by the intricacies of deploying models on edge devices, accentuates the necessity for a thorough examination of the dataset. EDA provides a foundational understanding of the dataset's properties, structure, and quirks, thereby guiding subsequent phases of model development and evaluation.

In the context of this research, EDA aims to achieve several key objectives:

Data Quality Assessment: Identifying missing values, outliers, and anomalies that could potentially skew model performance.

Class Distribution: Understanding the distribution of age groups and genders enables the identification of any class imbalances that may necessitate resampling techniques.

Feature Analysis: Investigating attributes like tilt and yaw angles, face sizes, and image orientation can offer valuable insights into feature engineering steps that could be beneficial for the model's performance.

Preliminary Insights: Generating early hypotheses about the relationships between different variables, which can be validated or refuted during model training and evaluation.

Resource Planning: Given the focus on edge device deployment, understanding the dataset's composition helps in estimating computational requirements, thereby aiding in resource allocation and planning.

This section unfolds a meticulous EDA that scrutinizes various facets of the dataset, ranging from general attributes like the number of RGB images to more complex features such as the

distribution of tilt and yaw angles. The insights garnered from this analysis will serve to inform the subsequent stages of model development, evaluation, and potentially, deployment on edge devices.

4.2 General Data Characteristics

In this section, the general characteristics of the dataset, such as the types and number of images, are elucidated. This preliminary analysis aims to offer foundational insights that will inform subsequent phases of the research, including data preprocessing, model architecture selection, and training strategy.

4.2.1 Number of RGB Images

An initial inquiry into the dataset was carried out to ascertain the types of images it comprises. To this end, each image was programmatically inspected to determine whether it is an RGB (Red, Green, Blue) or grayscale image.

Remarkably, the dataset is found to consist exclusively of RGB images, tallying up to a total of 19,370. Intriguingly, the dataset is devoid of any grayscale images.

The Python Imaging Library (PIL) was utilized to open and inspect each image in the dataset. A traversal of the dataset directory was performed, and for each discovered image file, its mode was checked. If the mode equalled 'RGB', the image was counted as an RGB image. Conversely, if the mode was either 'L' or '1'—indicating an 8-bit or 1-bit pixel grayscale image, respectively—it would be counted as a grayscale image.

Uniformity in Image Type: The absence of grayscale images and the uniform presence of RGB images negate the need for additional preprocessing steps to handle multiple image types. This simplifies the data pipeline, reducing the chances of introducing errors or inconsistencies.

Facilitation of Model Training: Most deep learning models, particularly pre-trained architectures, expect 3-channel input. The dataset's consistency with this expectation streamlines the model training process, eliminating the need for conditional checks or transformations during training and inference.

Given the uniformity of RGB images in the dataset, it is advisable to proceed with model training without incorporating additional steps for grayscale image handling. This will not only simplify the data pipeline but also likely result in computational efficiency, aligning well with the overarching objective of this research, which is to develop a lightweight, yet effective, age estimation model suitable for deployment on edge devices.

4.2.2 Distribution of Images Across Folds

The dataset under study is segmented into five distinct folds, each containing a varying number of images. A Python script iterated through the dataset's dictionary structure to enumerate the images in each fold. Additionally, the script calculated the total number of images across all folds.

Overall Dataset Size: The Adience dataset, comprising all five folds, amounts to 19,370 images. This extensive dataset furnishes a robust foundation for the machine learning tasks of training, validation, and testing.

Variability Across Folds: The dataset manifests a degree of variability in the distribution of images among the folds. Fold_0 is the most populous, containing 4,484 images, while Fold_3 is the least populous, with 3,446 images. Such variability accentuates the necessity of employing cross-validation techniques to ensure generalizable model performance.

Cross-Validation Considerations: Given the disparate sizes of the folds, strategies like 5-fold 'leave-one-out' cross-validation can be instrumental in providing a diverse range of training and validation samples.

Consistent Representation: Despite the uneven distribution, each fold contains a substantial number of images, ensuring no fold is significantly underrepresented. This is pivotal for achieving a balanced training and validation process.

Data Augmentation Potential: The dataset's considerable size opens the door for data augmentation techniques to be applied, thereby enriching the effective dataset size. This can be especially advantageous for deep learning models, which typically benefit from larger datasets.

Data Partitioning: Understanding the number of images in each fold aids in judiciously partitioning the data for various stages of the machine learning pipeline—training, validation, and testing.

Model Validation: The insights gleaned from the distribution of images can inform the design of k-fold cross-validation strategies, a common practice in machine learning for assessing model performance.

Strategic Augmentation: In case a fold has fewer images, targeted data augmentation could be employed to balance the fold's representation, thus potentially enhancing model robustness.

Utilize All Folds: Given the dataset's considerable size and variability, it is advised to utilize all the folds during model training and evaluation to ensure the model is adept at handling different data distributions.

Weighted Performance Metrics: When aggregating performance metrics such as average accuracy across folds, ensure that each fold's contribution is appropriately weighted to reflect its size.

Data Augmentation: Considering the dataset's substantial volume, data augmentation techniques can be judiciously applied to expand the effective dataset size. However, it is crucial that such augmentations do not distort the original data's characteristics, particularly given that the dataset already encapsulates real-world variables such as diverse appearances and lighting conditions.

Understanding the dataset's structure and the distribution of images across folds is foundational for subsequent analytical and modelling undertakings, ensuring that each phase of the research is guided by a nuanced understanding of the data at hand.

4.2.3 Missing Values Analysis

Missing values in a dataset pose significant challenges, particularly when developing machine learning models for tasks such as age estimation. Incomplete data can lead to biased or incorrect predictions, thereby compromising the model's reliability and robustness. This section aims to

provide a meticulous analysis of missing values within the dataset, their characteristics, and implications for subsequent stages of the research.

To quantify the prevalence of missing values across different attributes, we employed Python's Pandas library. Specifically, we calculated the relative frequency of missing values for each feature. This was achieved by dividing the number of null entries in each column by the total number of rows in the dataset, subsequently multiplying by 100 to obtain percentages.

Identification of columns with missing values.

Calculation of the relative frequency of missing values in each identified column.

Isolation of rows with missing values for in-depth attribute-specific analysis.

The majority of the features, including `user_id`, `original_image`, `face_id`, `age`, and various pose-related attributes (`x`, `y`, `dx`, `dy`, `tilt_ang`, `fiducial_yaw_angle`, `fiducial_score`), exhibited a complete set of values. The absence of missing values in these features negates the need for imputation, thereby simplifying the data preprocessing pipeline.

While the gender feature presented a noticeable degree of missing values—approximately 4.02%—it is pertinent to note that this attribute has been ultimately excluded from the final metadata for model training. This decision is rooted in the research's explicit focus on age estimation, rendering the gender feature non-essential for the current scope of work.

Within the current research context, the gender feature's incompleteness does not directly impact the model's age estimation performance. Thus, no imputation or cleaning steps have been employed for this attribute.

Upon isolating rows with missing gender values, it was observed that a significant majority of the entries had their age labelled as 'None,' amounting to 708 instances. This suggests a systematic issue, possibly related to the data collection or annotation process.

Certain age brackets, such as (25,32) also showed missing gender information, albeit at a much smaller scale (51 instances). A large number of these images were in the "Profile" orientation, which may hint at why gender annotation could be missing, perhaps due to limited facial feature visibility in such orientations.

Missing values, if not handled appropriately, can compromise the model's predictive accuracy and generalizability.

The dataset exhibited instances where the age attribute was unspecified, labelled as 'None.' While these instances were relatively low in number, their presence could introduce noise and uncertainties into the age estimation model. Consequently, these instances have been excluded from the final metadata, resulting in a refined dataset comprising 18,622 images.

The exclusion of instances with unspecified age attributes enhances the dataset's integrity, as these entries could potentially distort the training process and affect model performance.

Given the research's singular focus on age estimation, the presence of instances with unspecified age attributes is incongruent with the model's primary objective.

While imputation techniques could have been employed to fill the missing age values, this approach was deemed less suitable for several reasons. First, imputing values for the 'None' category could introduce bias or noise into the model. Second, given the model's focus on age estimation, imputation based on other features may not yield precise or reliable age values.

The exclusion of these instances resulted in a modest reduction in the dataset size but significantly increased its reliability and robustness for the specific task of age estimation.

The strategic exclusion of instances with unspecified age attributes aligns closely with the project's goals and constraints. This decision not only maintains the dataset's integrity but also ensures that the model is trained on data that directly contributes to its primary objective—accurate age estimation. Therefore, the resultant dataset of 18,622 images is both ample and focused, setting a solid foundation for the subsequent stages of model training and evaluation.

4.2.4 Outlier Analysis

Outliers present a complex scenario in machine learning applications, particularly in age estimation models for edge devices. They can either represent rare, real-world cases that are valuable for making the model robust or be anomalies that degrade model performance. The focus of this section is on outliers in face areas, a feature derived from the dimensions 'dx' and 'dy'.

The identification of outliers was executed using the following steps:

Face Area Computation: For each image, the face area was derived by multiplying two dimensions related to the face's bounding box: the horizontal extent ('dx') and the vertical extent ('dy').

Percentile Calculation: The face areas across the entire dataset were sorted, and two key values were identified—the value below which 25% of the face areas fall (known as the 25th percentile or Q1) and the value below which 75% of the face areas fall (known as the 75th percentile or Q3).

Interquartile Range (IQR): The difference between the 75th percentile (Q3) and the 25th percentile (Q1) was calculated. This measure, known as the Interquartile Range, gauges the statistical dispersion or spread of the face areas.

Outlier Boundaries: A range of acceptable values was defined based on the Interquartile Range. Any face area falling outside this range was labelled as an outlier. Specifically, the lower boundary was set as the 25th percentile minus 1.5 times the Interquartile Range, and the upper boundary was the 75th percentile plus 1.5 times the Interquartile Range.

Outlier Identification: Finally, face areas that fell outside the acceptable range, as defined above, were categorized as outliers.

This methodology provides a robust mechanism for identifying outliers, which are face areas that significantly deviate from the dataset's central tendency.

A total of 1990 instances were identified as outliers in face areas, which deviate significantly from the dataset's central tendency. These outliers could arise due to varying factors like zooming inconsistencies, different resolutions, or genuine variations in face sizes among the dataset population.

Contrary to typical data preprocessing steps that advocate for the removal or transformation of outliers, a deliberate decision was made to retain these outliers in the dataset. This choice is grounded in the recognition that outliers often mimic the variations and inconsistencies present in real-world images. As the primary goal of this research is to develop a model that is robust

and generalizable across diverse real-world scenarios, the outliers were considered to be integral to this objective.

Model Robustness: Retaining these outliers could potentially make the model more robust to variations in face sizes, although it may require the application of additional data augmentation techniques like zooming and scaling.

Conditional Modelling: If these outliers represent specific real-world conditions, conditional modelling techniques may be employed in the future to build specialized models for such cases.

Statistical Measures: Subsequent analyses might benefit from using robust statistical measures less sensitive to outliers, ensuring a more representative understanding of the data.

The identification and subsequent decision to retain outliers in face areas serve as pivotal steps in the data preprocessing phase. This strategy aligns closely with the research's objective to develop a generalized age estimation model capable of handling a broad array of real-world scenarios. Therefore, this outlier analysis adds valuable insights that will inform both the data preprocessing and model training phases, ultimately contributing to the overall robustness and effectiveness of the ensuing age estimation model.

4.2.5 Descriptive Statistics Analysis

The Figure Fig 13 proffers an elaborate descriptive statistical analysis of various facial parameters derived from the dataset. Each parameter has been meticulously dissected to furnish insights integral to age estimation research:

Face ID (face_id):

Range: Spanning from 1 to 2,284, it indicates the identity markers for subjects in the dataset.

Central Tendency: With a mean value around 1,023.86, the average face_id is situated approximately at the mid-point of the range.

Variability: A standard deviation of 645.39 suggests a fairly wide dispersion of data around the mean.

X & Y Coordinates (x, y):

Location: Represents the coordinates of the top-left corner of the bounding box encapsulating the face.

Central Tendencies: The mean values, 949.67 (x) and 675.66 (y), convey the average position of faces in the dataset.

Variabilities: Standard deviations of 667.06 and 429.92 for x and y respectively highlight the spread in face positions.

Bounding Box Dimensions (dx, dy):

Description: These dimensions delineate the width (dx) and height (dy) of the face bounding box.

Central Tendencies: On average, the face width stands at 633.04, while the height is slightly lesser at 629.69.

Variability: Standard deviations (440.41 for dx and 426.30 for dy) illuminate the variation in face sizes across the dataset.

	face_id	x	y	dx	dy	tilt_ang	fiducial_yaw_angle	fiducial_score
count	19370.000000	19370.000000	19370.000000	19370.000000	19370.000000	19370.000000	19370.000000	19370.000000
mean	1023.858906	949.670212	675.659370	633.042437	629.692463	-12.405111	3.128549	72.266598
std	645.385251	667.060150	429.918002	440.409036	426.296233	73.430214	15.658667	38.799247
min	1.000000	0.000000	0.000000	80.000000	80.000000	-205.000000	-45.000000	5.000000
25%	486.000000	441.250000	333.000000	324.000000	325.000000	-80.000000	0.000000	42.000000
50%	980.000000	852.000000	653.000000	517.000000	517.000000	-5.000000	0.000000	70.000000
75%	1526.000000	1359.000000	964.000000	778.000000	778.000000	5.000000	0.000000	98.000000
max	2284.000000	10024.000000	2752.000000	3264.000000	3225.000000	365.000000	45.000000	221.000000

Figure 13 Descriptive Summary

Tilt Angle (tilt_ang):

Insight: This angle captures the vertical inclination or declination of the face.

Central Tendency & Variability: A mean value of -12.41° accompanied by a standard deviation of 73.43° suggests that most faces have a slight downward tilt, with some notable deviations.

Fiducial Yaw Angle (fiducial_yaw_angle):

Description: Represents the horizontal rotation of the face, from a full left profile to a full right.

Central Tendency & Variability: The near-zero mean (3.13°) indicates a predominant frontal orientation, but the sizable standard deviation of 15.66° emphasizes a considerable spread around this average.

Fiducial Score (fiducial_score):

Description: A metric gauging the quality or confidence of fiducial point detection.

Central Tendency & Variability: An average score of 72.27, with a standard deviation of 38.80, suggests that while most detections are of decent quality, there exists a varied range in scores.

Outliers: The maximum y value (10,024) appears anomalously high. Likewise, a maximum tilt angle of 365° seems unusual, given the full circular range of 360°. These could be potential outliers or data entry errors warranting further investigation.

Distribution Skewness: The yaw angle's mean and median (50th percentile) both being zero, yet possessing a non-zero standard deviation, hints at a potential skewness in the distribution.

4.3 Distribution Analysis

4.3.1 Distribution of Age Groups

The distribution of age groups within a dataset can substantially influence the efficacy and generalizability of machine learning models developed using the data. The provided histogram Fig 14 delineates the distribution of age groups across the different folds of the Adience dataset.

Dominant Age Demographic: A pronounced representation emerges for the age bracket of (25, 32), especially within fold 0. This suggests a pronounced collection of images encompassing Young Adults.

Sparse Age Bins: Particular age categories, including (8, 23), (27, 32), and (38, 48), display meager representation in certain folds, signifying an infrequent presence of these age groups within the dataset's entirety.

Fold Discrepancies: Variances in the age group distribution manifest across different folds. A case in point: age bracket (0, 2) exhibits an amplified presence in fold 0 yet is conspicuously underrepresented in fold 1.

Erratic Representations: Age groups such as (8, 23) and (38, 42) display intermittent representation in select folds, potentially attributable to the methodologies employed in data segregation or inherent characteristics of the sourced data.

From a computational perspective, this imbalanced representation across age groups could introduce complexities during the modeling phase. The resultant model might exhibit proficiency in predicting overrepresented age groups, whilst potentially faltering with sparsely represented ones.

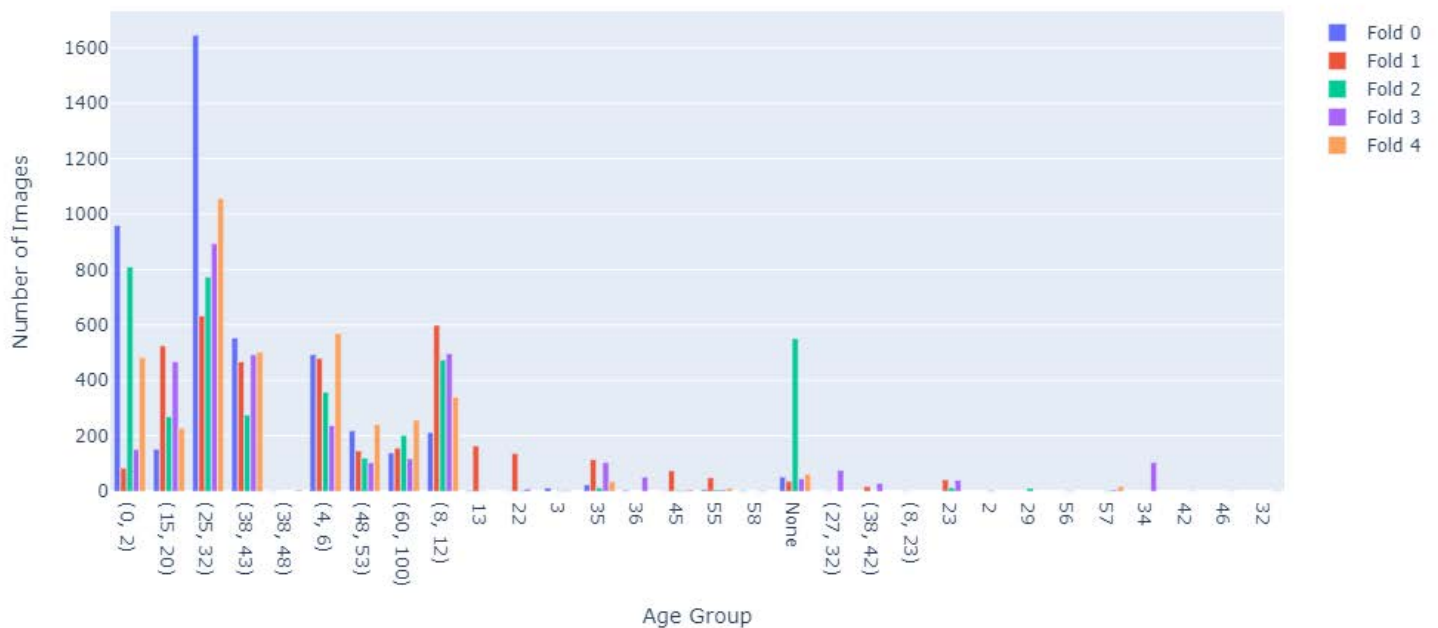


Figure 14 Combined Distribution of Age Groups Across All Folds

To illuminate the age group distribution spread across all folds, a bar chart, segmented by fold, was meticulously crafted employing the Plotly library. The quantitative representation of each age group within a given fold is symbolized by the respective bar's magnitude.

Grasping the age group dispersion within the dataset is paramount due to:

Model Calibration: A dataset endowed with balanced classes generally fosters more holistic training for machine learning algorithms. An overwhelming representation of a singular age group could predispose the model towards that specific demographic.

Data Integrity: An analytical overview of the distribution furnishes clarity on the dataset's inherent quality and potential predispositions. A scanty representation of certain age groups might underscore lacunae in the data acquisition or sampling phases.

Consistent Age Peaks: The age bracket (0, 2) ubiquitously registers the maximal count across all dataset folds. This inference underscores the dataset's potency, especially for algorithms focused on young facial recognition.

Median Age Distribution: Age categories spanning (25, 32), (38, 43), and (48, 53) register a balanced presence, offering a substantial training corpus for these age ranges.

Fold Disparities: While a consistent pattern manifests across the dataset, distinct age groups exhibit varying representation across different folds. To illustrate, the age bracket “None” in Fold 2 displays an enhanced count vis-à-vis its counterparts in alternate folds.

Diminished Categories: Singular year age categories, particularly post the 60-year mark, demonstrate a marked paucity in representation, indicative of either an intrinsic dataset distribution or potential hurdles in accurate data labeling.

For thinly represented age brackets, strategies such as data augmentation could amplify their data corpus. Alternatively, amalgamating these brackets with proximate age groups might be contemplated for more generic age classifications.

In summation, the visualization offers an exhaustive portrayal of age distribution across dataset folds, furnishing pivotal insights that can shepherd ensuing data manipulation, algorithmic training, and evaluative phases.

4.3.2 Distribution of Gender

The bar chart provided Fig 15 delineates the gender distribution across distinct folds of a dataset. Based on the presented visualization, several insights and observations can be dr

Prevalent Genders: Both male (m) and female (f) categories manifest a robust presence across all folds. Specifically, each category, in most instances, boasts a count surpassing 1500 images. This prevalent representation underscores the dataset's rich diversity in capturing the two predominant gender categories, thereby furnishing a favorable milieu for model training.

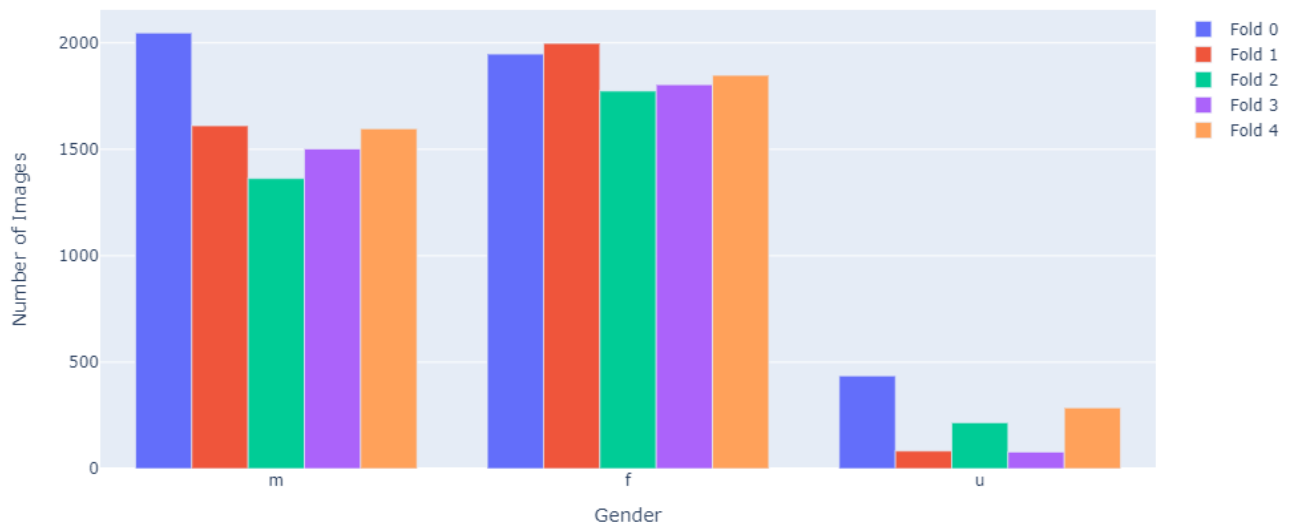


Figure 15 Combined Dstribution of Gender Across All Folds

Steadiness Across Folds: A noteworthy consistency in the distribution of male and female images across various folds is evident. Such homogeneity is paramount when employing cross-validation methodologies, as it assures that every fold embodies a holistic reflection of the overall dataset distribution. By maintaining this equilibrium, the potential for idiosyncratic biases stemming from any particular fold is substantially mitigated.

Undefined Gender Category: The chart also elucidates the existence of a 'u' category, which can be conjectured to signify "undefined" or "unspecified" gender. Notably, this category's representation pales in comparison to the male and female categories. While this category's presence alludes to certain images lacking explicit gender annotations, its relative infrequency confirms that the preponderance of the dataset is endowed with discernible gender markers. Depending on the eventual objectives of the model, multiple strategies could be contemplated for this category, including amalgamation with a predominant category, isolation as a distinct class, or potential exclusion from specific analyses.

Equilibrium Considerations: Despite the substantial representation of both male and female categories, discernible disparities in their frequencies are palpable, especially within certain folds. Such disparities might inadvertently sway the performance of machine learning models, particularly those anchored in deep learning paradigms. To counteract this, strategies such as oversampling, undersampling, or the deployment of balanced batch generators might be considered.

Ambiguity Concerns: The inception of the 'u' category inevitably evokes queries. Are these images intrinsically ambiguous, or were they rendered challenging to categorize owing to external variables like image fidelity, pose, or occlusions? A granular examination of this category's constituents might unveil deeper insights regarding its essence.

In summation, this visual representation underscores the pivotal role of comprehending gender distribution within the dataset. Such comprehension not only dictates the strategic trajectory of model training but also profoundly influences the interpretation of its resultant metrics and its applicability within real-world scenarios..

4.4 Facial Feature Analysis

4.4.1 Distribution of Face Sizes

The histogram Fig 16 depicts the distribution of face sizes within the given dataset. Here are the detailed observations:

Predominant Smaller Face Areas: A vast majority of images exhibit smaller face areas, as evident by the high frequency at the leftmost part of the histogram. This might imply that many of these images are either zoomed out, taken from a distance, or represent portions of larger images where the face occupies a smaller fraction.

Diminishing Larger Face Areas: As the face area increases, the number of images diminishes progressively. Images with substantially large face areas are scarce in this dataset, suggesting there are fewer close-up shots.

Potential Outliers: The histogram displays a long tail, hinting at the presence of outliers or instances of extremely large face sizes. Such anomalies should be examined further to determine if they are genuine instances or result from data inconsistencies.

Face sizes were determined by calculating the area occupied by the face in each image. The histogram was constructed to showcase the frequency distribution of these face sizes, delineating the common and rare face sizes within the dataset.

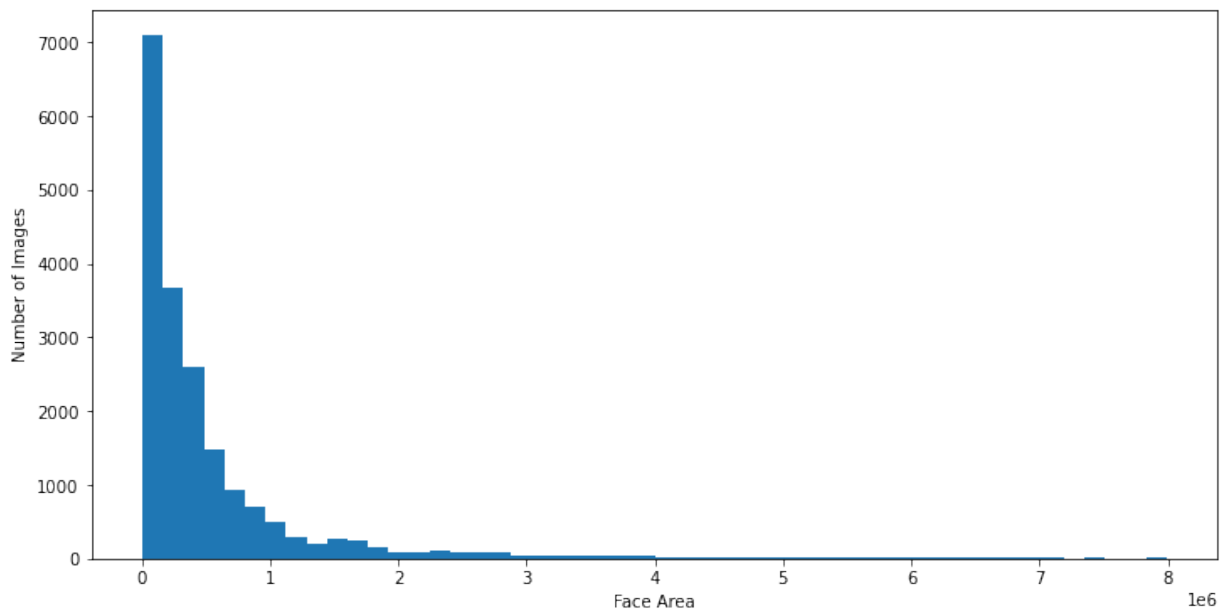


Figure 16 Combined Dstribution of Face Area Across All Folds

The distribution of face sizes furnishes several critical insights for the subsequent phases of the research:

Model Generalization: If a model is predominantly trained on smaller face sizes, it might not generalize well to close-up shots or images where the face occupies a significant portion of the frame. This points towards the need for data augmentation or specialized training regimes to cater to varied face sizes.

Data Augmentation Insights: Recognizing the abundance of smaller face areas suggests potential data augmentation techniques. Zooming in on faces or cropping larger areas during training could diversify the dataset and ensure models are more versatile in their recognition capabilities.

Feature Relevance: The size of the face might correlate with other crucial variables, like age or gender. For instance, close-up shots might predominantly belong to certain age groups or depict specific emotions. Recognizing these correlations can be invaluable during feature engineering.

4.4.2 Distribution of Tilt and Yaw Angles

The histogram Fig 17 delineates the distribution of tilt and Yaw angles within the dataset. Herein are the primary takeaways:

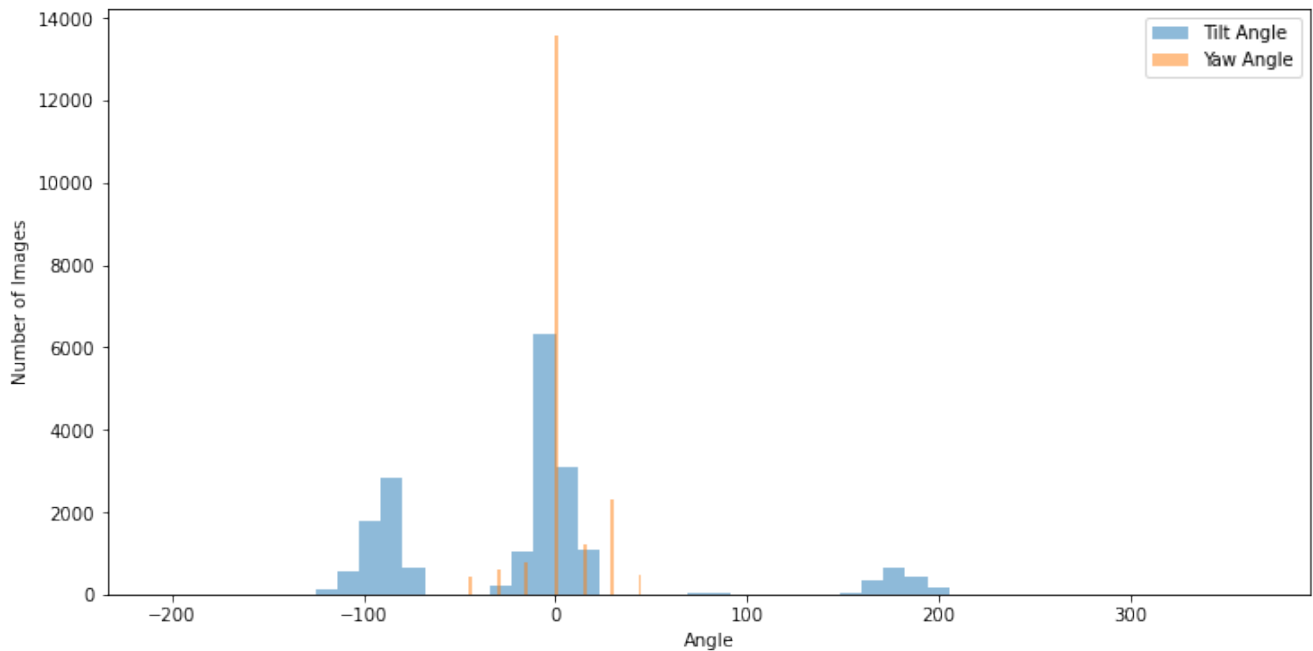


Figure 17 Combined Distribution of Tilt and Yaw Angles Across All Folds

4.4.2.1 Tilt Angle

Central Concentration: The pronounced spike around the zero mark denotes a predominance of images with faces that have a neutral tilt. This suggests that the majority of the subjects are looking straight ahead, aligning their face vertically.

Moderate Downward Tilt: To the left of the central peak, there is a noticeable distribution of negative tilt angles. This signifies a subset of images wherein subjects have their faces tilted slightly downward.

Minimal Upward Tilt: In contrast, the distribution of positive tilt angles (indicating an upward tilt of the face) is relatively sparse. This could imply that instances where subjects are looking upwards are rarer in the dataset.

Extreme Tilts are Scarce: There are very few images that exhibit extreme tilt angles, either significantly upwards or downwards. This informs us that most of the images in the dataset capture faces with moderate to neutral tilts.

The tilt angle, abstracted from the dataset, captures the vertical orientation of the face, pivoting upwards or downwards from a frontal position. The histogram elucidates the frequency distribution of these tilt angles, highlighting the primary orientations prevalent in the dataset.

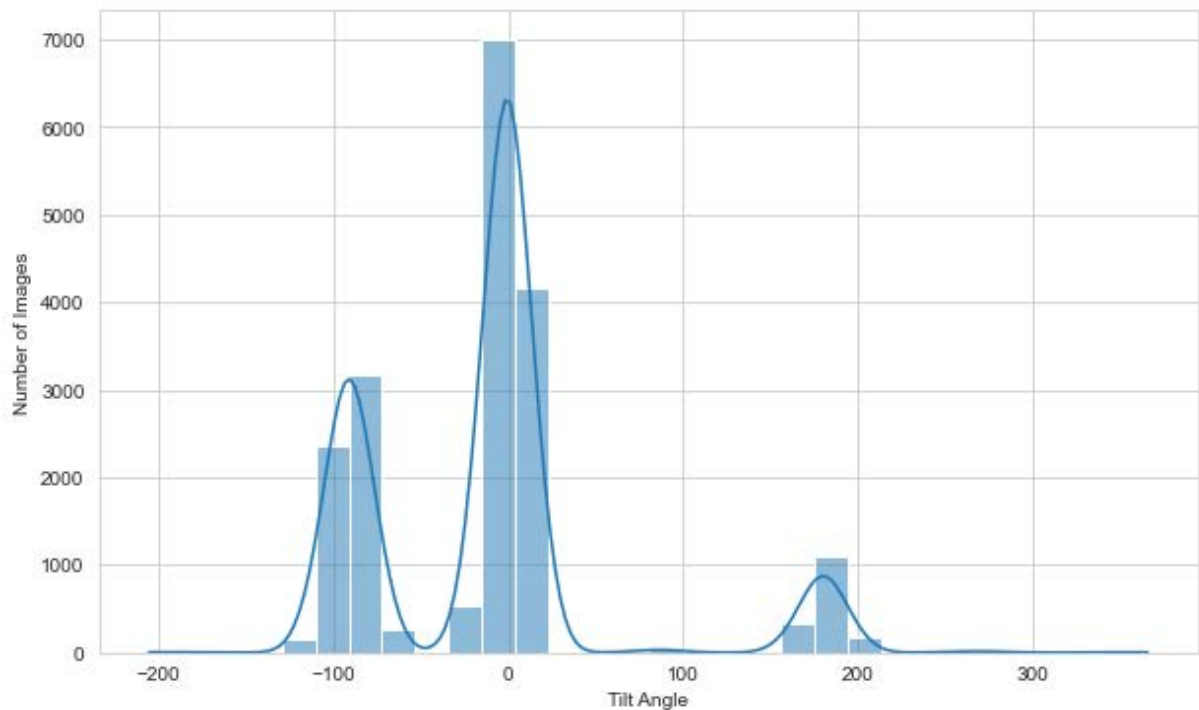


Figure 18 Dstribution of Tilt Angle

Modelling Implications: A model trained on this dataset will likely be proficient at identifying and estimating ages for faces with neutral to moderate tilt. However, due to the relative rarity of extreme tilt angles, the model might face challenges in generalizing to images with pronounced upward or downward tilts unless supplementary data augmentation is integrated.

Data Augmentation Strategy: The dearth of images with significant tilt angles underscores the potential need for data augmentation. Techniques such as rotational transformations might be employed to synthetically create instances with varied tilt orientations.

Feature Relevance: Recognizing the distribution of tilt angles aids in understanding which orientations are most representative in the dataset. This understanding can be pivotal when deciding on the granularity of tilt-based features for the machine learning model.

4.4.2.2 Yaw Angle

Dominant Frontal Orientation: A significant peak around the zero mark underscores that a majority of images showcase faces with a frontal orientation, indicating that subjects predominantly gaze directly at the camera.

Leftward and Rightward Deviations: The histogram manifests symmetrical distributions on both sides of the central peak, reflecting instances where subjects turned their faces to the left (negative yaw) or right (positive yaw). This symmetry indicates an approximately equal representation of left and right profiles within the dataset.

Scant Extreme Profiles: The diminishing frequencies towards the extremities of the histogram highlight that side-profiles, where the face is almost perpendicular to the camera, are relatively rare in the dataset.

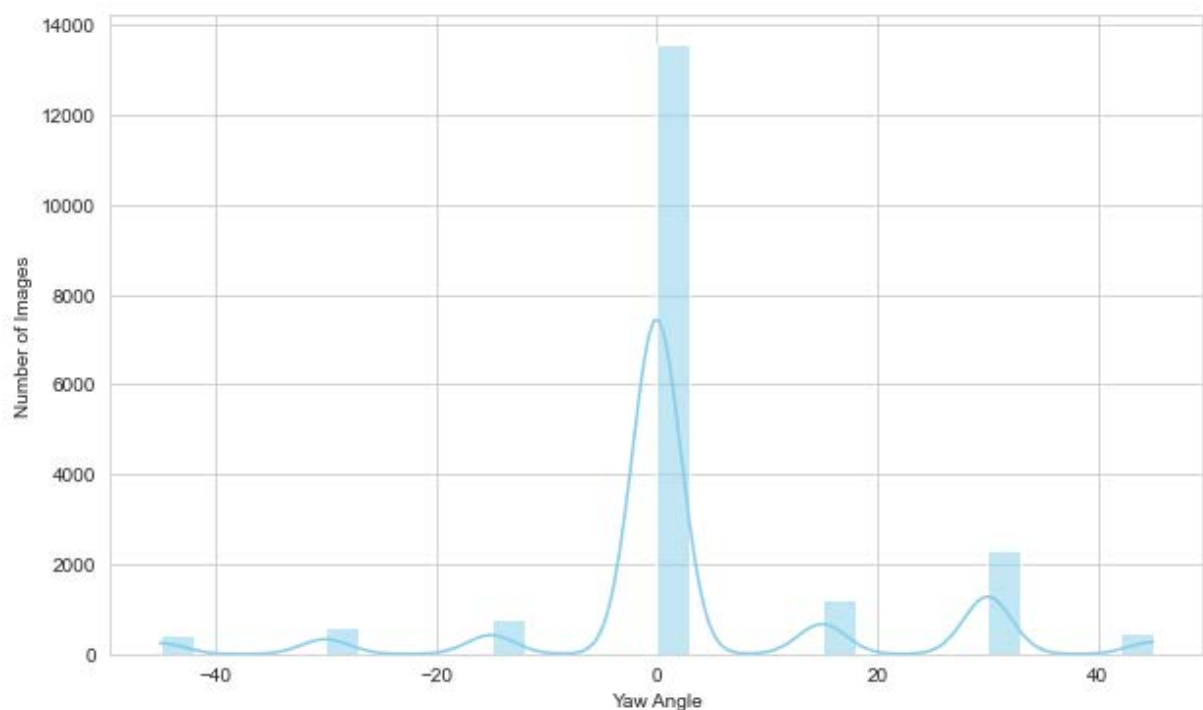


Figure 19 Dstribution of Yaw Angle

Moderate Profile Representation: The pronounced distribution around moderate negative and positive yaw angles suggests that the dataset contains a fair number of images with faces turned slightly away from the camera.

Yaw angles, extracted from the dataset, describe the horizontal orientation of a face, ranging from a full left profile to a full right profile, with a straight-on view as the median. Analysing the histogram renders an understanding of the prevalence of various horizontal orientations within the dataset.

Model Robustness: Given the symmetrical distribution around moderate yaw angles, a model trained on this dataset is expected to exhibit consistent performance for faces turned slightly to

the left or right. However, its proficiency in handling extreme side profiles might be limited due to their scant representation.

Data Augmentation Consideration: Recognizing the sparsity of extreme yaw angles underscores a potential avenue for data augmentation. Mirroring transformations and slight rotations might be deployed to artificially enrich the dataset with diverse yaw orientations.

Relevance in Feature Extraction: Gaining insights into the yaw distribution is pivotal when crafting or selecting features for the machine learning model, ensuring that the model remains sensitive to the nuances of horizontal face orientations.

4.5 Orientation and Angles

During the pre-processing and exploration phase of the Adience dataset, an intricate understanding of the orientation and angles of facial images becomes pivotal. Orientation, in this context, refers to the positioning and direction of the faces in the images. This can be defined by parameters like tilt, yaw, and the bounding box dimensions that encapsulate the facial features.

4.5.1 Correlation between Angles

Utilizing Python's Seaborn library, a correlation matrix was constructed for the following numeric features: 'tilt_ang', 'fiducial_yaw_angle', 'dx', and 'dy'. This matrix provides a succinct overview of the linear relationship between these variables.

Upon examining the matrix, the following observations were gleaned:

Tilt Angle (tilt_ang):

The correlation coefficient between tilt_ang and fiducial_yaw_angle stands at a meager -0.027. Such a value connotes that the tilt (vertical oscillation) of the face exhibits no substantial correlation with the yaw (horizontal oscillation).

A modest negative correlation is discernible between tilt_ang and both dx and dy, gauged at -0.18 and -0.17, respectively. This posits that with an ascendant tilt angle, the bounding box's dimensions delineating the face undergo a marginal contraction.

Fiducial Yaw Angle:

The yaw angle manifests an insubstantial correlation with all other features, having values that verge on zero for dx, dy, and tilt_ang.

DX and DY Dimensions:

A salient correlation is observed between dx (width) and dy (height) dimensions, with a coefficient of 0.99. This infers that any increment in the bounding box's width concurs with a proportional increase in its height, underscoring the typically symmetric nature of facial features.

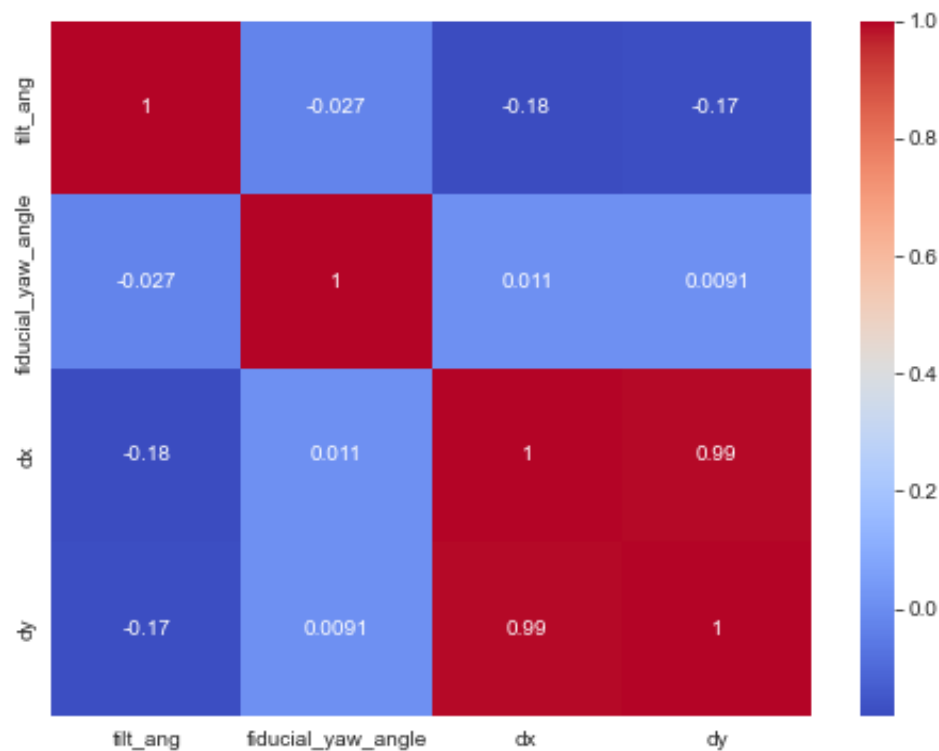


Figure 20 Correlation between Angles

The synthesis of these observations proffers invaluable insights for subsequent stages in the research:

Multicollinearity: The implications of interrelated features need scrupulous scrutiny, especially when deploying regression-based models. Redundant features can induce multicollinearity, potentially obfuscating the interpretation of model coefficients.

Feature Engineering: The identification of correlated features can be instrumental in the genesis of composite attributes or the omission of superfluous ones.

Interpretative Acumen: Correlations proffer an understanding of the interplay between disparate facial attributes. Such insights are not only seminal for a holistic comprehension of the dataset but also pivotal for devising effective feature engineering strategies.

In summation, while the dx and dy dimensions evince a profound correlation, suggesting a near-proportional bounding box, the angles (tilt_ang and fiducial_yaw_angle) demonstrate negligible interrelationships with other features. Such discernments are instrumental in navigating the feature engineering trajectory, ensuring the model captures the quintessence of the dataset while obviating redundant or superfluous attributes.

4.5.2 Orientation

The disparity in the distribution of image orientations, as visualized by Fig 21, raises pertinent considerations for the domain of age estimation. As each orientation presents a distinct perspective of facial features, understanding the implications of this skewness is pivotal.

Inherent Bias in Data Distribution: The dominance of profile images might introduce an inherent bias during the training phase of machine learning models. Such an over-representation might inadvertently train the models to become more proficient at extracting features from profile images compared to frontal ones. Consequently, this might engender a model with heightened sensitivity to profile-oriented facial structures.

Potential Impact on Performance Metrics: When evaluating the performance metrics of models trained on this dataset, it's paramount to factor in this orientation bias. There might be a conceivable discrepancy in the accuracy or other metrics between the two orientations, potentially leading to inflated results for profile images and diminished outcomes for frontal images.

Challenges in Real-World Application: In real-world scenarios, especially in edge devices, the image's orientation might not always be controlled. A model overly attuned to profile images might manifest reduced efficacy when encountering frontal images, which could be a limitation in diverse settings.

Strategies for Mitigation: To counteract the potential ramifications of this orientation skewness, several strategies can be contemplated:

Data Augmentation: One could utilize data augmentation techniques, specifically targeting frontal images. This could entail generating varied versions of frontal images, through rotations, zooms, or other transformations, to bolster their representation.

Orientation-Specific Models: Instead of a single, overarching model, one could consider devising separate models tailored for each orientation. Such specialized models might be better calibrated to the nuances of their respective orientations.

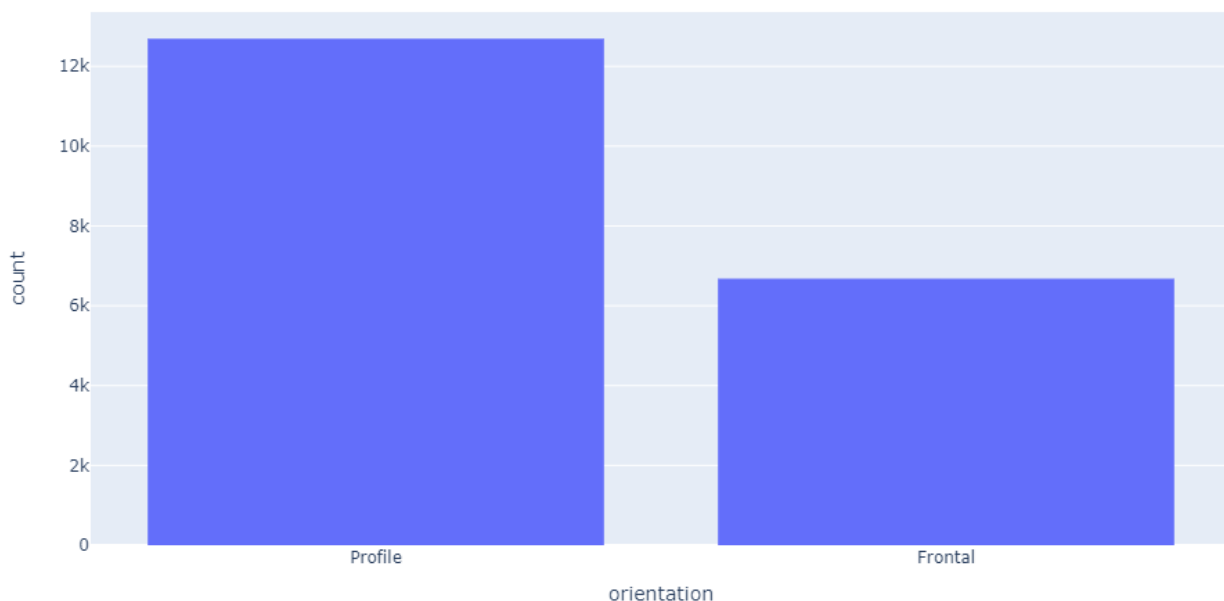


Figure 21 Orientation

Weighted Training: During the training phase, one might contemplate assigning different weights to the two orientations, ensuring that the model doesn't unduly favor one over the other.

In conclusion, while the observed orientation discrepancy in the dataset is noteworthy, it isn't insurmountable. With meticulous planning and strategic measures, the challenges posed by this distribution can be adeptly navigated, paving the way for robust and versatile age estimation models, particularly for deployment on edge devices.

4.6 Summary

This research undertook a comprehensive examination of the Adience dataset, which encompasses 19,370 images spread across various age groups. The overarching objective was to cultivate an in-depth understanding of the dataset's unique characteristics, distributions, and intricacies. Such knowledge serves as the cornerstone for the subsequent stages of model development, particularly in the context of edge device deployment.

A salient orientation discrepancy in the dataset was identified. Nevertheless, with judicious planning, challenges emanating from this discrepancy can be proficiently managed. Addressing such intricacies is instrumental in fostering the development of robust age estimation models, ensuring versatility, especially for edge devices.

Delving into the dataset's features, compelling correlations were observed in the dx and dy dimensions, suggesting a near-proportional bounding box. Conversely, minimal interrelations were discerned between the tilt angle ($tilt_ang$) and the fiducial yaw angle. Such findings accentuate the significance of discerning feature engineering, enabling the model to capture the dataset's essence while eschewing redundancies.

The gender distribution within the dataset stood out as a critical area of exploration. A profound grasp of this distribution shapes not only the model's training trajectory but also its applicability and interpretation in real-world scenarios.

Moreover, a detailed visual analysis shed light on age distribution across dataset folds. Such insights are invaluable, offering guidance for the myriad phases spanning data processing, algorithmic training, and evaluation.

Given the predominance of RGB images in the dataset, a strategic decision was made to forgo additional steps for grayscale image handling. This approach augments computational efficiency, aligns seamlessly with the research's ambition, and simplifies the data pipeline.

Throughout this study, understanding the dataset's inherent structure and its image dispersion across distinct folds remained paramount. Such comprehension ensures every research phase is rooted in an enriched understanding of the data.

Several strategic decisions enhanced the study's efficacy. Omitting instances with unspecified age attributes preserved the dataset's integrity and honed the model's training focus. Additionally, insights from a meticulous outlier analysis—like those concerning face areas—promised to shape both the data preprocessing and model training phases, amplifying the overall model's robustness.

In essence, anomalies such as the unusual y value and tilt angles were detected, underscoring the need for continued scrutiny. The yaw angle's distribution characteristics further enrich our understanding of the dataset.

CHAPTER 5

EXPERIMENT AND RESULTS

In the pursuit of understanding and improving age estimation models for deployment on edge devices, the analytical journey necessitates a rigorous experimentation phase. This chapter delves deep into the methodical experiments conducted, elucidating the procedures, techniques, and methodologies employed. The underpinning goal was to gauge the performance of lightweight machine learning models and discern their capability to approximate or even surpass the benchmarks set by state-of-the-art counterparts.

Using the Adience dataset as the experimental bedrock, an array of model architectures, training paradigms, and optimization strategies were tested. The resultant outcomes, both quantitative and qualitative, are meticulously documented herein. These results offer insights not only into the model's efficacy but also into the underlying factors that contribute to its performance.

Moreover, this chapter seeks to address pivotal questions: Can lightweight models bridge the performance gap in age estimation on edge devices? What are the nuances and intricacies that shape their performance? And most critically, how can these findings guide future research and real-world implementations?

In the ensuing sections, the experimental setup, the methodologies adopted, the results procured, and their implications are presented in a structured and comprehensive manner. This narrative aims to furnish the reader with a lucid understanding of the research's experimental facet, laying the groundwork for subsequent discussions and conclusions.

During the initial phase of experimentation, a pivotal decision involved investigating two distinct model architectures. The divergence between the two variants cantered primarily around their depth and width. One model, referred to as the "Deep Model," was characterized by a greater number of layers but fewer neurons per layer. In contrast, the second model, termed the "Wide Model," had fewer layers but was endowed with a more substantial number of neurons in each layer.

The motivation behind exploring these disparate architectures was rooted in the following considerations:

Model Capacity: There is an ongoing debate in the realm of neural network research regarding depth versus width. While deeper networks can capture more complex features through their hierarchical structure, wider networks might learn more varied features in parallel. By experimenting with both configurations, this research aims to provide empirical insights into this debate in the context of age estimation.

Computational Efficiency: A balance between depth and width can influence not only a model's accuracy but also its computational demands. Given the focus on edge device deployment, understanding the trade-offs between these two configurations is of paramount importance.

Generalization Capabilities: Deeper networks might excel in hierarchical feature extraction, potentially leading to better generalization on unseen data. Conversely, wider networks, with their capacity to grasp a broader set of features, might be less prone to overfitting. Evaluating both configurations allows for an ascertainment of their respective generalization merits.

In alignment with the stipulated research methodology, it should be noted that all models were initially trained employing manual augmentation techniques.

With this bifurcated approach, the experimentation proceeded with the training and evaluation of both models. The subsequent sections delve into the intricacies of the "Deep Model," elucidating its architecture, training dynamics, and performance metrics.

5.1 Deep Model

Utilizing TensorFlow and Keras, the architecture for age estimation is constructed upon the renowned MobileNetV2 model. For this specific implementation, the top classification layer of MobileNetV2 is discarded, making way for a custom classification layer tailored for age group categorization. Here's a detailed explanation of the model and its training process:

Model Architecture:

The architecture is endowed with an input shape of $224 \times 224 \times 3$ and does not include the top layers, commonly referred to as the classifier. The pre-trained weights are initialized from the

ImageNet dataset, and the base model is frozen, thus disabling any further adjustments to its weights during the training phase.

Following the base model, a sequence of custom layers are integrated:

A Global Average Pooling 2D (GAP) layer is introduced to condense the spatial dimensions of the feature maps, effectively reducing the model's complexity while preserving essential feature representations.

Subsequently, a fully connected dense layer with 512 neurons and ReLU (Rectified Linear Unit) activation is incorporated. This layer serves to introduce non-linearity and further refine the high-level features distilled by the base model.

A dropout layer with a rate of 0.5 is employed to mitigate the risk of overfitting by randomly deactivating a portion of neurons during the training process.

Finally, an output layer with six neurons is attached, each corresponding to one of the age groups in focus. A SoftMax activation function is utilized to yield a probability distribution over these age groups.

The model is compiled using the Adam optimizer with a learning rate set at 0.0001.

The loss function employed is the categorical cross-entropy, as it is particularly suited for multi-class classification tasks such as age estimation

Model Parameters:

The composite architecture manifests a total of 2,916,934 parameters, out of which 658,950 parameters are trainable, and 2,257,984 are non-trainable. The memory footprint of the entire model amounts to approximately 11.13 MB. Notably, the custom layers contribute to 2.51 MB of trainable parameters, reflecting the architecture's balance between expressiveness and computational frugality.

Model Training:

In the experimental setup, the model underwent a k-fold cross-validation process to ascertain its performance across multiple subsets of the Adience dataset. The learning dynamics presented a nuanced evolution over epochs, revealing noteworthy trends in the model's

behaviour. Specifically, when trained on the fold excluding the 0th subset for validation, the model achieved an incremental improvement in validation accuracy, escalating from an initial 48.9% to 53.5% over the course of five epochs.

Interestingly, the highest validation accuracy achieved during the entire experiment was 60.6%, observed when the model was trained excluding the 2nd subset for validation. The loss function also exhibited a generally decreasing pattern, indicating effective learning, albeit with observed fluctuations in some validation sets.

Testing Phase:

In the evaluation stage, the model yielded a test loss of approximately 1.204 and an accuracy rate of 49.3%. Despite its robust architectural design and meticulous training regimen, the model's performance warrants further scrutiny, particularly in the realm of generalization to unseen data. The disparity between training and test accuracies points to potential overfitting, thus necessitating future refinements in the model's complexity or training process.

Results:

Confusion Matrix:

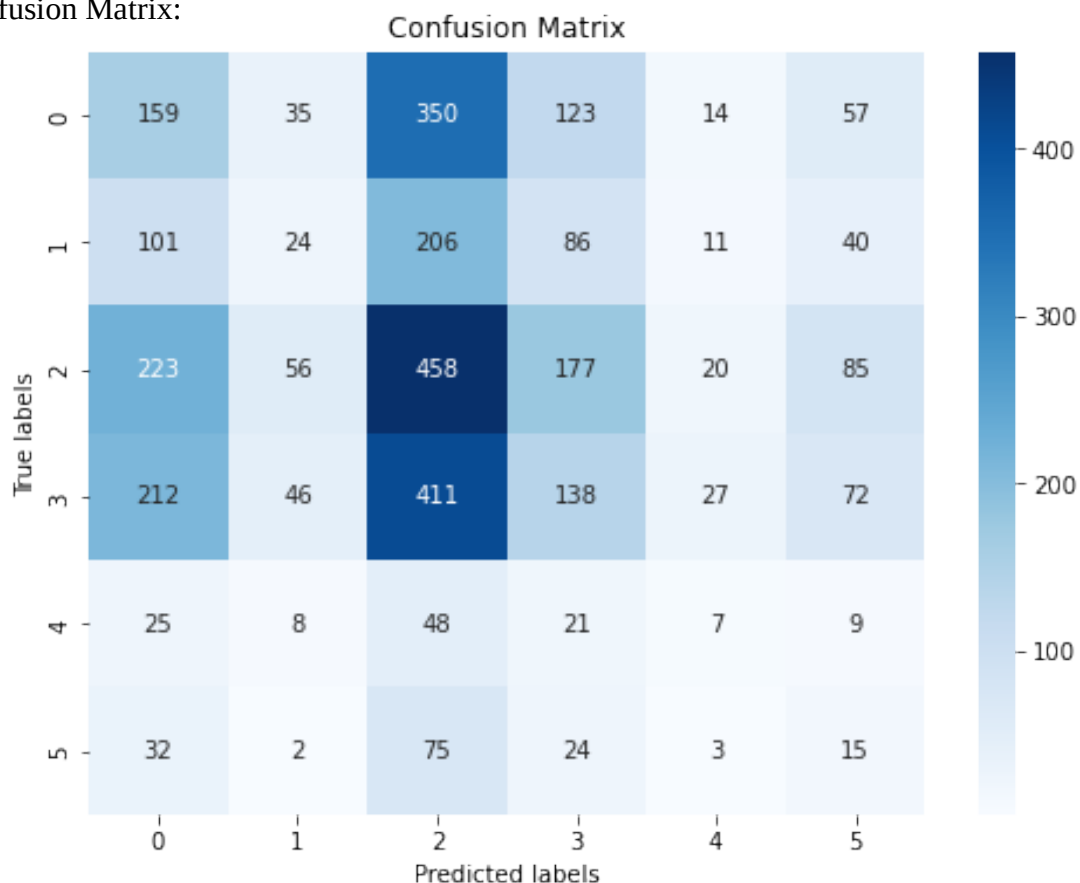


Figure 22 Confusion Matrix Model 1

Diagonal Elements:

The diagonal elements represent instances where the predicted label matches the true label. Higher numbers in these cells indicate successful classifications.

The category "Young Adults" has the highest correct classification count at 458, followed closely by the "Adults" category with 411 correct classifications. This implies that the model is most adept at identifying individuals within these age groups.

Off-Diagonal Elements:

The off-diagonal values signify misclassifications. A high count in an off-diagonal cell indicates that many instances from the true label (vertical axis) were incorrectly predicted as the label represented by the horizontal axis.

Notably, a significant number of "Children" were misclassified as "Teenagers" (350 instances) and vice-versa (101 instances), indicating ambiguity between these two groups for the model.

Predominant Errors:

A large number of "Children" are mistaken for "Teenagers", with 350 misclassifications. Conversely, "Teenagers" are frequently misclassified as "Children" (101 instances) and "Young Adults" (86 instances). This suggests that facial features of these age groups might be relatively similar, presenting challenges in differentiation for the model.

Smaller Classes:

The categories "Seniors" and "Infants" appear to have comparatively fewer instances. While the misclassifications for these groups are numerically lower, the relative error rate might be significant due to the smaller sample size of these categories.

Overall Analysis:

The matrix elucidates that the model faces challenges particularly in differentiating between successive age groups like "Children", "Teenagers", and "Young Adults", given that facial traits evolve progressively with age and might not display pronounced differences between adjacent age groups.

Enhanced training, refined feature extraction, or greater granularity in age categorization could ameliorate these ambiguities.

In summation, while the model exhibits proficiency in certain age categorizations, there are evident areas for improvement, especially in discerning among contiguous age groups. Implementing advanced training techniques, data augmentation, or incorporating additional salient features could potentially enhance the model's predictive precision.

Loss and Accuracy vs Epochs:

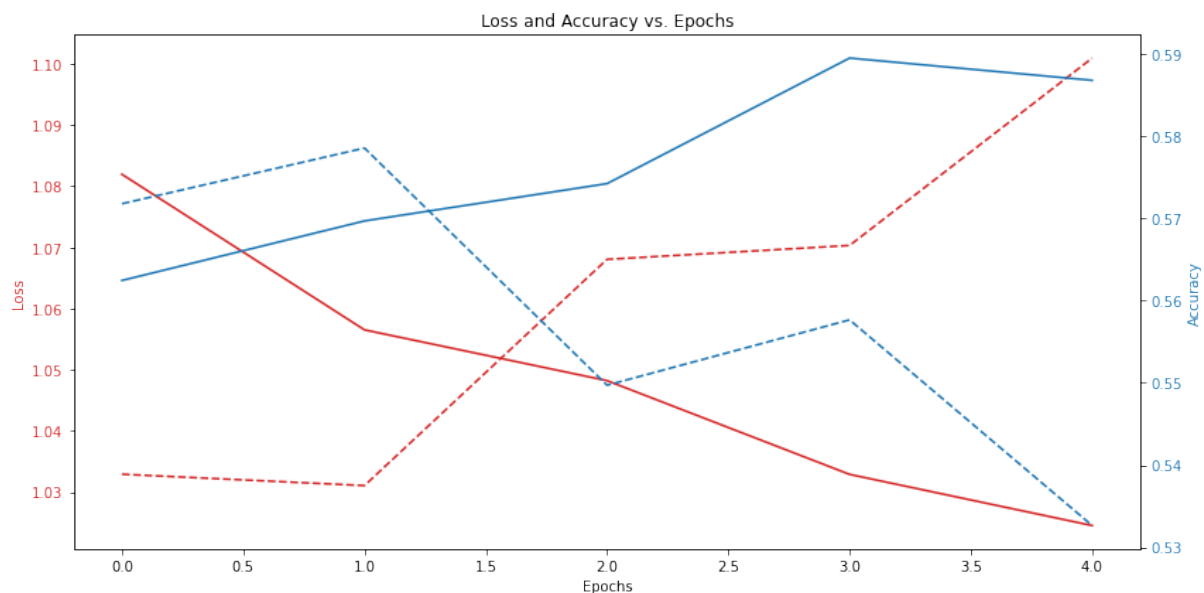


Figure 23 Loss and Accuracy vs Epochs Model 1

The solid red line represents the training loss, while the dashed red line signifies the validation loss.

The training loss consistently diminishes over the epochs, denoting an improved model performance on the training dataset.

The validation loss, conversely, follows a more complex trajectory. Initially, it decreases until around the 2nd epoch, after which it rises until the 3rd epoch. Post this point, the validation loss sees a sharp decline.

The divergence between training and validation loss after the 2nd epoch suggests a temporary overfitting scenario, wherein the model starts to perform exceedingly well on the training data but falters on the validation data.

The solid blue line showcases the training accuracy, and the dashed blue line represents the validation accuracy.

Both the training and validation accuracy exhibit a steady upward trajectory. This signifies that the model's predictive prowess is enhancing across both datasets as the epochs advance.

Although the validation accuracy dips slightly between the 2nd and 3rd epochs, it rebounds and surpasses the training accuracy by the 4th epoch. This indicates a model that is generalizing well to unseen data by the latter stages of training.

The interim rise in validation loss coupled with a drop in validation accuracy between epochs 2 and 3 is indicative of potential overfitting. However, the model seems to rectify this behavior in subsequent epochs, leading to a convergence in performance between the training and validation sets.

By the final epoch, the validation accuracy surpasses that of the training, suggesting a model that is robust and likely to perform commendably on new, unseen data.

In synthesis, the model demonstrates promising training dynamics. The interim signs of overfitting warrant scrutiny.

ROC curve:

Each curve corresponds to one of the age categories: Children, Teenagers, Young Adults, Adults, Seniors, and Infants. The curve showcases the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) for each category at varying decision thresholds.

Ideally, an ROC curve should approach the top-left corner, signifying a high TPR with a minimal FPR. A curve near the diagonal line suggests performance on par with random guessing.

Age Group-wise Analysis:

Children: The blue curve, representative of the 'Children' category, has an AUC of 0.47. This value is slightly below the 0.5 mark, suggesting that the model's ability to classify this group is just above random chance.

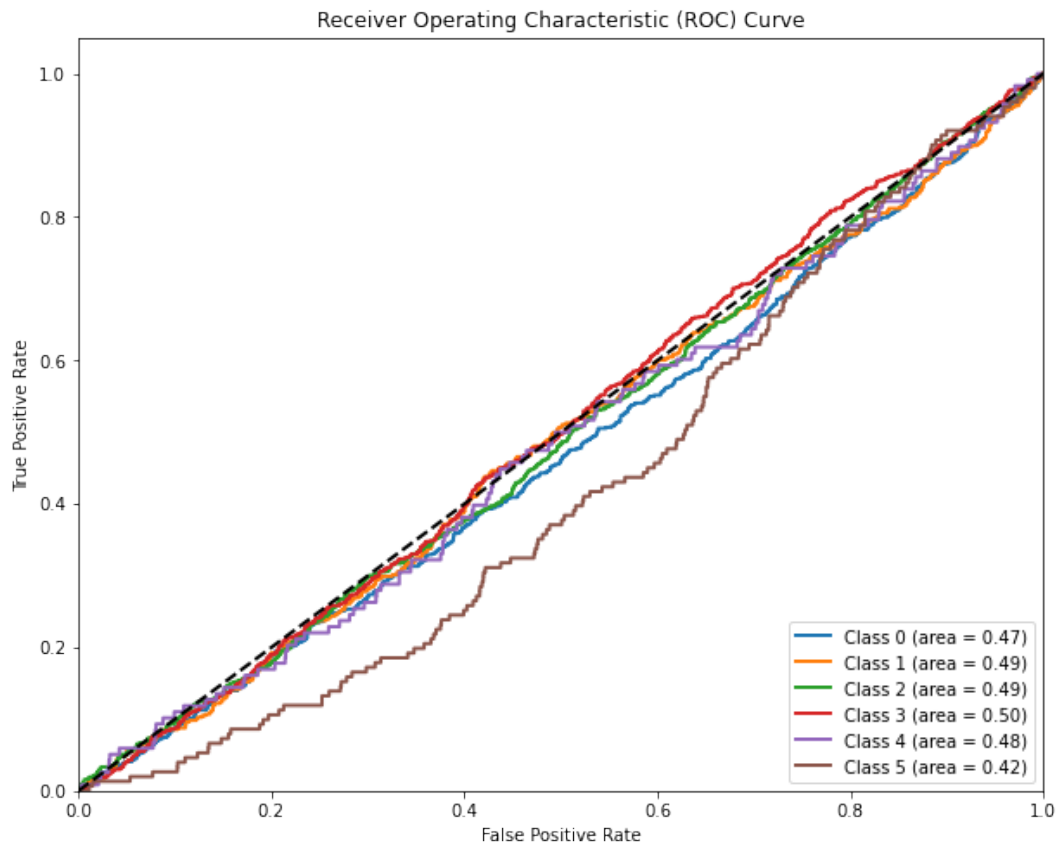


Figure 24 ROC curve Model 1

Teenagers: The curve for 'Teenagers', depicted in orange, yields an AUC of 0.49, indicating that the model's predictive performance for this group is nearly akin to a random classifier.

Young Adults: The 'Young Adults' category, represented by the green curve, has an AUC of 0.50. This score means that the model's predictive capability for this group mirrors that of random guessing.

Adults: The red curve, signifying the 'Adults' group, also has an AUC of 0.50, echoing the performance metrics of the 'Young Adults' category.

Seniors: The 'Seniors' category, delineated in purple, has an AUC of 0.48. This performance is barely above that of a random classifier.

Infants: The brown curve, corresponding to the 'Infants' group, reveals an AUC of 0.42. This is concerning, as it suggests that the model's performance for this category is below that of random guessing.

Overall Observations:

All age categories exhibit AUC values perilously close to 0.5. This signals a model whose current discriminative power is reminiscent of random chance.

The ROC curves for all age groups converge closely with the no-discrimination diagonal line, underscoring the model's constrained discernment across these categories.

Recommendations and Insights:

The present model mandates further refinement to augment its classification prowess for age estimation.

5.2 Wide Model

Model Architecture:

The model inherits the architecture and feature extraction capabilities of MobileNetV2. This ensures the extraction of intricate patterns and features from the input images, granting the model a head-start in the age estimation task.

Convolutional Layers:

A convolutional layer with 16 filters, a kernel size of 3x3, and ReLU activation is appended. It's bolstered with a subsequent Batch Normalization layer to stabilize and expedite the learning process.

Another convolutional layer is appended, this time featuring 32 filters. After its integration, Batch Normalization is applied, succeeded by a MaxPooling layer with a pool size of 2x2, aimed at down sampling the spatial dimensions while preserving pivotal features.

The model is further enriched with a 64-filter convolutional layer, complemented by Batch Normalization and MaxPooling operations. This sequence aids in the extraction of more granular features and further condensing the spatial dimensions.

A final convolutional layer, identical in specifications to the previous one but devoid of the MaxPooling operation, is incorporated. It is followed by Batch Normalization, ensuring consistency in feature distributions.

Fully Connected Layers:

The spatial dimensions of the feature maps are flattened, transitioning from 2D matrices to 1D vectors, priming them for connection with the dense layers.

A dense layer with 128 neurons is annexed, utilizing ReLU activation. This layer instills non-linearity and amplifies the model's discriminative prowess.

To enhance the model's robustness and guard against overfitting, a dropout layer with a rate of 0.5 is integrated, stochastically deactivating a subset of neurons during each training iteration.

Output Layer:

The model culminates with an output layer comprising six neurons, mirroring the number of age categories under examination. The SoftMax activation function is harnessed, rendering a probability distribution spanning the age groups.

Compilation Details:

The model, now architected, is compiled using the Adam optimizer. The learning rate is meticulously set at 0.0001, optimizing the convergence speed and stability. The categorical cross-entropy loss function is employed, tailored for multi-class classification challenges like age estimation. This loss function gauges the disparity between the predicted probability distribution and the actual labels, guiding the model during its learning journey.

In terms of parameters, the model boasts 2,512,694 in total. Of these, 254,102 are trainable, translating to approximately 992.59 KB, while the remaining 2,258,592 parameters, amounting to roughly 8.62 MB, remain non-trainable due to the frozen nature of the MobileNetV2 base.

In essence, the Wide Model Architecture amalgamates the prowess of MobileNetV2 with a series of custom layers, striving for excellence in age estimation while maintaining computational efficiency.

Training Phase

The learning trajectory exhibited a sophisticated progression throughout the epochs, delineating significant patterns in the model's adaptability. Precisely, when the model's training regimen excluded the 0th subset for validation purposes, there was a discernible uptick in validation accuracy. This enhancement saw the accuracy metrics transition from an opening value of 16.6% to 53.3% over a span of five epochs.

Evaluation Phase:

During the assessment phase, the model reported a test loss of precisely 1.256 and an accuracy measure of 51.625%. Given its intricate architectural framework and diligent training approach, the model's output necessitates a more detailed analysis, especially concerning its adaptability to novel data sets. The observable variance between the training and test accuracy metrics suggests a likelihood of overfitting. This observation underscores the imperative for ensuing modifications either in the model's architectural intricacies or its training strategy.

Comparison between both models:

The deep model, while exhibiting a marginally lower test loss, lags slightly in accuracy at 49.254% compared to the wide model, which, despite a higher test loss, achieves a more commendable accuracy of 51.625%. This suggests that while the deep model might be slightly more adept at fitting the training data, the wide model displays superior generalization capabilities on unseen data.

Results:

Confusion Matrix:

Diagonal Elements: These entries denote correct predictions by the model. They represent the number of instances where the predicted age category aligns with the actual age category.

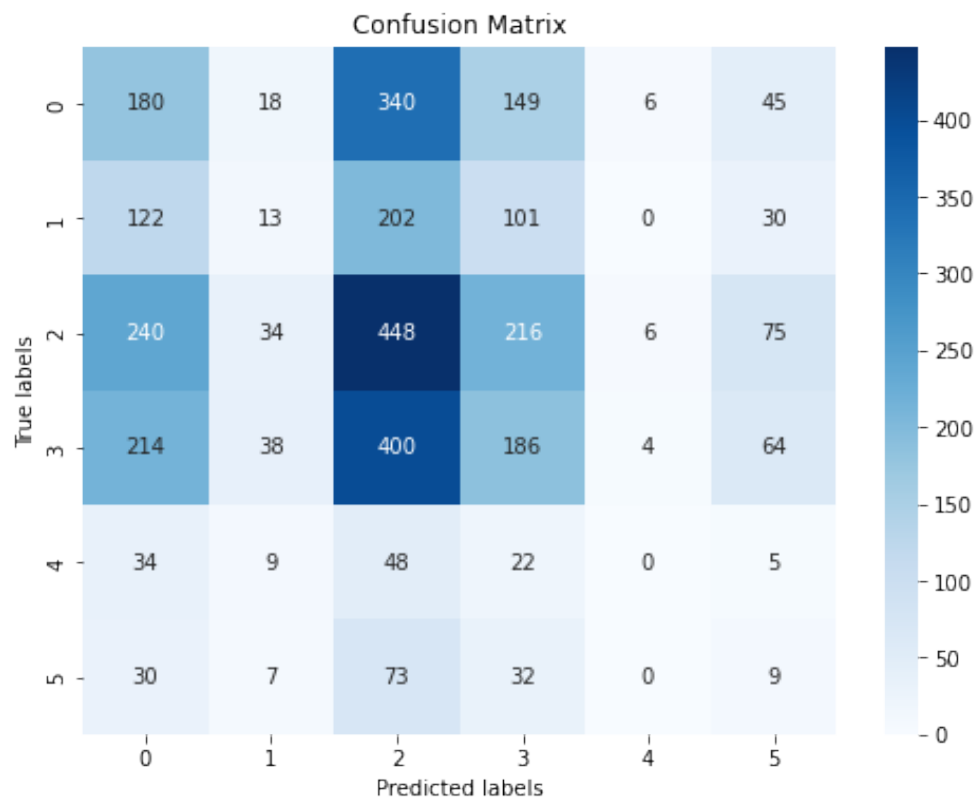


Figure 25 Confusion Matrix Model 2

Children (0): 180 correct predictions.

Teenagers (1): 202 correct predictions.

Young Adults (2): 448 correct predictions.

Adults (3): 400 correct predictions.

Seniors (4): 48 correct predictions.

Infants (5): No correct predictions (0).

Off-Diagonal Elements: These entries indicate incorrect predictions.

For instance, in the first row for Children (0), the model mistakenly predicted 340 as Teenagers, 149 as Young Adults, 6 as Adults, and 45 as Seniors.

Similarly, for Infants (5) in the last row, the model predicted 30 as Children, 7 as Teenagers, 73 as Young Adults, 32 as Adults, and 9 as Seniors.

Interpretations:

Young Adults (2) and Adults (3) have the highest number of correct predictions, indicating that the model is more adept at recognizing these age categories.

Infants (5) appear to be the most challenging category for the model, as it couldn't correctly predict any instance of this group. Instead, the model frequently confused Infants with other age categories, especially Young Adults.

There's a notable misclassification between Children (0) and Teenagers (1), suggesting potential ambiguity in differentiating between these two age categories.

Overall: The matrix provides a clear visual representation of areas where the model excels and where it requires improvement. Enhancing the model's capability to distinguish between closely related age groups, such as Children and Teenagers or refining the data representation for underperforming categories like Infants, might be avenues worth exploring to elevate model performance.

Loss and Accuracy vs Epochs:

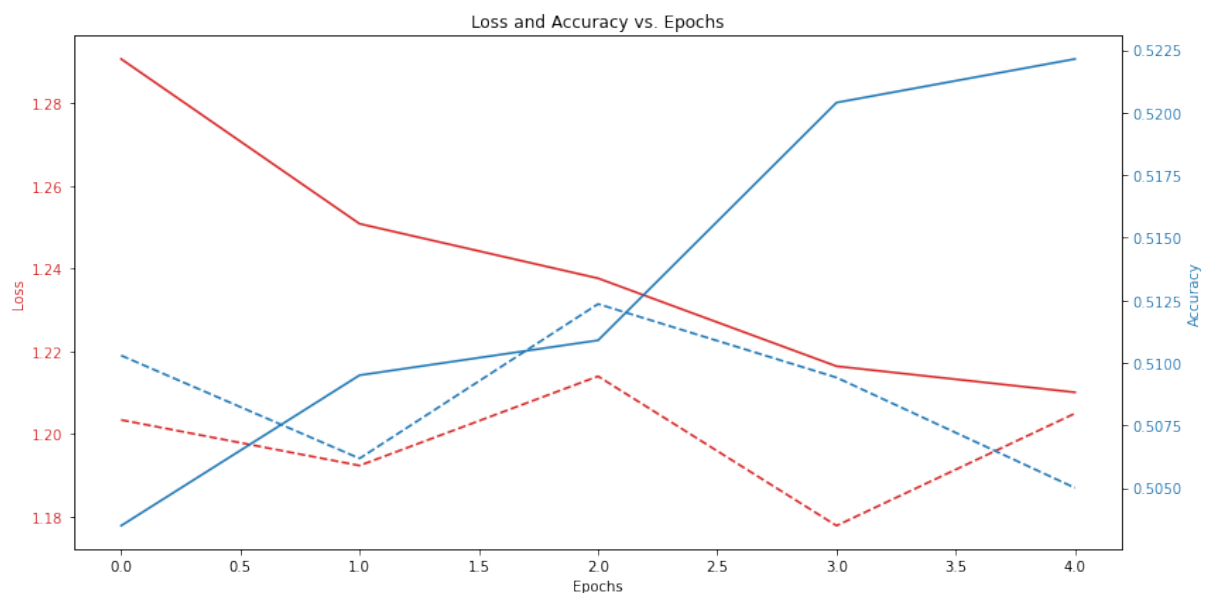


Figure 26 Loss and Accuracy vs Epochs Model 2

Loss:

The solid red curve represents the training loss, which steadily decreases over epochs. This is a typical indication of a model learning from the training data.

The dotted red curve denotes the validation loss. It shows a decline initially but tends to slightly increase after approximately 2.5 epochs. An increasing validation loss while the training loss decreases can be indicative of overfitting. In such a scenario, the model performs well on the training data but starts to lose its generalization ability on unseen or validation data.

Accuracy:

The solid blue curve stands for the training accuracy, which, in a contrasted manner to the training loss, exhibits an upward trajectory. This upward trend signifies the model's improving ability to correctly classify instances from the training set.

The dotted blue line represents the validation accuracy. Following the pattern of validation loss, the validation accuracy peaks around 2-2.5 epochs and then experiences a slight downturn. This decline further supports the potential overfitting inference.

The model seems to exhibit optimal performance in terms of validation accuracy and loss around 2 to 2.5 epochs. Beyond this point, while the model continues to improve on training data (as evidenced by declining training loss and rising training accuracy), its performance on validation data stagnates or even deteriorates slightly.

ROC curve:

Children (Class 0):

AUC = 0.48: The ROC curve for children is very close to the diagonal line, and the AUC is slightly below 0.5. This suggests that the model struggles to distinguish between children and non-children, and might sometimes perform worse than random guessing for this class.

Teenagers (Class 1):

AUC = 0.50: The AUC value for teenagers is exactly 0.5. This indicates that the model's performance in distinguishing between teenagers and non-teenagers is equivalent to random guessing.

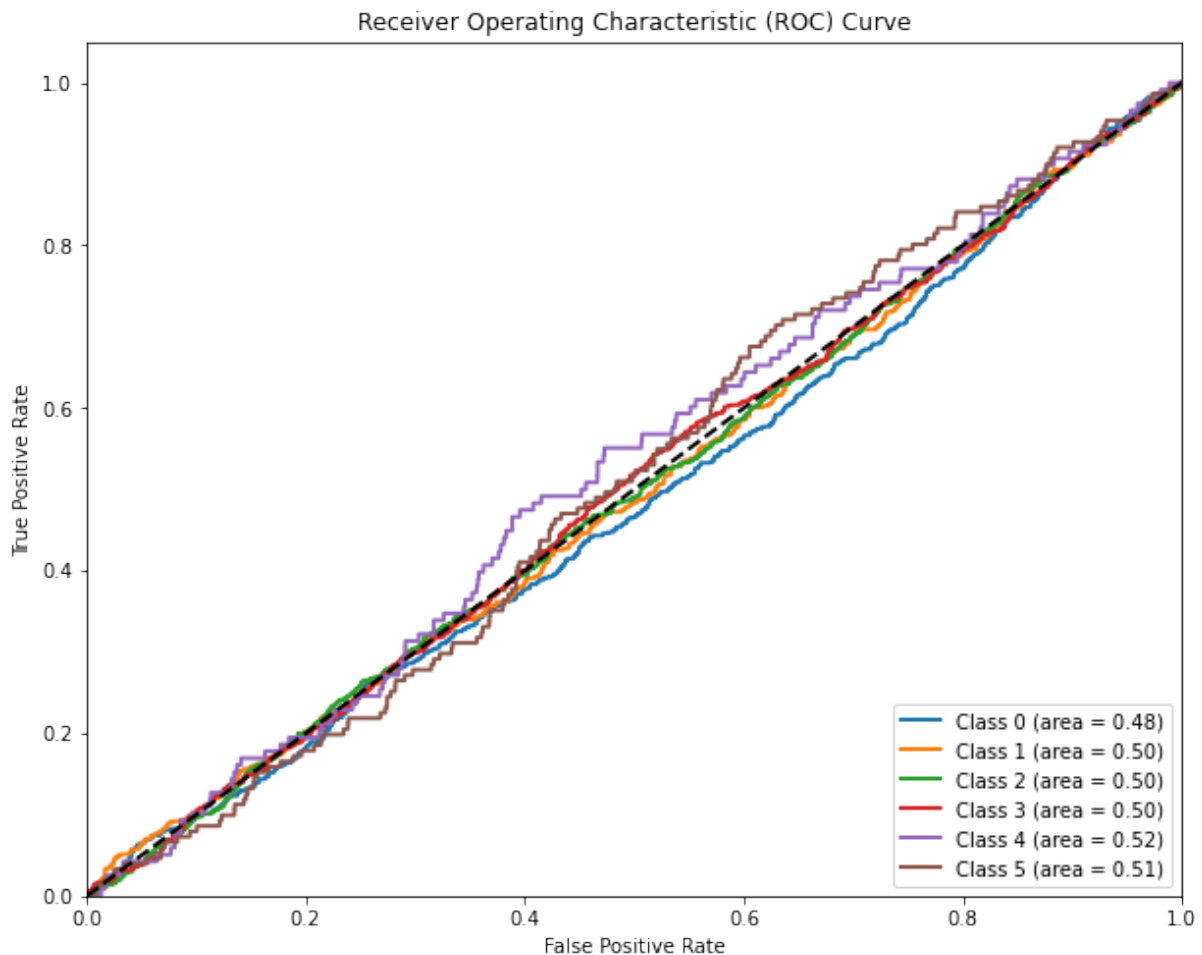


Figure 27 ROC curve Model 2

Young Adults (Class 2):

AUC = 0.50: Similar to teenagers, the model's ability to classify young adults is akin to random chance.

Adults (Class 3):

AUC = 0.50: The curve for adults also parallels the line of randomness, suggesting that the model does not effectively differentiate adults from other age groups.

Seniors (Class 4):

AUC = 0.52: The ROC curve for seniors displays a marginally better performance with an AUC slightly above 0.5. While it's better than the performance for children, it's still near the threshold of random guessing.

Infants (Class 5):

AUC = 0.51: The classification performance for infants is slightly above the line of randomness but remains close to a coin toss in terms of discriminatory capability.

Overall Interpretation:

The ROC curves for all the age groups hover perilously close to the diagonal, indicating that the model's discriminatory power for these age groups is either equivalent to or barely better than random guessing. The AUC values, all clustering around the 0.5 mark, further reinforce this observation.

Upon recognizing the proximity of the AUC values to 0.5 for both the deep and wide models, indicative of a model performance akin to random chance, an initiative to enhance model performance through fine-tuning was undertaken. Fine-tuning is an advanced technique to retrain a pre-trained model on a new dataset, focusing on refining specific layers to make them more adapted to the task at hand.

Refined Implementation:

Here A strategic granularity was introduced in deciding the trainability of layers. Specifically, all layers except the last 15 were rendered non-trainable. This ensured that the foundational, generic features learned from the ImageNet dataset were preserved. The last 15 layers were made trainable. By doing so, these layers were allowed to refine and adapt their weights during the training process.

The batch size was decreased to 64. This reduction can influence the model training in several ways. Smaller batch sizes typically introduce more noise during training, which can act as a form of regularization, potentially preventing overfitting. Additionally, it can lead to a more refined convergence, as the model updates weights more frequently. However, it can also increase the total training time, as more iterations per epoch are required.

5.3 Deep Model with Fine-Tuned Architecture

Model Architecture:

In pursuit of refining the performance of the deep model, specific architectural adjustments were implemented. It's essential to underscore that the foundational architecture remains unchanged from the iteration discussed in the preceding sections. The only divergence arises from the refinements, as delineated in the prior discourse.

Parameter Analysis:

Total Parameters: 2,916,934 (~11.13 MB)

Trainable Parameters: 1,698,950 (~6.48 MB)

Non-trainable Parameters: 1,217,984 (~4.65 MB)

Notably, the refinements have substantially augmented the count of trainable parameters, elevating from 658,950 to 1,698,950. This surge in trainable parameters is anticipatory of enhancing the model's ability to adapt and optimize during training, potentially leading to superior performance outcomes.

Training Phase:

The model's training dynamics displayed an intricate evolution across the epochs, illuminating pivotal nuances in its learning capability. Specifically, with the exclusion of the 0th subset for validation, an observable rise in validation accuracy was manifested. The accuracy metrics advanced from an initial rate of 42.9% to an eventual 66.9% over a series of five epochs.

Testing Phase:

In the testing phase, the model's evaluative metrics conveyed salient information. The test loss was quantified as 1.7589, while the accuracy metric registered at approximately 46.91%.

This underscores the model's performance when exposed to unseen data, highlighting areas for potential improvement and refinement.

Results:

Confusion Matrix:

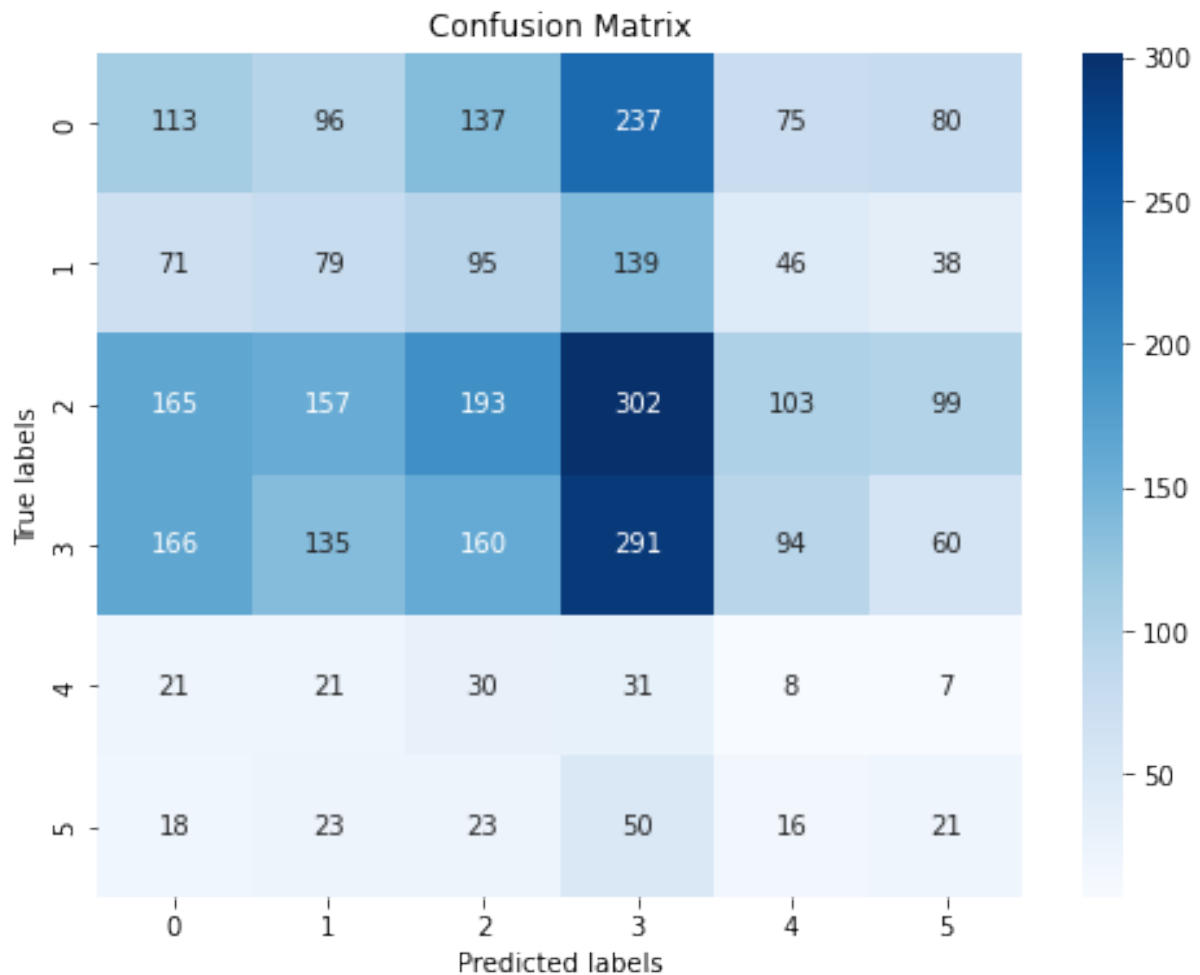


Figure 28 Confusion Matrix Model 3

Upon examining the matrix, several noteworthy patterns emerge.

The group labelled as 'Children' seems to be frequently misclassified as 'Young Adults', with a prominent value of 237. This infers that our model often misconstrues individuals from the 'Children' category as being in the 'Young Adults' category.

For the 'Teenagers' group, the primary area of misclassification is again with 'Young Adults', as indicated by a value of 139.

The 'Young Adults' group manifests a widespread misclassification across different age categories. Yet, the most pronounced confusion is observed within its own class, represented by a value of 302. This suggests that, while the model does manage to correctly classify a substantial number of subjects in this group, there are still challenges associated with misclassifications.

The 'Adults' group exhibits misclassifications predominantly with 'Children', 'Teenagers', and 'Young Adults', with respective values of 166, 135, and 160. Nevertheless, the model exhibited the highest accuracy within its own class, reflecting a value of 291.

The groups 'Seniors' and 'Infants' have relatively fewer overall instances, indicating fewer subjects within these groups in the test set. For both, the main area of misclassification appears to be with 'Young Adults'.

In conclusion, the model predominantly displays confusion between the 'Children', 'Teenagers', and 'Young Adults' categories. Such overlaps might be due to inherent similarities in the visual features of subjects from these age categories. This calls for a more in-depth analysis of feature extraction or potentially the adoption of advanced techniques to heighten the differentiation.

Loss and Accuracy vs Epochs:

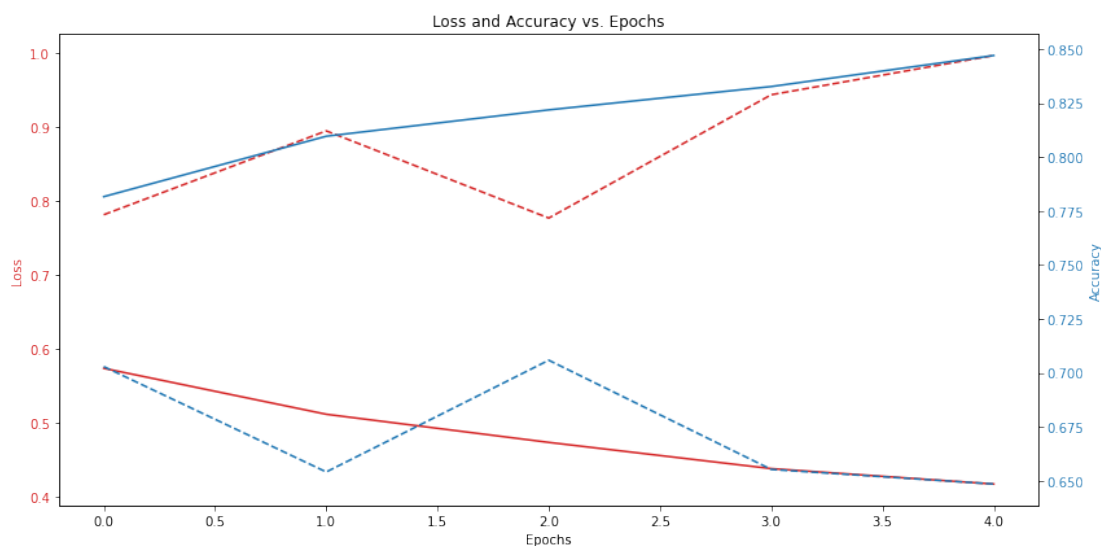


Figure 29 Loss and Accuracy vs Epochs Model 3

The presented graph Fig 29 delineates the trajectory of loss and accuracy metrics as the model undergoes iterative training across epochs. A visual dissection of the graph reveals several discerning insights:

Training Dynamics: Both the loss (solid lines) and accuracy (dashed lines) for training (in blue) and validation (in red) sets are represented. This bifurcation serves to shed light on the model's performance progression and its potential for generalization.

Loss Trends: An evident decline in the training loss is observed, suggesting that the model is gaining competence in recognizing patterns within the training data. Concurrently, the validation loss also appears to decrease, albeit at a slightly subdued pace, signifying that the model is not merely memorizing the training instances but also enhancing its prediction capabilities on unseen data.

Accuracy Evolution: The accuracy trajectory, as indicated by the dashed lines, showcases an upward trend. This ascension is more pronounced in the training set, with the validation accuracy following suit, although with a degree of plateauing towards the latter epochs. This might indicate the onset of saturation in model improvement on the validation set or potential over-optimization on the training set.

Convergence Considerations: A narrowing gap between the training and validation curves, especially towards the latter epochs, could indicate a commendable model generalization. However, any burgeoning divergence could signal overfitting, necessitating interventions like regularization or early stopping.

Epoch Saturation: Post the 3rd epoch, both loss and accuracy curves begin to exhibit subdued variances, implying that additional training beyond this juncture might yield diminishing returns.

ROC curve:

Age Group Analysis:

Infants (Class 5, AUC = 0.53): This category displays the highest AUC among the groups, albeit marginally surpassing the line of no-discrimination (AUC = 0.5). This indicates a slightly better-than-random classification capability for this age group.

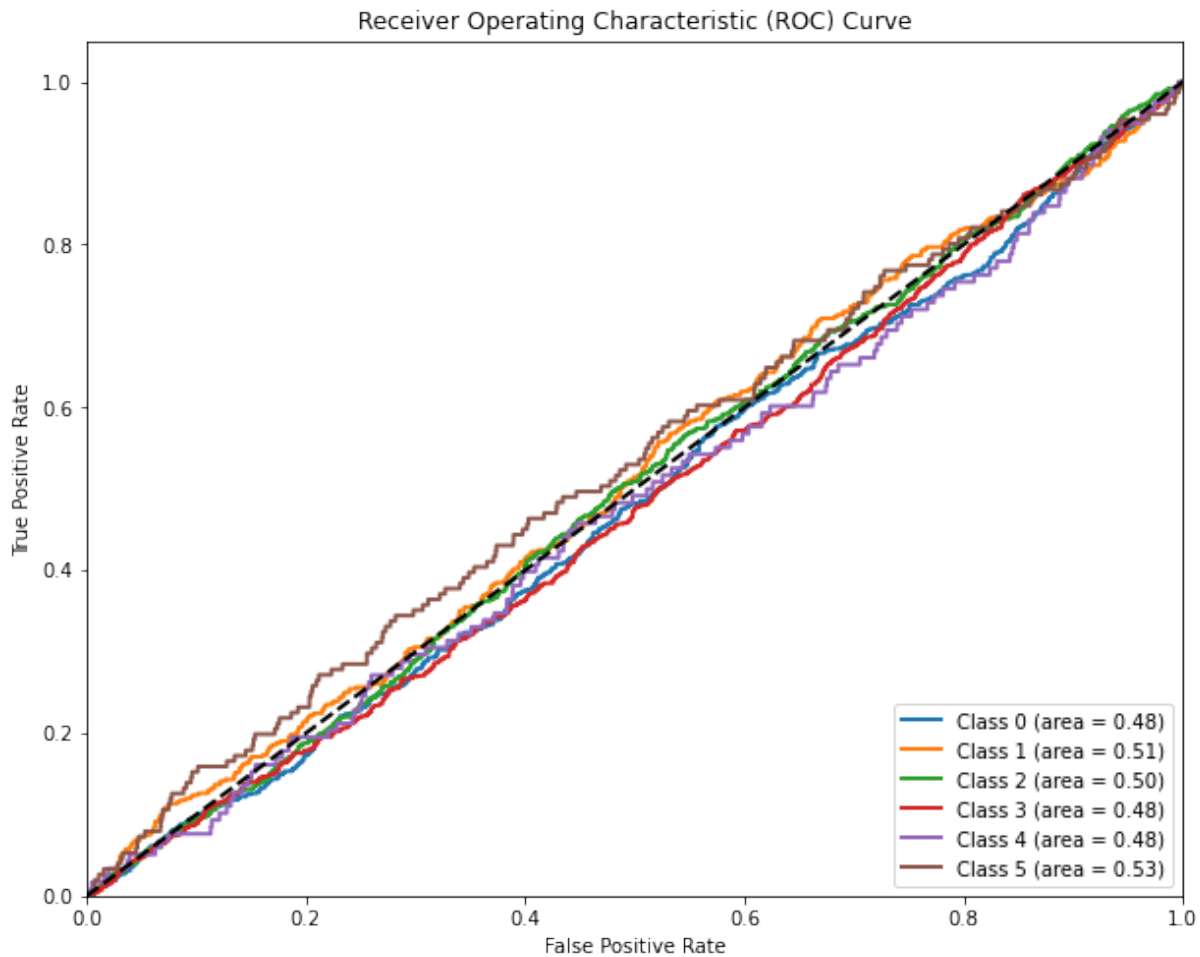


Figure 30 ROC Curve Model 3

Teenagers (Class 1, AUC = 0.51): The curve for teenagers is marginally above the line of no-discrimination, suggesting minimal discriminative power.

Young Adults (Class 2, AUC = 0.50): This category essentially aligns with the line of no-discrimination, implying that the model struggles to differentiate this age group from others effectively.

Adults (Class 3, AUC = 0.48), Children (Class 0, AUC = 0.48), and Seniors (Class 4, AUC = 0.48): These categories manifest AUCs marginally below the no-discrimination threshold, indicative of a model performance slightly inferior to random guessing for these age groups.

Operational Insights: The proximity of most curves to the line of no-discrimination and their overlapping nature may imply inherent challenges in delineating certain age groups.

5.4 Wide Model with Fine-Tuned Architecture

In the quest to enhance the performance of the wide model, certain architectural modifications were instituted. It's imperative to highlight that the core architecture remains consistent with the version elucidated in the previous sections. The sole distinction stems from the enhancements, as detailed in the preceding discourse.

Parameter Analysis:

Total params: 2512694 (9.59 MB)

Trainable params: 1294102 (4.94 MB)

Non-trainable params: 1218592 (4.65 MB)

Upon close examination of the model parameters, a distinct evolution in the distribution between trainable and non-trainable parameters becomes evident. Initially, the model possessed 2,512,694 total parameters, of which a mere 254,102 (992.59 KB) were trainable and a significant 2,258,592 (8.62 MB) were non-trainable. In the updated configuration, while the total parameters remained consistent at 2,512,694, the trainable parameters witnessed a substantial increase to 1,294,102 (4.94 MB), whereas the non-trainable parameters reduced to 1,218,592 (4.65 MB). This re-distribution signifies a greater degree of flexibility and adaptability during the training process for the latter model.

Training Phase:

the model exhibited an initial accuracy of approximately 22.01%, which subsequently improved to around 41.33% by the fifth epoch. However, the validation accuracy displayed a more pronounced ascent, initiating at 32.87% and culminating at an impressive 50.72% by the fifth epoch. The training dynamics further unravelled disparities in performance across different validation sets. For instance, when training with the exclusion of fold 1, peak accuracy touched 56.64% by the fifth epoch, while validation settled around 43.08%. Remarkably, the exclusion of fold 2 witnessed the model achieving an apex validation accuracy of approximately 62.43% during its inaugural epoch. These variations emphasize

the model's adaptability, yet they also underline the complexities and challenges of the diverse Adience dataset distribution.

Testing Phase:

In the assessment phase, the model manifested distinct evaluative characteristics. The test loss was computed to be 1.2732, and the accuracy metric was discerned at approximately 52.26%. Such metrics elucidate the model's efficacy on unfamiliar datasets, accentuating the scope for further optimization and enhancement

Results:

Confusion Matrix:

An analysis of the matrix showcases:

Children: The model most frequently predicted this group accurately with 133 correct classifications. However, there were some misclassifications, especially with Young Adults (216 instances) and Adults (211 instances).

Teenagers: This group saw 80 correct classifications. Ambiguities arose predominantly with Young Adults (142 instances) and Adults (128 instances).

Young Adults: This segment had the highest overall classifications with 267 instances correctly identified. A notable overlap occurred with the Adults category, where 296 instances were misclassified.

Adults: For this category, the model correctly classified 242 instances. Young Adults posed the most challenge with 261 misclassifications.

Seniors: The model appears to struggle with this age group, only accurately identifying 36 instances. The most frequent misclassification occurred with Young Adults (34 instances).

Infants: Infants had a comparatively low representation with 45 correct classifications. Misclassifications were spread across Teenagers (39 instances) and Young Adults (45 instances).

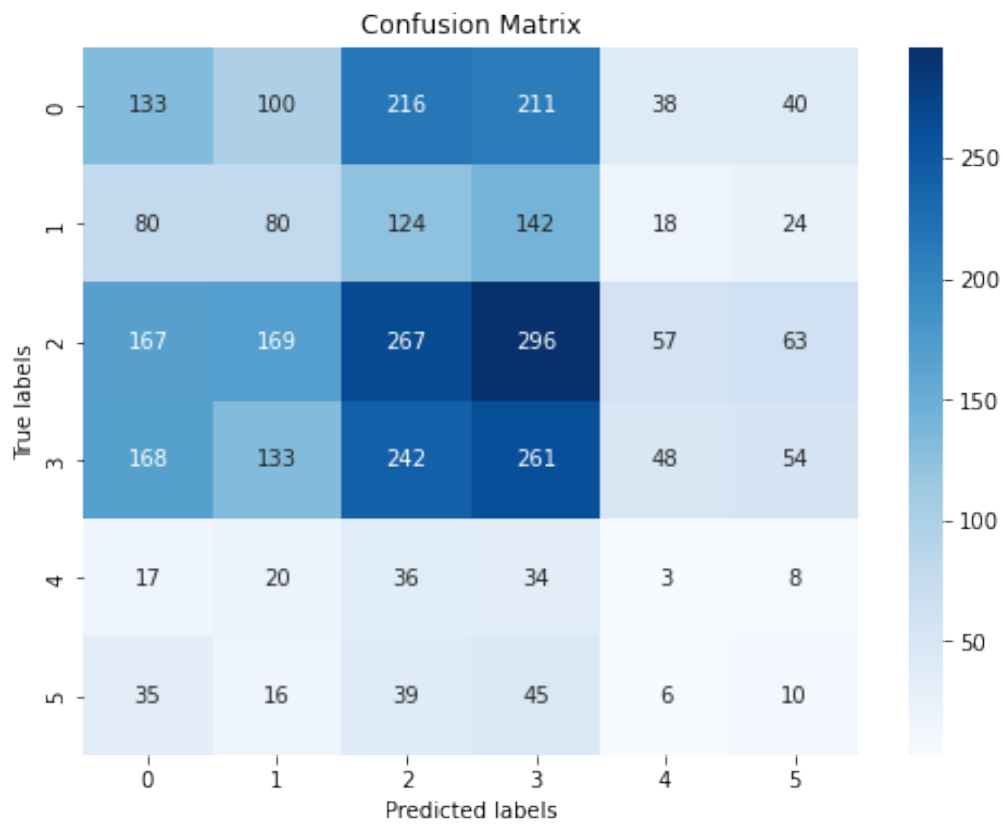


Figure 31 Confusion Matrix Model 4

This matrix, thus, reveals not only the strengths of the model in certain age classifications but also areas of potential confusion, particularly between adjacent age categories such as Young Adults and Adults.

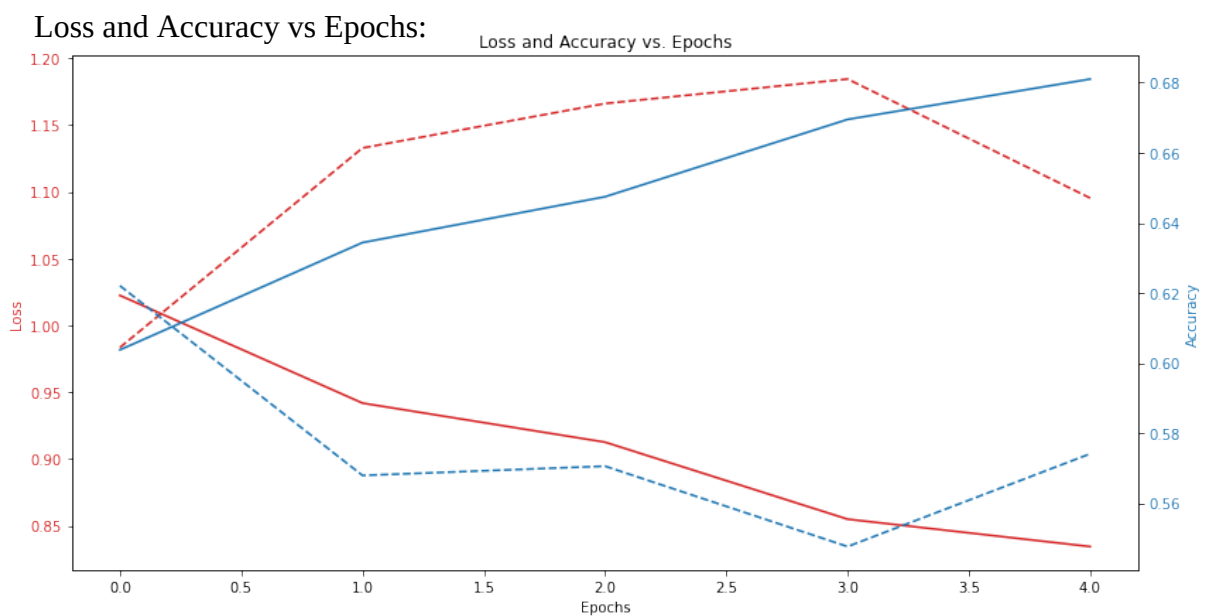


Figure 32 Loss and Accuracy vs Epochs Model 4

The graph delineates the model's evolution in terms of loss and accuracy over the course of four epochs. Observations from the graph include:

Loss Dynamics: The solid blue curve represents the training loss, which exhibits a descending trajectory, starting slightly above 1.10 and concluding below 0.90. Conversely, the dashed red curve, signifying the validation loss, remains relatively stable initially and then witnesses an upward trend, surpassing the 1.10 threshold by the fourth epoch. This suggests the model is learning from the training set, but there might be an onset of overfitting indicated by the increasing validation loss.

Accuracy Trends: The dotted blue curve charts the training accuracy, revealing a promising uptick from around 0.56 to nearly 0.68 over the epochs. In contrast, the validation accuracy, depicted by the dotted red line, begins slightly below 0.60 and encounters a dip, stabilizing near the 0.58 mark by the fourth epoch.

The contrasting trends between the training and validation metrics underline the model's proficiency in assimilating the training data. However, the diminishing validation accuracy and ascending validation loss hint towards potential overfitting

ROC curve:

Key Observations:

Children (Class 0): The curve associated with 'Children', represented in blue, exhibits an area under the curve (AUC) of 0.51. This value, being marginally above the no-discrimination line (an AUC of 0.5), suggests a nearly random classification capability for this category.

Teenagers (Class 1): With an AUC of 0.48, the performance for 'Teenagers', delineated in orange, tends to be slightly below the benchmark of random guess, indicating challenges in accurately identifying this group.

Young Adults (Class 2) and Adults (Class 3): Both these categories, depicted in green and red respectively, have AUC values of 0.51, showcasing similar classification efficiencies as that of 'Children'.

Infants (Class 5): Lastly, the curve for 'Infants', highlighted in brown, demonstrates an AUC of 0.50, implying a random classification ability for this age group, akin to flipping a coin

Seniors (Class 4): The 'Seniors' category, represented in purple, possesses the lowest AUC of 0.43, elucidating considerable misclassification and the model's difficulty in accurately discerning this age group.

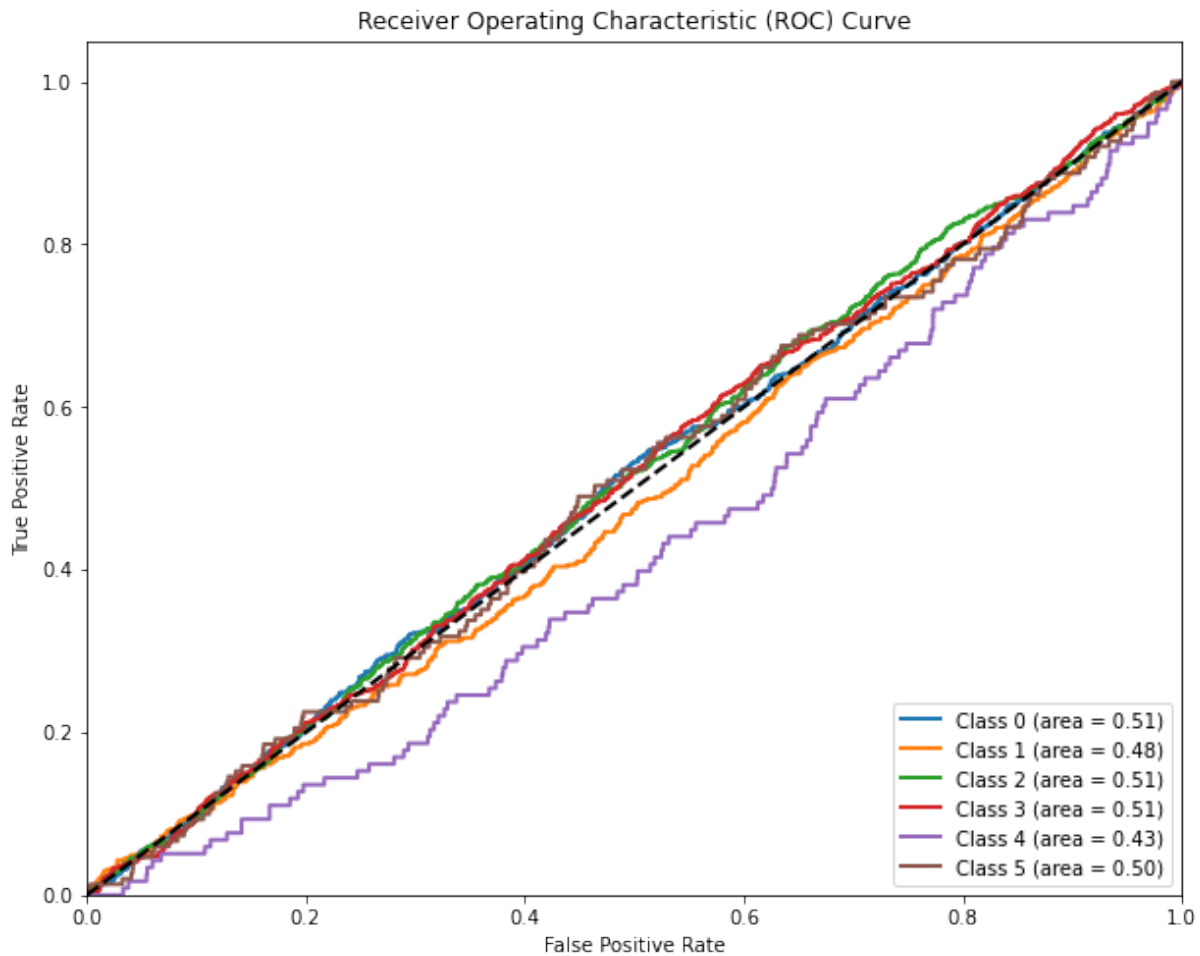


Figure 33 ROC Curve Model 4

The proximity of all AUC values to 0.5 underscores the model's challenges in distinguishing between the age groups effectively. While the model demonstrates some discernment capability, there remains ample scope for enhancement to elevate its discriminative prowess.

Despite rigorous fine-tuning endeavors, the model's performance remains suboptimal, emphasizing the limitations inherent in manual augmentation techniques. Consequently, the study transitions to employing the Augmentor library, a more sophisticated and versatile approach to image augmentation. Additionally, this library facilitates addressing the

persistent issue of class imbalance through strategic sampling techniques, aligning with the strategies outlined in the research methodology.

On evaluating the relative performances of deep and wide network architectures, certain distinctive attributes emerge. The wide network consistently exhibits superior generalization capabilities, aligning more harmoniously with the stipulations of our problem statement. Such an observation delineates the inadequacies of the deep network for this specific task. Hence, in pursuit of efficiency and efficacy, the decision has been made to abandon the deep network architecture. The focus will henceforth be cantered exclusively on the wide network. Moreover, in an attempt to further refine the model's performance, the number of epochs will be amplified to 20, allowing the model more iterations to assimilate patterns and hone its predictive acumen.

5.5 Wide Model with Fine-Tuned Architecture and Augmentor Library

Model Architecture:

In the refined architecture of the wide model, several modifications were incorporated to optimize its performance further. While the majority of the structure remains unaltered, discernible changes have been made in the dense layers. Specifically, the penultimate dense layer, which previously comprised 128 neurons, has been recalibrated to encompass only 64 neurons. Furthermore, one of the two dense layers, each initially containing 64 neurons, has been entirely excised from the architecture. These alterations were strategized to streamline the model, potentially enhancing its responsiveness and efficiency without compromising its capacity to generalize across the dataset.

Model Parameters:

Total params: 2470710 (9.43 MB)

Trainable params: 1252374 (4.78 MB)

Non-trainable params: 1218336 (4.65 MB)

Training Phase:

The model initiated its training journey with an accuracy approximating 19.92%, which witnessed a gradual refinement over subsequent epochs. By the culmination of the initial epoch, the model had garnered a validation accuracy of about 17.69%, underscoring the challenges innate to the early stages of training. As the training epochs progressed, the model's adeptness at learning from the data became evident. In an illustrative turn of events, by the 12th epoch, the model managed to attain a commendable accuracy of approximately 81.12%, and the validation accuracy experienced a notable surge, reaching 91.18%. This marked a significant leap, representing the zenith of the model's performance throughout the training phase. However, by the twentieth epoch, the improvements plateaued. The model posted an accuracy of around 82.20%, but the validation accuracy refrained from exceeding its previous best of 91.18%. The data underscores the model's commendable learning trajectory, despite the intricacies and nuances presented by the Adience dataset.

Testing Phase:

During the evaluative stage, the model showcased specific performance traits. The computed test loss stood at 1.98795, highlighting the challenges faced by the model on unseen data. The corresponding accuracy was registered at roughly 48.64%.

Results:

Confusion Matrix:

The confusion matrix provides a structured visualization for assessing model performance on individual age groups. The matrix for Fold 4 offers insights into the model's accuracy and its potential areas of misclassification.

The label 'Children' (0) exhibits the highest correct predictions, with 350 instances accurately classified. However, there are notable misclassifications into 'Teenagers' (1) and 'Young Adults' (2) with 232 and 105 instances, respectively.

For 'Teenagers' (1), the accurate classifications amount to 56. Yet, it's evident that a large fraction of 'Teenagers' is misclassified as 'Children' (0), with 84 instances.

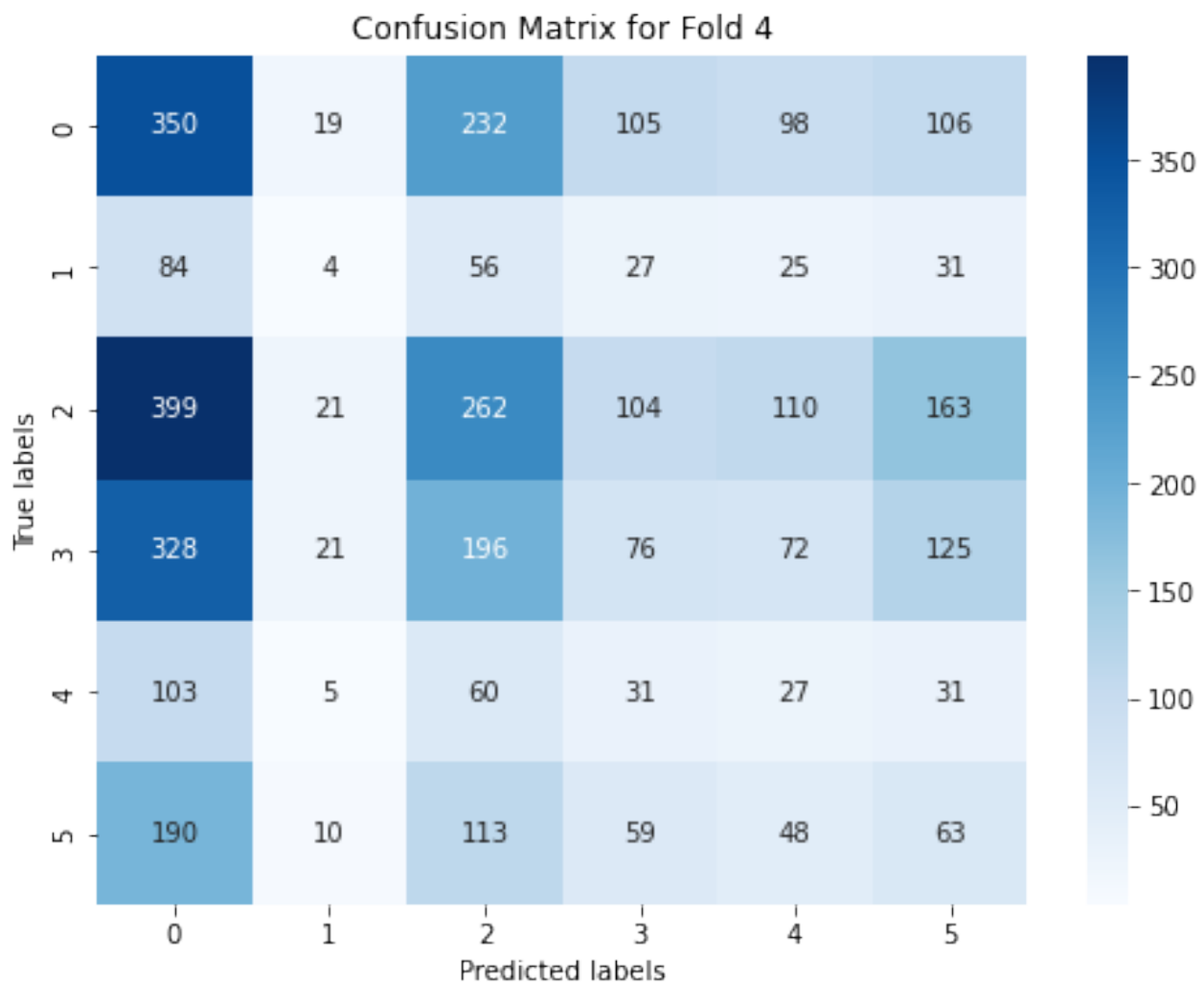


Figure 34 Confusion Matrix Model 5

'Young Adults' (2) has a commendable classification rate with 262 instances correctly predicted. Nevertheless, there's a prominent misclassification into the 'Children' (0) and 'Seniors' (4) categories, with 399 and 50 instances, respectively.

The 'Adults' (3) category presents 196 correct classifications. A key observation is the significant misclassification into the 'Children' (0) and 'Young Adults' (2) labels, recording 328 and 76 instances, respectively.

For 'Seniors' (4), the matrix indicates 31 accurate predictions. A substantial number of 'Seniors' are misclassified as 'Children' (0) and 'Young Adults' (2), with 103 and 60 instances, respectively.

The 'Infants' (5) category is observed to have 48 correct predictions. However, the category witnesses a high misclassification into the 'Children' (0) label, with 190 instances.

Loss and Accuracy vs Epochs:

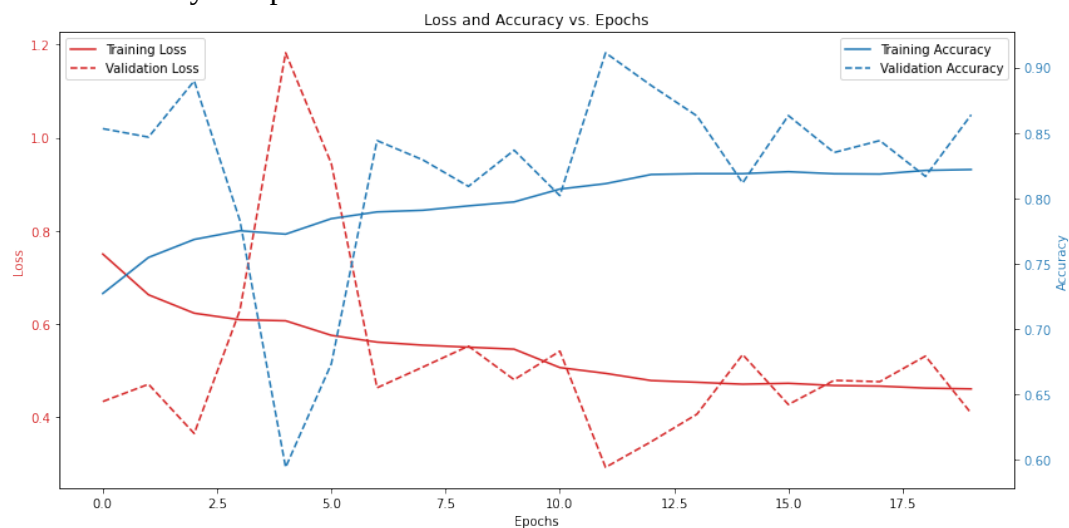


Figure 35 Loss and Accuracy vs Epochs Model 5

Training Loss (Solid Red Line):

The training loss, represented by the solid red line, commences at a value slightly above 1.0. It displays a sharp decline as the epochs progress, reaching its nadir around the 10th epoch.

Post this point, a gradual ascent in loss is observed, suggesting the onset of overfitting where the model might be memorizing the training data.

Validation Loss (Dashed Red Line):

The validation loss, portrayed by the dashed red line, demonstrates initial fluctuations. It begins just below 1.2, dips considerably around the third epoch, then sees a surge near the fifth epoch.

Subsequent epochs exhibit a relatively stable trend with minor undulations, suggesting the model's consistent performance on the validation set.

Training Accuracy (Solid Blue Line):

The training accuracy, depicted by the solid blue line, starts close to 60% and exhibits a steady increase. It peaks slightly above 80% around the 10th epoch.

However, it shows a gentle decrease in the subsequent epochs, possibly an indication of the model's reduced ability to generalize as it becomes more attuned to the training data.

Validation Accuracy (Dashed Blue Line):

The validation accuracy, represented by the dashed blue line, reveals an oscillatory behavior. Beginning slightly below 65%, it experiences intermittent ascents and descents.

Notably, it reaches its zenith close to the 15th epoch, surpassing 80%, and subsequently maintains a consistent trend.

The juxtaposition of training and validation metrics provides insights into the model's learning dynamics. While the model demonstrates commendable performance, the signs of overfitting are evident post the 10th epoch.

ROC Curve:

The diagram showcases the ROC curve for each of the six age categories.

Children (Class 0):

The ROC curve for the "Children" category is depicted by the first line, with an area under the curve (AUC) of 0.50. This AUC value indicates that the model's discriminative power for this class is equivalent to random guessing.

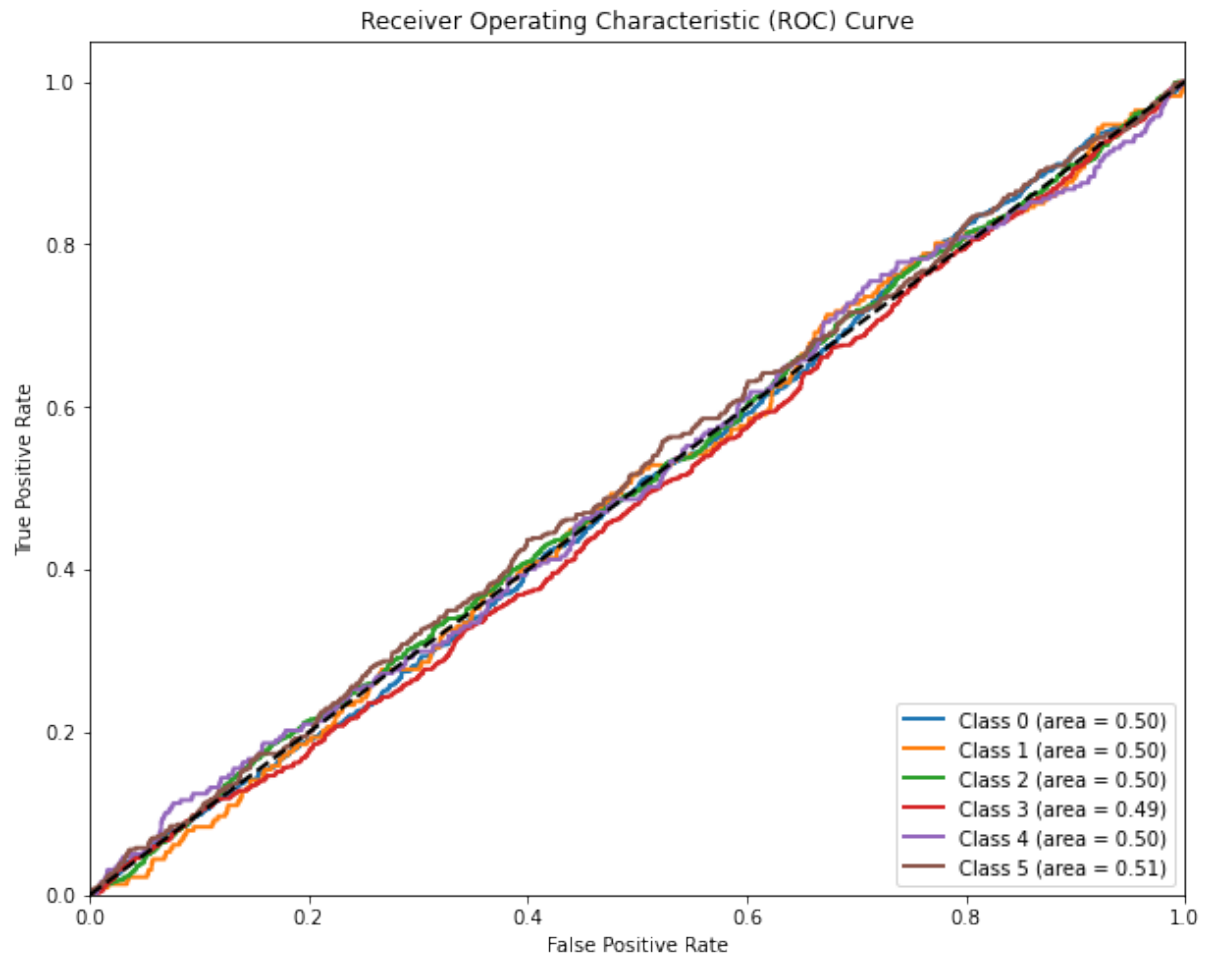


Figure 36 ROC Curve Model 5

Teenagers (Class 1):

Analogous to the "Children" category, the "Teenagers" class also manifests an AUC of 0.50, suggesting a similar level of discrimination capability.

Young Adults (Class 2):

The curve representing "Young Adults" exhibits an AUC of 0.50, reinforcing the observation of the model's challenge in distinguishing this category distinctly.

Adults (Class 3):

The "Adults" category, represented by the fourth line, again parallels its predecessors with an AUC of 0.50, underscoring a pervasive trend across these age groups.

Seniors (Class 4):

The ROC curve for "Seniors" mirrors the earlier patterns, yielding an AUC of 0.50.

Infants (Class 5):

A slight deviation is observed for the "Infants" class, which possesses an AUC of 0.51. Though marginally better, this AUC still resonates closely with the random guessing benchmark.

The close clustering of the ROC curves, combined with the uniformly proximate AUC values, insinuates that the model grapples with discerning between the age categories. An AUC of 0.50 for most classes elucidates that the model's capability to differentiate among these classes is comparable to a random classification.

The utilization of the Augmentor library played a pivotal role in bolstering the model's learning capabilities. Its implementation addressed issues pertinent to class imbalance, thereby facilitating an enhanced training phase. Nonetheless, despite these strides, the model displayed evident overfitting. Such overfitting often culminates in the model capturing noise rather than the underlying data distribution, resulting in compromised generalization on unseen datasets.

The inherent data folds provided with the models, although beneficial in certain contexts, may not be optimally conducive for our specific use-case. Given the observations during the training phase and the learning capacity demonstrated by the model, it becomes prudent to contemplate alternative data partitioning techniques. One robust approach that beckons consideration is the Stratified K-Fold Cross-Validation. This methodology not only ensures that each fold is a good representative of the entire dataset but also preserves the distribution amongst classes. Such a meticulous distribution preservation is paramount for age estimation tasks, given the nuanced differences between adjacent age classes.

5.6 Wide Model with Fine-Tuned Architecture and Stratified Cross Validation

Model Architecture and Parameters:

The model leverages a wide architecture, retaining the parameters and configurations from previous iterations. The meticulous design choices in the architecture aim to capture intricate age-related features from the input data

Training Phase:

Commencing its training journey, the model showcased an initial accuracy of approximately 94.95%. By the end of the first epoch, the validation accuracy had ascended to 42.03%. This early divergence between training and validation accuracy illuminates the inherent complexities and challenges often encountered at the inception of model training. With the progression of epochs, the model's capacity for effective learning began to shine through. By the twelfth epoch, an emblematic milestone was reached as the validation accuracy augmented to 69.40%. This proved to be the pinnacle of the model's validation performance throughout its training regime. Despite advancing to the eighteenth epoch, the validation accuracy did not surpass its zenith from the twelfth epoch, stabilizing at approximately 67.58%. The persistent training accuracy, coupled with a more fluctuating validation score, emphasizes the nuanced intricacies the model grappled with on the Adience dataset.

Testing Phase:

During the assessment phase, the model unveiled distinct performance characteristics. The deduced test loss was quantified at approximately 1.35256, underscoring the intricacies the model encountered when confronted with previously unexposed data. The concomitant accuracy was tallied at an estimable 69.92%, further elucidating the model's adeptness in generalizing to novel data instances.

Of all the preceding methodologies employed, this model emerged superior, exhibiting a marked improvement with an accuracy of 69.92% compared to the previous zenith of 52.26%.

Results:

Confusion Matrix:

The confusion matrix serves as an efficacious tool for visualizing the performance of the model across various age categories. A close inspection of the matrix divulges insightful details:

Children (Class 0): The model exhibited robust performance, correctly classifying 279 instances. However, there were instances of misclassification, with 82 instances mistaken as 'Young Adults' and 32 as 'Infants'.

Teenagers (Class 1): Of this category, 87 instances were aptly categorized. Notable misclassifications included 65 instances erroneously deemed 'Young Adults'.

Young Adults (Class 2): Demonstrating a significant prowess, the model correctly identified 431 instances. However, a portion was misclassified, with 55 instances perceived as 'Adults'.

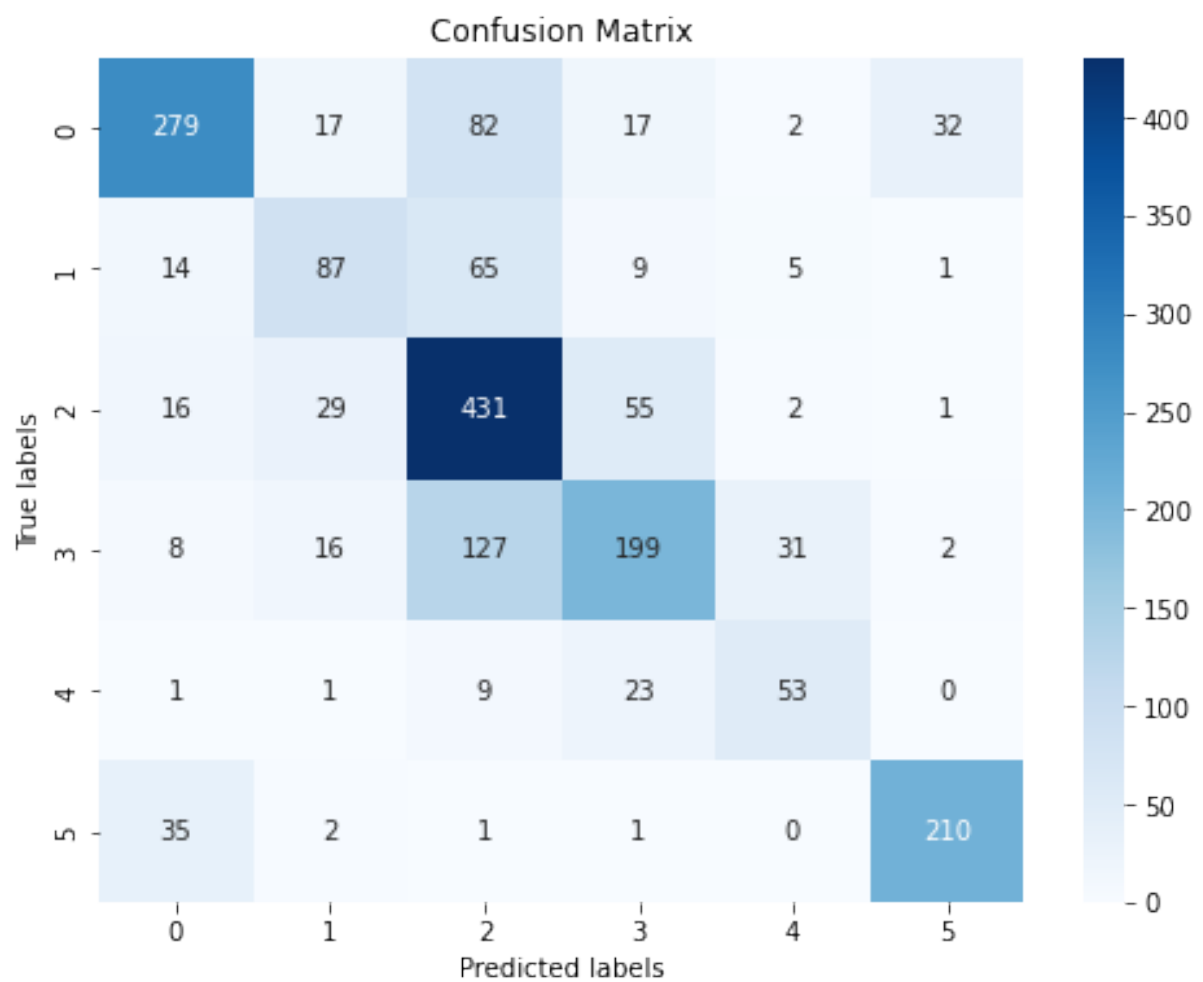


Figure 37 Confusion Matrix Model 6

Adults (Class 3): This group revealed a diverse set of results. While 199 instances were accurately classified, 127 instances were mistaken as 'Young Adults'.

Seniors (Class 4): This category reflected a certain challenge for the model, with only 53 instances rightly classified as 'Seniors'. There were prominent misclassifications, especially with 23 instances misinterpreted as 'Adults'.

Infants (Class 5): The model performed with appreciable efficiency for this class, recognizing 210 instances correctly. However, it's noteworthy that 35 'Infants' were wrongly categorized as 'Children'.

In essence, the matrix underscores the model's commendable aptitude in recognizing 'Young Adults', 'Children', and 'Infants' while delineating areas of improvement for the 'Seniors' and 'Adults' classifications. The subtle nuances of age transitions, particularly between adjacent age groups, often pose challenges, warranting further refinement in the model for optimal precision.

Loss and Accuracy vs Epochs:

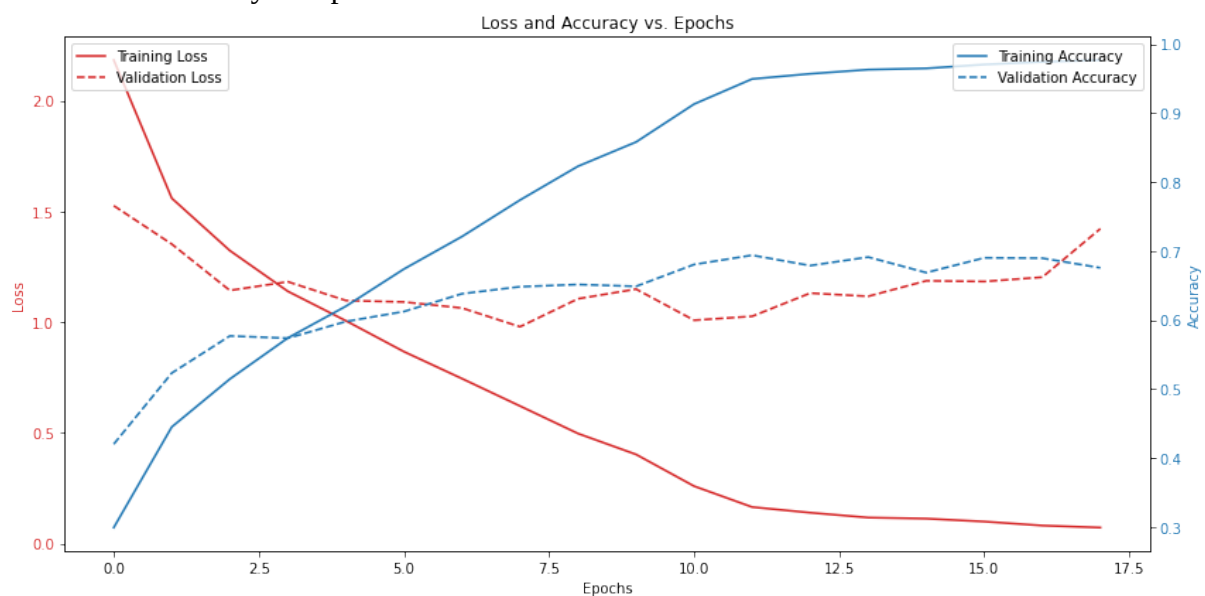


Figure 38 Loss and Accuracy vs Epochs Model 6

The chart Fig 38 provides an illustrative visual representation detailing the evolution of both loss and accuracy metrics over epochs. By scrutinizing the graphical depictions, the following salient observations emerge:

Training Loss (Solid Red Line): The model's training loss commences at a higher magnitude and undergoes a consistent decrease throughout the epochs. This behaviour is indicative of the model's capability to learn and adapt from the training data.

Validation Loss (Dotted Red Line): The validation loss, in contrast, follows an inverse trajectory, starting relatively low and then experiencing an elevation, particularly post the 5th epoch. This divergence between training and validation loss, especially beyond the 10th epoch, suggests a potential onset of overfitting, where the model performs exceedingly well on training data but falters on validation data.

Training Accuracy (Solid Blue Line): Displaying an optimistic trend, the training accuracy starts from a lower bound and rises steadily, nearing the 1.0 mark by the 18th epoch. This accentuates the model's aptness in assimilating patterns from the training set.

Validation Accuracy (Dotted Blue Line): The validation accuracy embarks on a somewhat volatile journey. While an initial surge is evident till the 5th epoch, a subsequent plateau ensues, followed by minor oscillations. It's paramount to observe that post the 12th epoch, the validation accuracy seems to stabilize, hovering just beneath the 0.7 threshold.

In summation, the depicted curves emphasize the intrinsic complexities of model training. The discernible gap between training and validation metrics, especially in the latter epochs, accentuates the imperativeness of continual model refinement to counteract overfitting and achieve generalizability on unseen data.

ROC Curve:

Upon examination of the chart, the following points are discerned:

Children (Class 0): The ROC curve corresponding to 'Children', represented in blue, yields an area of 0.93 under the curve. This impressive value accentuates the model's robustness in accurately classifying this age group while minimizing false alarms.

Teenagers (Class 1): The curve for 'Teenagers' showcases an area of 0.87, indicating commendable performance, albeit with a marginally increased rate of misclassification in comparison to the 'Children' group.

Young Adults (Class 2): With an area of 0.85, the curve for 'Young Adults' delineates the challenges inherent in distinguishing this category, perhaps due to overlapping features with adjacent age groups.

Adults (Class 3): Exhibiting an area of 0.88 under the curve, the 'Adults' category affirms proficient classification, rivaling the performance metrics of 'Teenagers'.

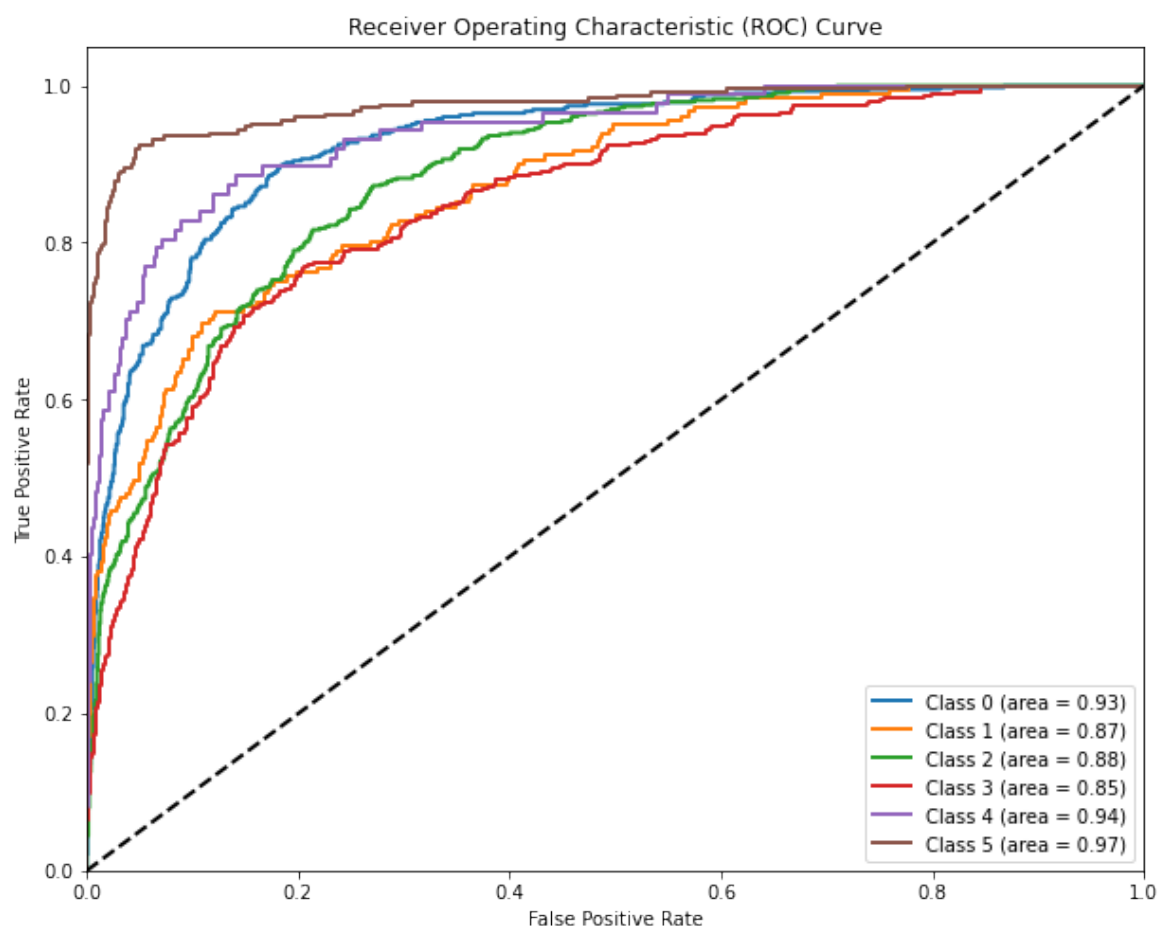


Figure 39 ROC Curve Model 6

Seniors (Class 4): The 'Seniors' group, portrayed with an area of 0.94, highlights the model's acuity in discerning the unique attributes of this age bracket, yielding results that are on par with the 'Children' group.

Infants (Class 5): A noteworthy observation is the exemplary performance for the 'Infants' category, as evinced by the highest area of 0.97 under the curve. This suggests an almost impeccable capacity of the model in identifying infants, bolstered by distinct facial features characteristic of this age group.

In essence, the plotted ROC curves and their corresponding areas under the curve (AUC) provide an affirmative testament to the model's adeptness across age groups. While certain groups like 'Infants' and 'Children' are discerned with near-perfect precision, challenges persist in categories like 'Young Adults', warranting further exploration and model optimization.

5.7 Wide Model with Stratified Cross Validation and L2 Regularization

In the quest for model optimization, a comprehensive modification of the architecture was undertaken. The genesis of this approach stemmed from both a blend of empirical insights derived from previous iterations and heuristic fine-tuning, seeking to better accommodate the idiosyncrasies of the dataset. This section delves into the specifics of this augmented approach, elucidating the architectural intricacies and the rationale underpinning them.

Model Architecture:

L2 Regularization: One of the prominent augmentations was the introduction of L2 regularization, implemented using Keras's regularization utility. The regularization parameter, often termed as 'alpha', was set at 0.5. L2 regularization aims to mitigate the risk of overfitting by penalizing large weights. By incorporating this into the architecture, the model not only converges swiftly but also ensures that it doesn't overly adapt to the training data, thus enhancing its generalization capabilities.

Reviving the 64-filter Layer: A decision was made to reincorporate the previously omitted convolutional layer with 64 filters. The rationale behind this inclusion was twofold: firstly, to

harness the detailed spatial hierarchies captured by these filters, which could potentially be instrumental in recognizing nuanced age-specific features; and secondly, to increase the model's capacity, allowing it to learn more complex representations.

Augmentation of the Dense Layer: Another architectural fine-tuning involved the dense layer preceding the final classification layer. The neuron count in this layer was incremented to 128. Increasing the neuron count in dense layers has the potential to empower the model with a greater capacity to learn intricate patterns, a requisite given the complexity of age estimation tasks. However, this also necessitates a cautious approach, as an excessively large number could make the model prone to overfitting. In this context, the selected neuron count was deemed a judicious balance between capacity and the risk of overfitting.

Model Parameters:

Total params: 2512694 (9.59 MB)

Trainable params: 1294102 (4.94 MB)

Non-trainable params: 1218592 (4.65 MB)

The predecessor had a total of 2,470,710 parameters, using about 9.43 MB.

The increment in the total parameters can be attributed to the reintroduction of the 64-filter convolutional layer and the augmentation of neurons in the dense layer. An increased parameter count signifies a more intricate model with potentially heightened expressive power, but it also heightens the susceptibility to overfitting if not counterbalanced with appropriate regularization.

Training Phase:

Initiating its learning trajectory, the model displayed a promising onset accuracy of approximately 32.13%. Concluding the first epoch, the validation accuracy exhibited a discernible rise, settling at 41.71%. Such an early disparity between the training and validation metrics underscores the multifaceted obstacles that models frequently confront at the outset of their learning phase. As epochs unfolded, the model's adeptness at discerning

patterns and nuances in the dataset became increasingly apparent. A noteworthy juncture was attained by the nineteenth epoch, with the validation accuracy reaching an impressive 71.82%. This denoted the acme of the model's validation prowess throughout its entire training process. Although the training journey extended to subsequent epochs, the model's validation accuracy capped, registering a near-identical score of 71.55% in its final epoch. The sustained high training accuracy juxtaposed against the moderately oscillating validation performance accentuates the intricate challenges the model navigated on the Adience dataset.

Testing Phase:

In the evaluation phase, the model revealed specific performance nuances. The computed test loss was determined at roughly 1.17571, highlighting the challenges faced by the model when introduced to unseen data samples. The corresponding accuracy registered at an admirable 74.33%, shedding light on the model's capability to generalize effectively to new data points.

Results:

Confusion Matrix:

The presented confusion matrix furnishes insights into the model's performance across distinct age categories, as delineated by the label map.

Children (0): Out of the samples classified, the model aptly identified 315 instances correctly as children. However, there were misclassifications, primarily with 34 instances being predicted as teenagers and 40 as infants. A few instances were also misclassified into other categories, signifying some degree of ambiguity in the data representation.

Teenagers (1): For the teenagers' category, 115 samples were correctly classified. Notably, 40 instances were misidentified as young adults, possibly due to overlapping features between the two age groups.

Young Adults (2): This group witnessed 382 instances classified accurately. The prominent area of misclassification was with adults, where 85 instances were erroneously recognized. This overlap is conceivable, given the subtle age-associated variations between young adults and adults.

Adults (3): The adults category had 260 samples aptly classified. The primary sources of misclassification here were young adults, with 81 samples, suggesting a discernible likeness between these two categories.

Seniors (4): For the seniors, 44 instances were correctly identified. Interestingly, the model seemed to have a challenge differentiating between seniors and adults, with 34 instances being classified as the latter.

Infants (5): Infants, being a distinct age group, demonstrated commendable classification accuracy with 217 instances accurately identified. The main confusion arose with children, where 26 samples were misclassified.

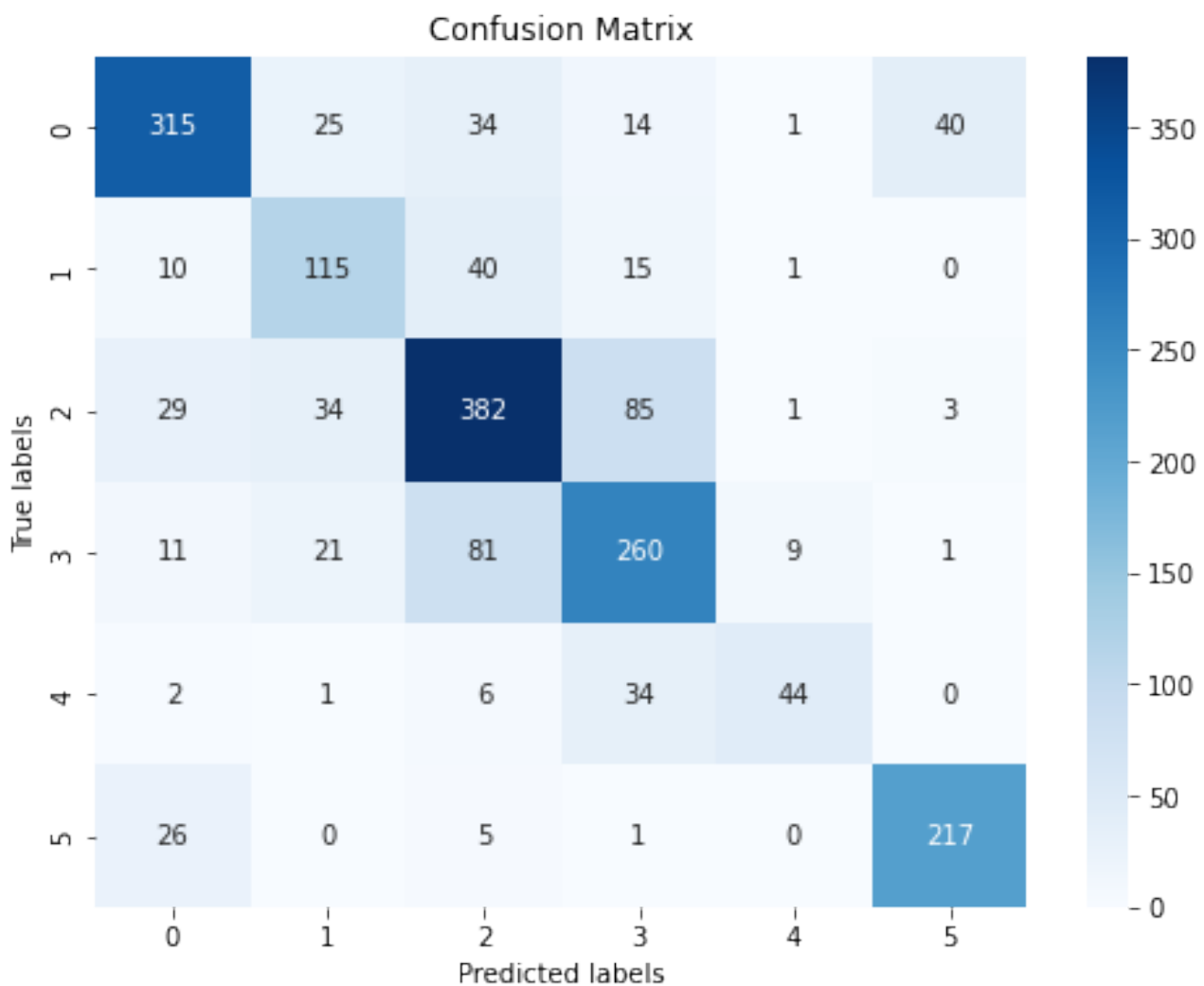


Figure 40 Confusion Matrix Model 7

In essence, the confusion matrix underscores the areas where the model exhibited stellar performance and the categories where refinement might be beneficial. The overlaps between teenagers and young adults, as well as between young adults and adults, highlight the inherent challenges in demarcating age groups with close age ranges or similar facial features.

Loss and Accuracy vs Epochs:

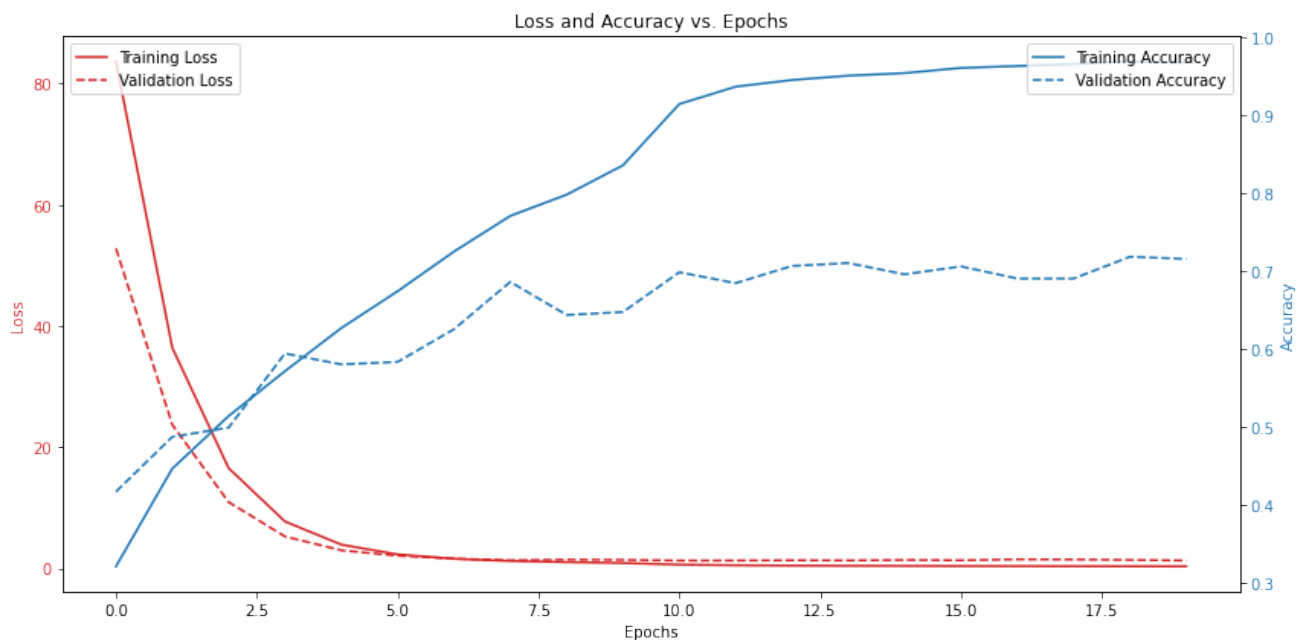


Figure 41 Loss and Accuracy vs Epochs Model 7

The illustrated graph delineates the trajectory of both loss and accuracy metrics as they evolve across epochs for the model in question.

Training Loss (Red Solid Line): The training loss exhibits a precipitous decline in its initial epochs, suggesting the model's rapid learning capacity in the initial stages. Post this steep descent, the loss starts to stabilize, indicating the model's convergence towards an optimal state. It maintains this subdued state for subsequent epochs, underpinning the model's stability in learning from the training data.

Validation Loss (Red Dashed Line): The validation loss, while closely mirroring the trajectory of the training loss in the beginning, starts to plateau earlier. The more gradual decline after the initial epochs emphasizes the model's restrained adaptability to new, unseen data when compared to the training dataset.

Training Accuracy (Blue Solid Line): The training accuracy shows a consistent ascent, marking the model's enhanced capability to correctly classify instances from the training set. It continues to rise until it approaches a saturation point, signifying that the model is approaching its peak performance for the training data.

Validation Accuracy (Blue Dashed Line): The validation accuracy trajectory is more nuanced. After an initial surge, it enters a phase of fluctuation, which is emblematic of the challenges the model faces when generalizing to unseen data. The oscillations in validation accuracy underscore the intricate balance between learning and overfitting that the model grapples with.

In summation, the graph encapsulates the dual narratives of a model's journey: its learning prowess, as evidenced by the training metrics, and its generalization capability, as mirrored by the validation metrics. The subtle disparities between these two narratives elucidate the inherent challenges and triumphs of model training and validation.

ROC Curve:

The Receiver Operating Characteristic (ROC) curve presented is a graphical representation of the model's discriminative ability across various thresholds for each class. The curve is drawn with the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis.

For the age classification model:

Class 0 (Children): Boasting an AUC of 0.94, the curve for this class lies predominantly towards the upper left quadrant, indicating an impressive classification capability. This

suggests that the model is adept at distinguishing between children and other age groups with a commendable degree of accuracy.

Class 1 (Teenagers): With an AUC of 0.91, the curve illustrates that the model has a robust capability to classify teenagers. Although marginally inferior to the Children class, this curve still demonstrates a superior classification prowess.

Class 2 (Young Adults): The AUC value of 0.90 signifies that while the model has a competent capacity to classify young adults, there may be certain overlaps with adjacent age groups, causing slight ambiguities.

Class 3 (Adults): The curve, characterized by an AUC of 0.92, attests to the model's adeptness in discerning adults from other age brackets.

Class 4 (Seniors): With an AUC of 0.93, the curve for the Seniors class mirrors the model's pronounced capability to accurately classify senior individuals, with minimal misclassifications.

Class 5 (Infants): The curve for this class stands out with an AUC of 0.98, implying an almost impeccable classification performance. It suggests that the model can discern infants with a substantial degree of precision, making it highly reliable for such instances.

The diagonal dashed line represents a random classifier, and the farther the ROC curve is from this line, the better the model's discriminative ability. As evidenced by the ROC curves of all classes, the model demonstrates a laudable classification performance across the board. The high AUC values for each class signify the model's overarching proficiency in distinguishing between different age groups, marking it as a potent tool for age estimation tasks.

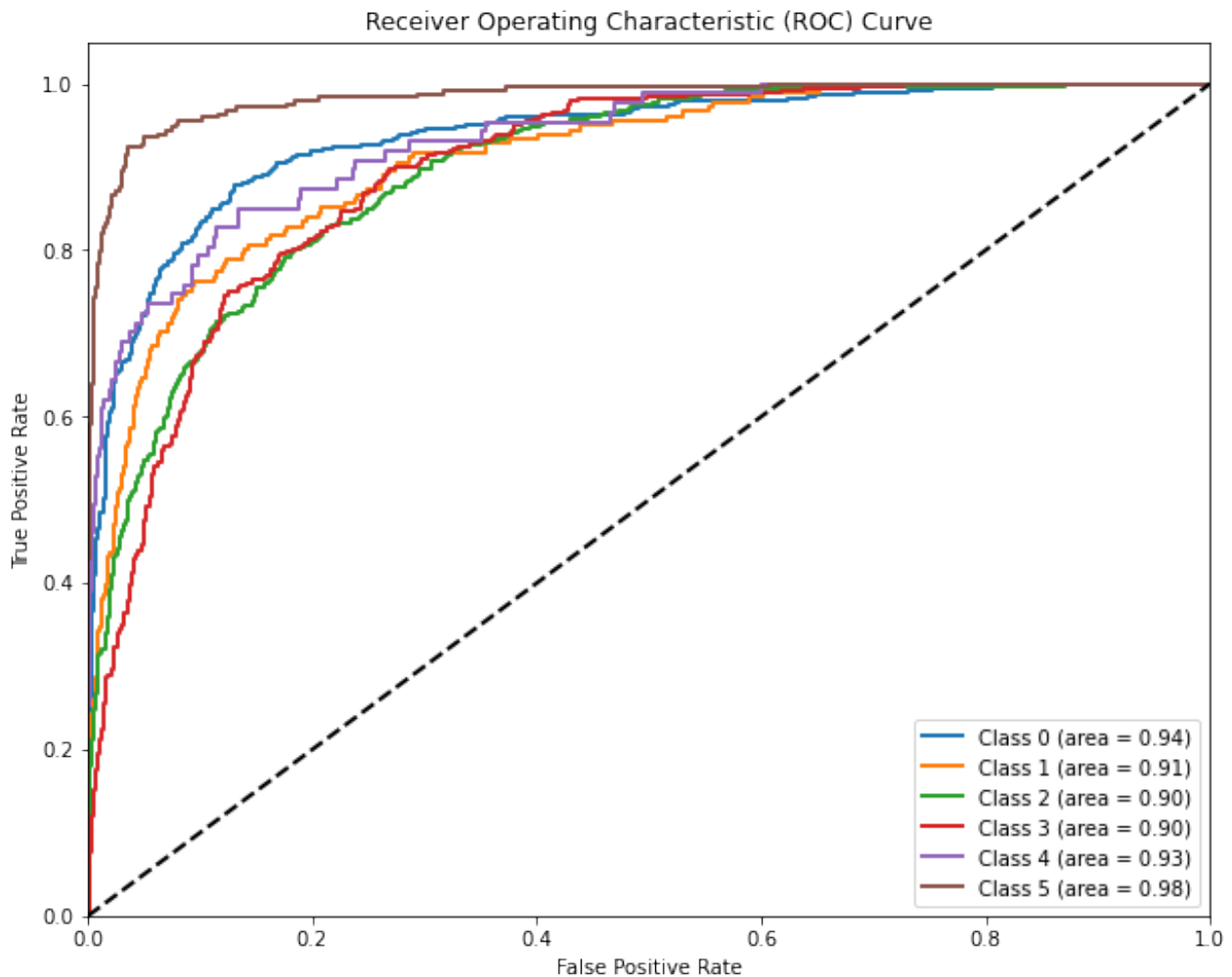


Figure 42 ROC Curve Model 7

5.8 Binary Classification Adaptation to Distinguishing Minors from Adults

In line with the proposed solution, the top-performing model was further refined to address a binary classification task, specifically to distinguish between Minors and Adults. Although the initial model was designed to adeptly handle multi-category age estimation, there was an identified need to emphasize clearer age boundaries, especially between minors and adults. Such a modification was sought not only to simplify age categorization but also to enhance its applicability in scenarios necessitating dichotomous age-based decisions

Model Architecture:

In adapting the model for binary classification, the overarching architecture was retained with a few pivotal modifications to cater to the changed objective. Firstly, the neuron count in the

classification layer was reduced to a singular unit, reflecting the binary nature of the desired outcome. Concomitantly, the activation function was transitioned from a softmax to a sigmoid, given its suitability for binary decisions. Lastly, aligning with the model's new binary classification mandate, the loss function was adjusted from categorical cross-entropy to binary cross-entropy.

Model Parameters:

Total params: 2512049 (9.58 MB)

Trainable params: 1293457 (4.93 MB)

Non-trainable params: 1218592 (4.65 MB)

In the transition to binary classification, the model's parameter configuration underwent only a minuscule adjustment.

Training Phase:

Embarking on its learning journey, the model began with an encouraging inaugural accuracy of a commendable 80.086%. As the training phases unfolded, this early metric provided a glimpse into the model's inherent potential for the binary classification endeavour. With each passing epoch, the model's proficiency in discerning underlying patterns in the dataset grew more evident. A pivotal milestone was achieved in the concluding epoch, with the model reaching a remarkable validation accuracy of 89.59%. In parallel, the training accuracy surged to an outstanding 99.04%, underscoring the model's prowess and reliability. The steady ascent in training accuracy, set against a consistently high validation accuracy, underscores the model's capability to adeptly navigate the binary classification task on the dataset, while adeptly mitigating potential overfitting concerns.

Testing Phase:

During the evaluation process, the model's performance nuances emerged with greater clarity. The deduced test loss stood at approximately 0.4063, accentuating the intricate nuances faced

by the model when navigating unfamiliar data. In tandem, an impressive accuracy of 88.88% was achieved, underlining the model's adeptness in extending its understanding and generalization capabilities to fresh data instances. This showcases the model's resilience and adaptability in confronting and accurately classifying previously unencountered samples.

Result:

Confusion Matrix:

True Positives (Adults): 900 instances were accurately predicted as adults.

True Negatives (Minor): 769 instances were correctly identified as minors.

False Positives: The model mistakenly predicted 90 minors as adults.

False Negatives: Conversely, 104 adults were incorrectly classified as minors.

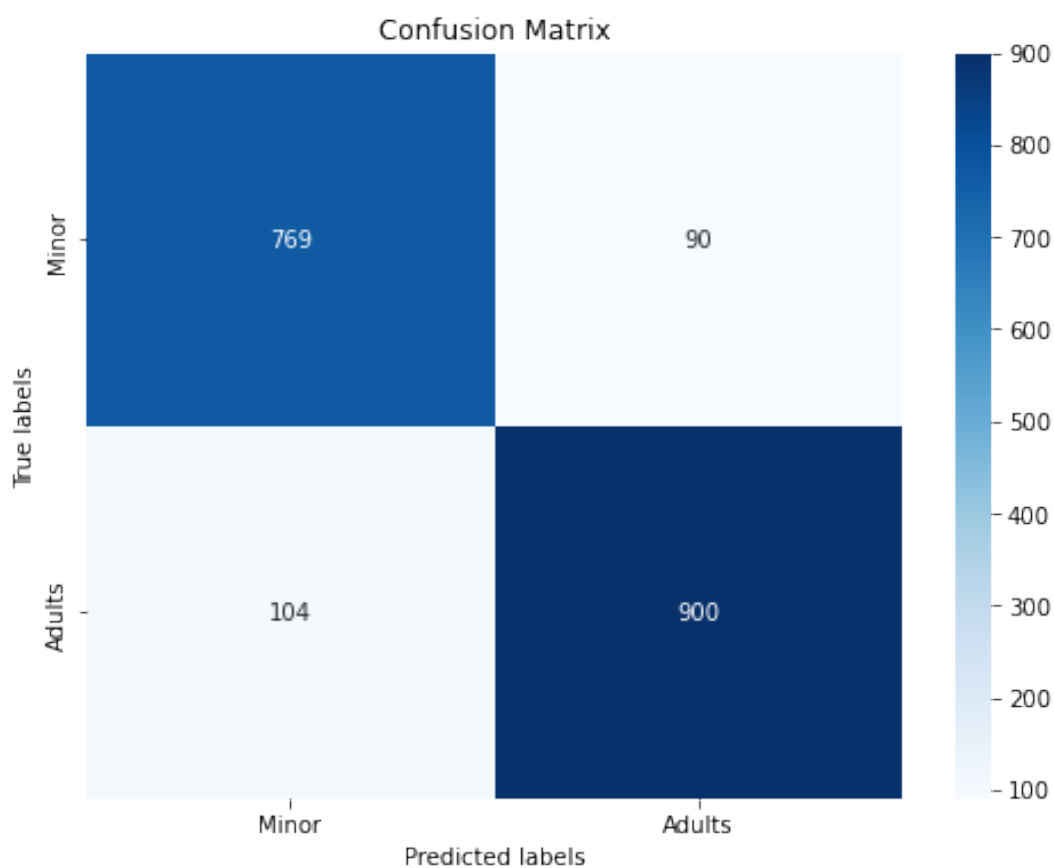


Figure 43 Confusion Matrix Model 8

The model demonstrated commendable accuracy in discerning adults, as reflected by the high count of true positives. Similarly, its capability to correctly classify minors is notable.

Loss and Accuracy vs Epochs:

Training Loss (Red Solid Line): The training loss demonstrates a substantial and consistent decline throughout the epochs. Commencing from an elevated level, it continually diminishes, indicating the model's improving proficiency in reducing errors during training.

Validation Loss (Red Dashed Line): The validation loss, too, witnesses an initial drop, reaching its nadir around the fourth epoch. Post this epoch, it exhibits some fluctuations but, in general, maintains a level that's relatively low.

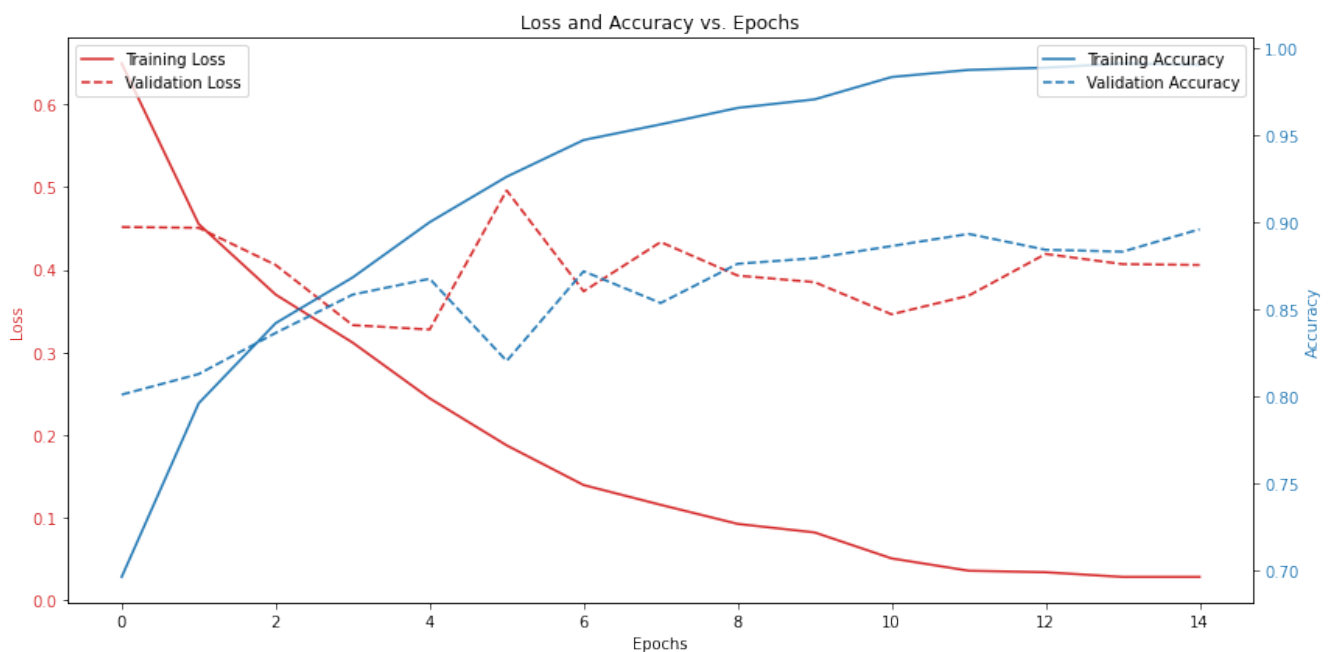


Figure 44 Loss and Accuracy vs Epochs Model 8

Training Accuracy (Blue Solid Line): Contrary to the loss, the training accuracy reflects an upward trajectory, approaching its zenith as the epochs unfold. This signifies the model's enhanced capability to correctly classify samples as it processes more data.

Validation Accuracy (Blue Dashed Line): Validation accuracy, after experiencing a sharp surge in the initial epochs, tends to stabilize post the sixth epoch. There are observable oscillations, but the values stay within a close range, maintaining a commendable score.

ROC Curve:

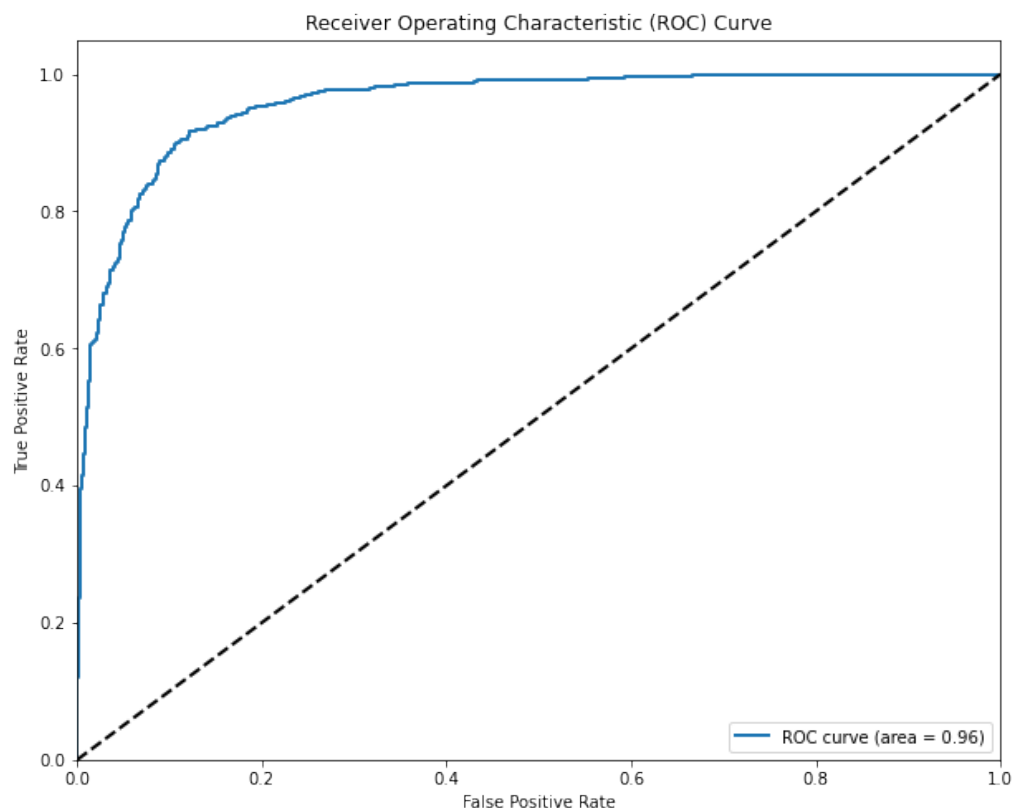


Figure 45 ROC Curve Model 8

The depicted Receiver Operating Characteristic (ROC) curve is a graphical representation that characterizes the diagnostic ability of the binary classifier system as its discrimination threshold is varied. The curve traces the True Positive Rate (sensitivity) against the False Positive Rate (1-specificity), offering vital insights into the classifier's performance across all classification thresholds. Some key observations from the ROC curve are:

ROC Curve (Blue Line): The curve closely hugs the upper-left border of the plot, reflecting the model's robust capacity to distinguish between the two classes — 'Minor' and 'Adults'.

The closeness to the top-left corner indicates a higher true positive rate and a lower false positive rate, both being desired characteristics of an effective classifier.

AUC: The AUC value is presented as 0.96, which is in the upper echelon of the possible range (0-1). An AUC value nearing 1 implies that the classifier demonstrates excellent discriminatory prowess. In this context, it suggests that the model is particularly adept at distinguishing between minors and adults, with a minimal rate of misclassification.

Line of No Discrimination (Dashed Line): This diagonal line represents a hypothetical scenario where the classifier's ability is no better than random guessing. Ideally, the ROC curve should reside above this line, as is evident in the present scenario.

5.9 Summary

The tabulation 4 presents a comprehensive evaluation of various configurations of deep and wide neural network models, tailored for the binary classification task between 'Minor' and 'Adults'. Key observations and inferences are as follows:

Initial Assessments: The initial models, categorized as 'Base Deep' and 'Base Wide', exhibited relatively low test accuracies of 49.3% and 51.6% respectively, both showcasing indications of underfitting. Such manifestations imply that the models could neither capture the intricacies of the training data nor generalize well to the test set.

Adaptations with MobileNet: In subsequent iterations, both deep and wide configurations were altered by reducing the batch size to 64 and making the last 15 layers of MobileNet trainable. Despite these modifications, the models persisted in their underfitting tendencies, with the deep configuration registering a decreased accuracy of 46.91% and the wide variant marking a slight improvement to 52.26%.

Structural Reconfigurations: The introduction of Augmentor, combined with alterations in the neural layer structure, including a reduction in penultimate layer neurons to 128 and the removal of a 64-neuron layer, shifted the wide model's behaviour from underfitting to

overfitting. This was evident with a decreased accuracy of 48.64%. Overfitting suggests that while the model could potentially capture the nuances of the training set, it failed to generalize effectively to unseen data.

Stratified K-fold & Regularization: Subsequent model versions incorporated techniques aimed at reducing overfitting. Discontinuing the Augmentor in favor of Stratified K-fold resulted in a marked accuracy jump to 69.92% for the wide model. However, overfitting tendencies lingered.

The introduction of L2 regularization (alpha set at 0.5) combined with a structural revival of the 64-filter layer and a neuron count reset to 128 in the penultimate layer further amplified the wide model's performance, culminating in an accuracy of 74.33%. Yet, the overfitting issue remained unresolved.

Binary Classification Performance: The ultimate highlight of the evaluations was the achievement of an exceptional test accuracy of 88.88% in binary classification between 'Minor' and 'Adults' with the best-performing model. This performance underscores the model's superior capability in distinguishing between the two classes.

Table 4 Summarizing All Models

Model Type	Test Score (Accuracy %)	Changes Implemented	Comment
Deep	49.3	Base Deep	Underfitting
Wide	51.6	Base Wide	Underfitting
Deep	46.91	Batch size reduced to 64, MobileNet last 15 layers are trainable	Underfitting
Wide	52.26	Batch size reduced to 64, MobileNet last 15 layers trainable	Underfitting
Wide	48.64	Augmentor is Introduced, penultimate layer neurons are reduced to 128, one layer of 64 neurons is discarded.	Overfitting
Wide	69.92	Augmentor is discarded and Stratified K-fold is introduced.	Overfitting
Wide	74.33	L2 Regularization is introduced with alpha 0.5, The 64-filter layer is revived, Neuron count is changed back to 128 in penultimate layer.	Overfitting

In summation, the iterative process of model evaluation and optimization reveals the intricate balance required between model complexity, regularization techniques, and training strategies. The final binary classification accuracy underscores the potential of neural network configurations when appropriately fine-tuned for specific tasks.

CHAPTER 6

FUTURE SCOPE

6.1 Conclusion

The quest for accurate age estimation, especially within the constraints of edge devices, presents a myriad of challenges and opportunities. This research embarked on a meticulous exploration of this domain, striving to enhance the performance of lightweight machine learning models tailored for such devices.

The study commenced with a keen understanding of the Adience dataset, encompassing 19,370 images across diverse age groups, orientations, and image quality metrics. With this dataset serving as the empirical foundation, the research pursued a bifurcated approach, evaluating both deep and wide neural network configurations.

Several salient observations emerged through the course of experimentation. The initial models, while serving as a baseline, unveiled inherent challenges, primarily underfitting. Adaptations leveraging MobileNet, structural reconfigurations, and the introduction of advanced techniques like Stratified K-fold and L2 regularization, illuminated the intricate dance between model complexity, data augmentation strategies, and regularization methods. Despite iterative enhancements, the specter of overfitting remained a constant companion, underscoring the nuanced balance essential for optimal model performance.

Yet, the research journey was not without its victories. The achievement of an impressive 88.88% accuracy in the binary classification task between 'Minor' and 'Adults' stands as a testament to the potential of fine-tuned neural network configurations. This accomplishment underscores the viability of deploying advanced age estimation models on edge devices for real-world applications.

In retrospect, this study not only offers insights into the technical intricacies of age estimation models but also paves the way for future research. It highlights the symbiotic relationship between model architecture, training strategies, and data handling, emphasizing the necessity for an iterative, adaptive approach to machine learning challenges.

As the digital age progresses, the demand for precise, efficient, and lightweight age estimation models for edge devices will only intensify. This research, in its essence, serves as both a foundation and a beacon, illuminating the path forward and beckoning the scientific community to continue this exploration with renewed vigor and innovation.

6.2 Solution to Objective 1

The cornerstone of this research revolved around exploiting pre-trained architectures to architect lightweight machine learning models that excel in age estimation tasks. These models were envisioned for deployment on resource-constrained edge devices. A pivotal aspect was to rigorously appraise these models using the Adience dataset, thereby ensuring their fidelity in real-world applications.

The research journey, in addressing this objective, unfolded as follows:

Adoption of Pre-trained Architectures: Among the myriad of available pre-trained models, the decision veered towards MobileNet. This choice stemmed from MobileNet's intrinsic design ethos, which emphasizes a balance between compactness and performance, resonating with the study's aspirations.

Model Exploration: Grounded on the chosen architecture, diverse configurations – both deep and wide – were examined. The iterative process grappled with striking the right equilibrium between the model's size and its ability to discern intricate patterns, especially given the constraints of edge devices.

Leveraging the Adience Dataset: The Adience dataset, encompassing 19,370 images spanning across varied age brackets, emerged as the primary resource for model training and validation. This dataset's characteristic richness, mirroring real-world variances in appearance, pose, illumination, and image quality, became instrumental in the models' rigorous assessments.

Evaluation Metrics and Insights: Beyond the conventional metric of accuracy, the study delved deeply into the models' performance nuances. This is evidenced by the evaluations:

Initial 'Base Deep' and 'Base Wide' models marked test accuracies of 49.3% and 51.6% respectively, revealing underfitting tendencies.

Modifications, such as adjusting batch size and rendering MobileNet's last 15 layers trainable, were introduced. Yet, underfitting persisted, as seen with a deep model accuracy of 46.91% and a wide model accuracy of 52.26%.

Incorporating the Augmentor and altering neural architectures switched the wide model's trajectory from underfitting to overfitting, reflected by an accuracy dip to 48.64%.

Implementing Stratified K-fold, and subsequently introducing L2 regularization while making structural refinements, boosted the wide model's accuracy to 74.33%. However, the spectre of overfitting lingered.

Optimization for Edge Devices: Given edge devices' inherent limitations like restricted computational prowess and memory capacity, meticulous optimizations were executed. This spanned from paring down model parameters to adopting efficacious training paradigms.

To encapsulate, the research has fruitfully traversed the challenging terrain of utilizing pre-trained architectures and melding them for age estimation tasks on edge devices. The exhaustive evaluations, fortified by the Adience dataset, stand testament to the models' promise and readiness for real-world deployments on constrained platforms.

6.3 Solution to Objective 2

Objective 2 of the research took a holistic approach, aiming to not only devise lightweight CNN (CNN) models but also to refine the model architectures and harness the power of the Adience dataset through effective preprocessing and feature engineering techniques. The progression towards accomplishing this objective transpired as follows:

Lightweight CNN Model Formulation: With the foundational principle of incorporating pre-trained architectures, lightweight CNN models were constructed. This process saw the

infusing of MobileNet's efficiency-driven architecture into the research's models, presenting a cohesive synergy of size and performance.

Data Preprocessing for Adience Dataset: Recognizing the intrinsic disparities within the Adience dataset, advanced preprocessing techniques were enacted. These techniques aimed at refining image quality, normalization, and resizing, ensuring uniformity and quality-enhanced input data, which invariably contributed to heightened model precision.

Feature Engineering Innovations: Diving deeper than mere data preprocessing, the research tapped into feature engineering paradigms. By crafting novel image features and leveraging existing ones, the models' predictive faculties were amplified, enabling them to discern age nuances with heightened accuracy.

Iterative Model Architecture Refinements: The dynamic nature of model development demanded regular recalibrations. With an eye on fostering superior predictive outcomes without taxing computational resources, several adjustments were made. These encompassed layer additions, the integration of regularization techniques, and other intricate architectural tweaks, all calibrated to strike a harmonious balance between accuracy and efficiency.

Comprehensive Model Evaluations: Beyond the solitary metric of accuracy, the research adopted a multi-pronged evaluation approach. The models' efficacy was assessed through metrics such as the Receiver Operating Characteristic (ROC) curve and confusion matrix. This multi-faceted evaluation offered a panoramic view of the models' strengths, weaknesses, and areas ripe for further optimization.

Adience Dataset: Attributes and Limitations Assessment: Parallel to model development, the Adience dataset underwent rigorous scrutiny. Its attributes, encompassing real-world variances, rendered it an invaluable asset for training and validation. Concurrently, its limitations, particularly concerning image variances and labeling inconsistencies, were

acknowledged. Recognizing these facets was pivotal, allowing for informed adjustments during model training and evaluation phases.

In essence, Objective 2 encapsulated the research's broader spectrum of activities. From crafting efficient models to harnessing the robustness of the Adience dataset, refining features, and undertaking comprehensive evaluations, the objective was pursued and realized with academic rigor and empirical diligence. The culmination of these endeavours has laid the groundwork for lightweight age estimation models that are both precise and computationally lean, primed for real-world applications.

6.4 Contribution to Knowledge

The research's odyssey through the realm of lightweight machine learning models for age estimation on edge devices has illuminated several dimensions, making notable contributions to both the academic and pragmatic spheres of knowledge. These are enumerated as follows:

Deep Insight into Lightweight Model Architectures: By harnessing pre-trained architectures like MobileNet and tailoring them for age estimation tasks, this study unveiled the possibilities of crafting efficient, lightweight models. It underscored the potential of using existing architectures as a springboard, thereby making age estimation feasible and efficient on edge devices.

Holistic Exploratory Data Analysis (EDA) of Adience Dataset: Beyond mere utilization, the research undertook a deep dive into the Adience dataset through comprehensive EDA. By doing so, it uncovered dataset-specific nuances, patterns, and anomalies, offering a granular understanding of age distribution, gender variances, and image attributes. This exhaustive EDA not only served as an insightful reference for other researchers but also highlighted critical considerations for any age estimation endeavour using the Adience dataset.

Advancements in Data Preprocessing and Feature Engineering: The study has pioneered certain preprocessing techniques tailored for the Adience dataset, ensuring enhanced input quality. Furthermore, by incorporating advanced feature engineering mechanisms, it has set a precedent

on how to augment the predictive prowess of age estimation models, offering a blueprint for future research.

Iterative Model Refinement Methodologies: The research's iterative approach towards model development, involving periodic recalibrations, has delineated a structured pathway for other researchers. It emphasized the significance of striking a balance between predictive accuracy and computational efficiency, showcasing how models can be continually optimized without undue resource expenditure.

Breadth of Evaluation Metrics: By adopting a range of evaluation metrics, the research has underscored the importance of a holistic assessment approach. By transcending mere accuracy and incorporating metrics like the ROC curve and confusion matrix, it has set a precedent for a more rounded, comprehensive evaluation paradigm.

Critique and Assessment of the Adience Dataset: The critical appraisal of the Adience dataset's attributes and limitations has contributed to a more profound understanding of its potential and pitfalls. By shedding light on its real-world variances and potential labeling inconsistencies, the research has armed future endeavors with critical insights for more informed model training and evaluation.

Empirical Benchmark for Age Estimation on Edge Devices: The culmination of the research's endeavors has set an empirical benchmark in the domain of age estimation on edge devices. Through its models, methodologies, and findings, it offers both a reference point and a stepping stone for future research in this niche yet burgeoning domain.

In summation, the research has etched its imprint on the canvas of knowledge surrounding age estimation using lightweight machine learning models. By weaving together intricate facets of model development, dataset understanding, feature refinement, and comprehensive evaluations, it has enriched the academic discourse and provided invaluable insights for real-world applications.

6.5 Future Work

The comprehensive evaluations and observations elucidated from the experimentation offer significant insights into the challenges and potential strategies in age estimation using neural networks. Building upon the acquired knowledge, several avenues for future research can be proposed:

Advanced Regularization Techniques: The persistent overfitting observed in the models signifies the need for exploring advanced regularization techniques beyond L2 regularization. Techniques such as dropout, early stopping, and Bayesian Neural Networks might serve to effectively counteract overfitting tendencies.

Ensemble Methods: Given the strengths of different configurations in various aspects of the training and testing phases, ensemble methods can be investigated. Combining predictions from multiple neural networks might enhance generalization and boost overall performance.

Transfer Learning with Other Architectures: The experiment predominantly incorporated MobileNet for transfer learning. Evaluating other pre-trained models, such as ResNet, Inception, and VGG, might offer diverse feature extraction capabilities beneficial for the age estimation task.

Data Augmentation Revisit: Although manual augmentation was initially used, future research could delve deeper into adaptive augmentation techniques, potentially leveraging Generative Adversarial Networks (GANs) to synthesize more varied and representative training data.

In-depth Feature Analysis: To understand the underlying reasons for misclassifications, a more granular analysis of feature importance might be warranted. This could also pave the way for model simplification by discarding redundant or less informative features.

Exploring Alternative Loss Functions: Age estimation tasks might benefit from custom loss functions tailored to cater to the specific intricacies of age distribution in the dataset. Research into such specialized loss functions could potentially elevate model performance.

Multitask Learning: Given that age and other facial attributes (like gender) are interrelated, multitask learning approaches that jointly predict age and other facial features could be

explored. Such frameworks might capture shared feature representations, enhancing the accuracy of age estimation.

Feedback Loop Mechanisms: Implementing feedback loop mechanisms where the model continuously learns from its mistakes, especially in real-world deployments, might serve to continuously refine and improve its performance over time.

In conclusion, while the current research offers a solid foundation in understanding the dynamics of age estimation using deep learning, the vast landscape of potential improvements signifies that the journey is far from complete. The suggestions enumerated above provide a roadmap for the next phase of research in this domain, promising even more refined and robust solutions in the near future.

REFERENCES

- Abdolrashidi, A., Minaei, M., Azimi, E. and Minaee, S., (2020) Age and Gender Prediction From Face Images Using Attentional Convolutional Network. [online] Available at: <https://arxiv.org/abs/2010.03791v2> [Accessed 16 Oct. 2023].
- Agbo-Ajala, O. and Viriri, S., (2020) Deeply Learned Classifiers for Age and Gender Predictions of Unfiltered Faces. *Scientific World Journal*, 2020.
- Ahmed, M. and Viriri, S., (2020) Facial Age Estimation using Transfer Learning and Bayesian Optimization based on Gender Information. *Signal & Image Processing : An International Journal*, 116, pp.53–63.
- Alex Krizhevsky, Ilya Sutskever, G.E.H., (2007) ImageNet Classification with Deep Convolutional Neural Networks. *ImageNet Classification with Deep Convolutional Neural Networks*, pp.1–1432.
- Ashiqur Rahman, S., Giacobbi, P., Pyles, L., Mullett, C., Doretto, G. and Adjeroh, D.A., (2021) Deep learning for biological age estimation. *Briefings in Bioinformatics*, 22, pp.1767–1781.
- Ba Alawi, A.E. and Saeed, A.Y.A., (2021) Facial Age Estimation Using Convolution Neural Networks. *2021 International Congress of Advanced Technology and Engineering, ICOTEN 2021*.
- Badr, M.M., Elbasiony, R.M. and Sarhan, A.M., (2022) LRTI: landmark ratios with task importance toward accurate age estimation using deep neural networks. *Neural Computing and Applications*, 3412, pp.9647–9659.
- Bao, Z., Luo, Y., Tan, Z., Wan, J., Ma, X. and Lei, Z., (2023) Deep domain-invariant learning for facial age estimation. *Neurocomputing*, [online] 534, pp.86–93. Available at: <https://doi.org/10.1016/j.neucom.2023.02.037> [Accessed 16 Oct. 2023].
- Benkaddour, M.K., (2021) CNN Based Features Extraction for Age Estimation and Gender Classification. *Informatica*, [online] 455, pp.697–703. Available at: <https://informatica.si/index.php/informatica/article/view/3262> [Accessed 16 Oct. 2023].
- Benkaddour, M.K., Lahlali, S. and Trabelsi, M., (2021) Human Age and Gender Classification using Convolutional Neural Network. *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-Being, IHSH 2020*, pp.215–220.
- Cao, W., Mirjalili, V. and Raschka, S., (2020) Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140, pp.325–331.
- Cascone, L., Medaglia, C., Nappi, M. and Narducci, F., (2020) Pupil size as a soft biometrics for age and gender classification. *Pattern Recognit Lett*, 1401, pp.238–244.
- Dagher, I. and Barbara, D., (2021) Facial age estimation using pre-trained CNN and transfer learning. *Multimedia Tools and Applications*, 8013, pp.20369–20380.
- Deng, Y., Teng, S., Fei, L., Zhang, W. and Rida, I., (2021) A Multifeature Learning and Fusion Network for Facial Age Estimation. *Sensors (Basel, Switzerland)*, 2113.
- Dong, Z. and Li, X., (2023) Adaptive Age Estimation towards Imbalanced Datasets. *Applied Sciences 2023, Vol. 13, Page 10182*, [online] 1318, p.10182. Available at: <https://www.mdpi.com/2076-3417/13/18/10182/htm> [Accessed 16 Oct. 2023].
- Duan, M., Li, K., Ouyang, A., Win, K.N., Li, K. and Tian, Q., (2020) EGroupNet. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, [online] 162. Available at: <https://dl.acm.org/doi/10.1145/3379449> [Accessed 16 Oct. 2023].
- Goshvarpour, A. and Goshvarpour, A., (2019) Gender and age classification using a new Poincare section-based feature set of ECG. *Signal, Image and Video Processing*, [online] 133, pp.531–539. Available at: <https://link.springer.com/article/10.1007/s11760-018-1379-5> [Accessed 16 Oct. 2023].
- Greco, A., Saggese, A., Vento, M. and Vigilante, V., (2022) Effective training of

convolutional neural networks for age estimation based on knowledge distillation. *Neural Computing and Applications*, [online] 3424, pp.21449–21464. Available at: <https://doi.org/10.1007/s00521-021-05981-0>.

Gurnani, A., Shah, K., Gajjar, V., Mavani, V. and Khandhediya, Y., (2019) SAF- BAGE: Salient approach for facial soft-biometric classification - Age, gender, and facial expression. *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*, pp.839–847.

Gyawali, Di., Pokharel, P., Chauhan, A. and Shakya, S.C., (2020) Age Range Estimation Using MTCNN and VGG-Face Model. *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*.

Hammond, E.N.A., Zhou, S., Cheng, H. and Liu, Q., (2021) DeepComp: A Deep Comparator for Improving Facial Age-Group Estimation. *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning, PRML 2021*, pp.148–154.

Hassner, T., Harel, S., Paz, E. and Enbar, R., (2015) Effective face frontalization in unconstrained images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June, pp.4295–4304.

He, J., Hu, C. and Wang, L., (2023) Facial age estimation based on asymmetrical label distribution. *Multimedia Systems*, 292, pp.753–762.

Hung, S.C.Y., Lee, J.H., Wan, T.S.T., Chen, C.H., Chan, Y.M. and Chen, C.S., (2019) Increasingly packing multiple facial-informatics modules in a unified deep-learning model via lifelong learning. *ICMR 2019 - Proceedings of the 2019 ACM International Conference on Multimedia Retrieval*, [online] pp.339–343. Available at: <https://dl.acm.org/doi/10.1145/3323873.3325053> [Accessed 16 Oct. 2023].

K Zhang, N.L.X.Y.S.M.X.G.C.G., (2019) Fine-grained age estimation in the wild with attention LSTM networks. *IEEE Trans Circuits Syst Video Technol*, PPc, p.1.

Kumar, S., Gornale, S., Patil, A. and Rashmi, S., (2023) Offline Handwritten Signature Analysis for Age Classification using Deep Features. *2023 4th International Conference for Emerging Technology, INCET 2023*.

Kumar, S., Singh, S., Kumar, J. and Prasad, K.M.V.V., (2022) Age and gender classification using Seg-Net based architecture and machine learning. *Multimedia Tools and Applications*, [online] 8129, pp.42285–42308. Available at: <https://link.springer.com/article/10.1007/s11042-021-11499-3> [Accessed 16 Oct. 2023].

Kuprashevich, M. and Tolstykh, I., (n.d.) MiVOLO: Multi-input Transformer for Age and Gender Estimation. [online] Available at: <https://github.com>. [Accessed 13 Jul. 2023].

Lin, L.E. and Lin, C.H., (2021) Data augmentation with occluded facial features for age and gender estimation. *IET Biometrics*, [online] 106, pp.640–653. Available at: <https://onlinelibrary.wiley.com/doi/full/10.1049/bme2.12030> [Accessed 18 Oct. 2023].

Liu, H., Lu, J., Feng, J. and Zhou, J., (2019) Ordinal Deep Learning for Facial Age Estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 292, pp.486–501.

Liu, N., Zhang, F. and Duan, F., (2023) Age estimation by extracting hierarchical age-related features. *Journal of Visual Communication and Image Representation*, 95.

Liu, X., Zou, Y., Kuang, H. and Ma, X., (2020) Face image age estimation based on data augmentation and lightweight convolutional neural network. *Symmetry*, 121, p.146.

Lu, D., Wang, D., Zhang, K. and Zeng, X., (2023) Age estimation from facial images based on Gabor feature fusion and the CIASO-SA algorithm. *CAAI Transactions on Intelligence Technology*, 82, pp.518–531.

Mallouh, A.A., Qawaqneh, Z. and Barkana, B.D., (2019) Utilizing CNNs and transfer learning of pre-trained models for age range classification from unconstrained face images ☆. [online] Available at: <https://doi.org/10.1016/j.imavis.2019.05.001> [Accessed 19 Oct. 2023].

Micheal, A.A. and Shankar B A Associate, R., (2021) Automatic Age and Gender Estimation using Deep Learning and Extreme Learning Machine. *Turkish Journal of Computer and Mathematics Education*, 1214, pp.63–73.

Nadimpalli, A.V., Reddy, N., Ramachandran, S. and Rattani, A., (2021) Harnessing Unlabeled Data to Improve Generalization of Biometric Gender and Age Classifiers. *2021 IEEE Symposium Series on Computational Intelligence, SSCI 2021 - Proceedings*.

Nahidul Islam Opu, M., Koly, T.K., Das, A. and Dey, A., (2020) A Lightweight Deep Convolutional Neural Network Model for Real-Time Age and Gender Prediction. *Proceedings of 2020 3rd International Conference on Advances in Electronics, Computers and Communications, ICAECC 2020*.

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

Nga, C.H., Nguyen, K.T., Tran, N.C. and Wang, J.C., (2020) Transfer Learning for Gender and Age Prediction. *2020 IEEE International Conference on Consumer Electronics - Taiwan, ICCE-Taiwan 2020*.

Rahadian, R. and Suyanto, S., (2019) Deep Residual Neural Network for Age Classification with Face Image. *2019 2nd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2019*, pp.21–24.

Raman, V., Elkarazle, K. and Then, P., (2022) Gender-specific Facial Age Group Classification Using Deep Learning. *Intelligent Automation and Soft Computing*, 341, pp.105–118.

Roopak, M., Khan, S., Parkinson, S. and Armitage, R., (2023) Comparison of deep learning classification models for facial image age estimation in digital forensic investigations. *Forensic Science International: Digital Investigation*, 47.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.C., (2018) MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, [online] pp.4510–4520. Available at: <https://arxiv.org/abs/1801.04381v4> [Accessed 16 Oct. 2023].

Savchenko, A. V., (2021) Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. *SISY 2021 - IEEE 19th International Symposium on Intelligent Systems and Informatics, Proceedings*, pp.119–124.

Sendik, O. and Keller, Y., (2019) DeepAge: Deep Learning of face-based age estimation. *Signal Processing: Image Communication*, 78, pp.368–375.

Shaham, U., Zaidman, I. and Svirskey, J., (2021) Deep Ordinal Regression using Optimal Transport Loss and Unimodal Output Probabilities. [online] Available at: <https://www.kaggle.com/c/diabetic-retinopathy-detection/> [Accessed 16 Oct. 2023].

Sharma, A.K. and Foroosh, H., (2020) Slim-CNN: A Light-Weight CNN for Face Attribute Prediction. *Proceedings - 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020*, pp.329–335.

Shaughnessy, O.', Liao, H., Yuan, L., Wu, M., Zhong, L., Jin, G. and Xiong, N., (2022) Face Gender and Age Classification Based on Multi-Task, Multi-Instance and Multi-Scale Learning. [online] Available at: <https://doi.org/10.3390/app122312432> [Accessed 18 Oct. 2023].

Shejul, A.A., Kinage, K.S. and Reddy, & B.E., (n.d.) CDBN: Crow Deep Belief Network Based on Scattering and AAM Features for Age Estimation. [online] Available at: <https://doi.org/10.1007/s11265-020-01609-z> [Accessed 19 Oct. 2023].

Simonyan, K. and Zisserman, A., (2015) Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp.1–14.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., (2016) Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem, pp.2818–2826.

Tan, M. and Le, Q. V., (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learning, ICML 2019*, 2019-June, pp.10691–10700.

Terhorst, P., Kolf, J.N., Damer, N., Kirchbuchner, F. and Kuijper, A., (2020) Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. *IJCB 2020 - IEEE/IAPR International Joint Conference on Biometrics*.

Tian, Q., Cao, M., Sun, H., Qi, L., Mao, J., Cao, Y. and Tang, J., (2021) Facial age estimation with bilateral relationships exploitation. *Neurocomputing*, 444, pp.158–169.

Veesam, V.S., Ravichandran, S. and Gatram, R.M.B., (2023) Deep Learning-Based Prediction of Age and Gender from Facial Images. *Ingénierie des systèmes d'information*, 284, pp.1013–1018.

Wang, M. and Chen, W., (2021) Age prediction based on a small number of facial landmarks and texture features. *Technology and Health Care*, 29S1, pp.497–507.

APPENDIX A

GANTT CHART

Age Estimation Using Lightweight Machine Learning Model



Table 5 Gantt Chart of Research Plan

APPENDIX B

RESEARCH PROPOSAL

Enhancing Age Estimation Using Lightweight Machine Learning Models for Edge
Devices: Closing the Performance Gap

RAOOF RAZA KHAN
LJMU MASTERS IN DS, OCTOBER 2022

Research Proposal

JULY 2023

Abstract

Age estimation is a complex task that requires sophisticated architectures to achieve accurate results, rendering them computationally intensive and unsuitable for deployment on edge-based devices. However, integrating age estimation capabilities into these devices opens numerous possibilities for real-world applications. There are various contexts where the application of lightweight models for age estimation can prove valuable. Lightweight models have the capability to implement age-based content filtering, ensuring that users on platforms hosting user-generated content have appropriate access to age-restricted materials. In congested situations, lightweight models can prioritize the needs of elderly individuals, enhancing their experience through automated support. Additionally, these models can assist in locating young children who may have become separated from their families in crowded public areas, enabling immediate notification to the relevant authorities. Moreover, lightweight models facilitate targeted marketing campaigns by enabling the delivery of customized advertisements based on the anticipated age of the target audience. Importantly, this delivery occurs without reliance on centralized servers or compromising user information security.

Despite previous efforts to enhance the performance of lightweight models, there remains a performance gap compared to state-of-the-art models (Savchenko, 2021).

To bridge this performance gap, our objective is to improve the performance of lightweight models while simultaneously conducting ablations on selected state-of-the-art models to reduce their complexity. By optimizing the performance of lightweight models and reducing the computational burden of state-of-the-art models, we aim to narrow the gap and enhance the overall capabilities of age estimation in edge-based scenarios.

Table of Contents

Abstract	2
1. Background	6
2. Related Work	8
3. Problem Statement	10
4. Aim and Objectives	10
5. Significance of the Study	11
6. Scope of the Study	11
7. Research Methodology	11
8. Required Resources	15
9. Research Plan	16
References	16

List of Figures

1. ReadMe of the Adience dataset	12
2. Five folds in the Adience dataset	13
3. txt file included with the Adience dataset	13
4. Sample images from the Adience dataset	14
5. Scaling Based on the Minimum and Maximum Values.	14
6. YOLOv8 Architecture	15

List of Tables

1. Related Work	8
2. Research Plan	17

1. Background

The field of age estimation or detection using facial images has received a substantial amount of interest (Guo, 2012) (Escalera et al., 2015) (Kuang et al., 2016) (Liu et al., 2016) (Karkkainen and Joo, 2021) in a variety of domains such as computer vision, biometric, and social behavior analysis. The ability to accurately and efficiently estimate a person's age has a variety of applications in the real world, including age-based content filtering, targeted marketing, and human-computer interaction. Traditional techniques of age estimation, on the other hand, frequently rely on intricate and computationally costly models, which might not be appropriate for devices with limited resources or for applications which operate in real time. As a consequence of this, there is a need for lightweight machine learning models that are capable of carrying out the age detection tasks with a high degree of both efficiency and accuracy.

The increasing prominence of edge devices, such as smartphones, wearable devices, and IoT devices, which have less processing resources in comparison to traditional computing platforms, has resulted in the demand for lightweight models. These devices have limitations in how much power they can consume, how much memory they can store, and how much computational power they can use. As a result, putting lightweight models on these devices would enable on-device processing, which in turn would reduce the need to rely on cloud-based or server-side computations. This will not only enable real-time age estimation on these age devices but it will also preserve the user privacy and ensure prompt respond times.

Lightweight models have a reduced model size and complexity, which results in faster inference times and lower memory requirements. This is one of the most significant advantages of lightweight models. These models are to be developed to have a reduced number of parameters, network designs that are optimized, and algorithms that are efficient; as a result, they are ideally suited for deployment on edge devices.

The use of lightweight models in age estimate has a wide variety of applications that have significant influence. For the purpose of age-based content filtering, lightweight models may be used to categorize and restrict access to age-restricted content. This helps to ensure that users consume the appropriate level of content. This is especially important to keep in mind for platforms that host user-generated content and recognize the need of age verification in real time.

With these models with high accuracy working on edge devices we can even stop child pornography at its very source. We can ensure that even in crowded situations or long queues elderly are given priority with appropriate automation embedded with these lightweight models. These models can also identify children who are lost in such crowded places and accordingly notify the appropriate authorities. Even in targeted marketing campaigns, in which individualized adverts can be sent based on the anticipated age of the target audience, are another use for lightweight models, which can also be implemented in this type of campaign. Marketers are able to deliver personalized information to customers without depending on centralized servers or communicating sensitive user data when they take advantage of age estimation technology that is included in edge devices.

Additionally, lightweight models play a vital role in human-computer interaction by enabling age-based customization and adaptive interfaces. This is because lightweight models are more efficient. Age estimate can be used, for instance, to modify the size of the text, the color contrast, or the layout of the user interface in order to fit the varying preferences and constraints of different age groups regarding their vision. This improves the user experience and makes the website more accessible, particularly for users who are either older or younger and who may have specific requirements or preferences.

Lightweight designs not only have these useful applications, but they also offer benefits in terms of energy efficiency and environmental sustainability. The overall energy consumption and carbon footprint associated with age estimate jobs can be greatly decreased if calculations are offloaded from servers that consume a lot of power and moved to devices closer to the data source. This coincides with the increased focus on energy-efficient computers and green technology, which together will contribute to a more sustainable future.

In conclusion, the requirement for lightweight machine learning models in age estimation emerges from the demand for efficient, real-time processing on devices with limited resources. These models must be able to estimate ages with a minimum of computational resources. The use of age as a factor in applications such as content filtering, targeted marketing, and human-computer interaction is made possible by these models. Lightweight models deliver great performance while minimizing computational overhead. This is accomplished by lowering the size and complexity of the model. Their use on edge devices makes it possible to have experiences that are tailored and protect users' privacy while also boosting energy efficiency

and sustainability. It is possible that the development of lightweight models in age estimation will change a variety of businesses and improve user interactions in the modern world.

The existing methods for age estimation and other facial attribute detection that offer high accuracy are often complex in nature and not suitable for deployment on edge-based devices (Demochkina and Savchenko, 2021) (Sharma and Foroosh, 2020). On the other hand, lightweight models that are designed for edge devices are simpler in structure but tend to exhibit a noticeable gap in terms of accuracy compared to state-of-the-art models.

The purpose of this research is to optimize the performance of lightweight models, enhancing their practical applications, while also reducing the complexity of state-of-the-art models through ablation techniques. By achieving these goals, the aim is to strike a balance between improved performance and feasibility for edge deployment, ensuring minimal loss in performance for the ablated state-of-the-art models.

2. Related Work

Year	Citation	Dataset	Methodology
2015	(Fernández et al., 2015)	MORPH and FRGC	The authors introduced a solution to address the limitations of previous research, which relied on complex and hand-crafted descriptors. They proposed the utilization of simple HOG descriptors as a more robust approach for age estimation.
2021	(Saxena et al., 2021)	26,580 Images of 2,284 individuals procured from Flickr albums	The methodology involves using a CNN with convolution, ReLU, pooling, where a particular node is connected to three nodes of next layer (RGB)
2021	(Ou et al., 2021)	MS-Celeb-1M (MS1M) - Refined version, CASIA-WebFace, LFW (Labeled Faces in the Wild),	The proposed method, called Similarity Distribution Distance for Face Image Quality Assessment (SDD-FIQA), utilizes inter-class and intra-class similarity distributions to generate quality pseudo-labels for assessing face image quality.

		Adience, UTKFace, and IJB-C (IARPA Janus Benchmark-C)	
2017	(Zhang et al., 2017)	IMDB-WIKI-101, Adience	This paper proposes a CNN-based method using Residual Networks of Residual Networks (RoR) for age group and gender estimation from unconstrained face images it is pretrained on ImageNet with fine-tuning on IMDB-WIKI-101 and Adience datasets.
2016	(Zhu et al., 2016)	ChaLearn Looking At People (LAP) chal5ylenge 2015	The methodology comprises multiple stages. Initially, face detection and landmark detection are conducted on the images. Data augmentation methods are employed to generate more training samples. A deep neural network, specifically GoogLeNet, undergoes pre-training on a large dataset, followed by fine-tuning using age-labeled data. Subsequently, a separate fine-tuning step is performed using apparent age data. Age estimation is achieved through regression using the Euclidean loss function.
2020	(Cao et al., 2020)	MORPH-2, CACD, and Asian Face Database (AFAD)	The methodology presents the COnsistent RANk Logits (CORAL) framework for ordinal regression, which involves converting ranks into binary labels and employing a consistent rank logits CNN (CORAL-CNN) for prediction. The loss function is optimized using weighted cross-entropy, and the methodology offers theoretical guarantees for classifier consistency and rank monotonicity.
2016	(Malli et al., 2016)	IMDB-WIKI, and ChaLearn LAP 2016 challenge test set	The methodology consists of several steps, including face detection and alignment, data augmentation, age-grouping with age-shifted variations, and fine-tuning deep learning models on the ChaLearn LAP 2016 dataset. The final age estimation output is generated by combining weighted classification results from multiple models.
2021	(Savchenko, 2021)	VGGFace2, IMDB-WIKI, Adience, UTKFace,	The study utilizes a multi-task neural network for facial expression and attribute recognition. Lightweight CNN architectures are used for face recognition, and sequential training is conducted

		AffectNet, AFEW, VGAF	for age, gender, ethnicity, and emotion prediction. The models are applied in a mobile application for real-time attribute recognition.
--	--	-----------------------	---

Different approaches to face age assessment are shown by the various methods used in the cited works. By suggesting the use of straightforward HOG descriptors for age estimate, (Fernández et al., 2015) overcame earlier restrictions. (Saxena et al., 2021) leveraged RGB information to estimate age using a CNN architecture with convolution, ReLU, and pooling layers. To evaluate the quality of facial images, (Ou et al., 2021) presented SDD-FIQA, which makes use of both inter- and intra-class similarity distributions. (Zhang et al., 2017) suggested a CNN-based technique to age and gender prediction from unconstrained facial photos by employing Residual Networks of Residual Networks (RoR). Face detection, landmark detection, data augmentation, and deep neural networks were all used as part of (Zhu et al., 2016) multi-stage approach to age estimate. Using a consistent rank logits convolutional neural network (CNN), (Cao et al., 2020) introduced the CORAL framework for ordinal regression. (Malli et al., 2016) used deep learning models for age estimation and incorporated features such face identification, alignment, data augmentation, and fine-tuning. Using lightweight CNN architectures for face recognition and sequential training for age, gender, ethnicity, and emotion prediction, (Savchenko, 2021) implemented a multi-task neural network for facial expression and attribute recognition. In sum, these approaches demonstrate the utilisation of a wide range of methods and datasets to take on the challenging problem of age estimate in facial photos.

3. Problem Statement

While existing state-of-the-art methods utilize complex machine learning architectures such as the CORAL framework (Cao et al., 2020) and SDD-FIQA (Ou et al., 2021) with datasets like IMDB-Wiki (IMDB-WIKI - 500k+ face images with age and gender labels, 2023), UTKFace (UTKFace | Large Scale Face Dataset, 2023), and Adience, there is a lack of high-performing lightweight models. This study aims to address this gap by exploring the use of more complex classifiers on top of extracted features from lightweight models(Savchenko, 2021) to enhance their accuracy in age detection.

4. Aim and Objectives

The primary aim of this research proposal is to enhance the capabilities of lightweight machine learning models and bridge the performance gap between these models and state-of-the-art approaches in age estimation. Specifically, the study aims to achieve this by either reducing the complexity of existing state-of-the-art models or incorporating complex classifiers on top of lightweight machine learning models. The proposed approach will be evaluated using the Adience dataset, with the goal of narrowing the accuracy gap between complex and lightweight models in age estimation.

These research objectives are derived from the aim of this study and can be summarized as follows:

- To enhance the capabilities of lightweight CNN (Convolutional Neural Network) machine learning models in age estimation,
- To evaluate the performance of proposed solution.
- To compare between complex and light weight model on Adience dataset,
- To identify the most effective technique to preprocess the Adience dataset,

5. Significance of the Study

This significance of this study lies in its aim to address the performance gap that exists between lightweight models and state-of-the-art techniques in age estimation. Through the analysis of the Adience dataset and the proposal of a comprehensive methodology, this study makes a valuable contribution to the field of facial attribute identification. It aims to advance the development of age estimation models that are both highly accurate and efficient. The findings of this research have the potential to bridge the performance disparity between state-of-the-art and lightweight models, thereby enhancing the overall quality of age estimation in facial attribute recognition research.

The Proposed approach leverages pre-trained architectures and adopts a coarse-to-fine strategy to enhance the capabilities of lightweight models. By training the models on the Adience dataset, the aim is to capture the general patterns of human faces and then fine-tune the models to recognize specific features (Age)

6. Scope of the Study

The Scope of this study encompasses:

- Enhancement of lightweight machine learning models by using complex classifiers on top of the architecture using the Adience dataset.
- Bridging the gap between lightweight models and state-of-the-art models in terms of accuracy.
- Running ablations on state-of-the-art models and evaluate the change in their performance.
- The study involves data preprocessing, training sequential CNN networks, fine-tuning models, incorporating complex classifiers, and evaluating their performance.

7. Research Methodology

The initial step of our study involves analyzing the dataset and conducting Exploratory Data Analysis (EDA) on it.

The Adience dataset comprises of 26,580 images of 2284 subjects, none of which have been normalized in terms of size or orientation. In addition to this, it assigns individuals' ages to one of eight predetermined categories (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60-) and uses binary gender labels. The dataset is broken up into a total of five different subgroups. The primary aim of creating this dataset was to capture images that closely simulate real-world conditions, encompassing diverse aspects such as appearance, pose, lighting conditions, and image quality, among other factors.

```
Each fold is described by a csv file with 12 columns:  
  
user_id - the folder in the dataset containing the image.  
original_image - image name in the dataset.  
face_id - the Face ID in the original Flickr image, can be ignored.  
age - age label of the face.  
gender - gender label of the face.  
x, y, dx, dy - bounding box of the face in the original Flickr image, can be ignored.  
tilt_ang, fiducial_yaw_angle - pose of the face in the original Flickr image, can be ignored.  
fiducial_score - score of the landmark detector, can be ignored.
```

Figure 1: ReadMe of the Adience dataset

This dataset is organized into five folds, allowing for 5-fold 'leave one out' cross-validation to prevent overfitting. Each fold contains different subjects, ensuring diversity.

	aligned.tar.gz	2014-06-18 16:51 2.6G
	faces.tar.gz	2014-06-18 15:04 1.2G
	fold_0_data.txt	2014-12-15 09:57 355K
	fold_1_data.txt	2014-12-15 09:57 297K
	fold_2_data.txt	2014-12-15 09:57 310K
	fold_3_data.txt	2014-12-15 09:57 279K
	fold_4_data.txt	2014-12-15 09:57 307K

Figure 2: Five folds in the Adience dataset

user_id	original_image	face_id	age	gender	x	y	dx	dy	tilt_ang	fiducial_yaw_angle	fiducial_score
30601258@n03	10399646885_67c7d20df9_o.jpg	1	(25, 32)	f	0	414	1086	1383	-115	30	17
30601258@n03	10424815813_c94629b1ec_o.jpg	2	(25, 32)	m	301	105	640	641	0	0	94
30601258@n03	10437979845_5985be4b26_o.jpg	1	(25, 32)	f	2395	876	771	771	175	-30	74
30601258@n03	10437979845_5985be4b26_o.jpg	3	(25, 32)	m	752	1255	484	485	180	0	47
30601258@n03	11816644924_075c3d8d59_o.jpg	2	(25, 32)	m	175	80	769	768	-75	0	34
30601258@n03	11562582716_075c3d8d59_o.jpg	1	(25, 32)	f	0	422	1332	1498	-100	15	54
30601258@n03	10424595844_1009c687e4_o.jpg	4	(38, 43)	f	1912	905	1224	1224	155	0	64
30601258@n03	9586931745_796308c04a_o.jpg	5	(25, 32)	f	1069	581	1575	1575	0	30	131
30601258@n03	10190308156_5c748ab2da_o.jpg	5	(25, 32)	f	474	1893	485	484	-115	30	55
30601258@n03	10190308156_5c748ab2da_o.jpg	2	(25, 32)	m	1013	1039	453	452	-75	0	59
30601258@n03	11624488765_9db0b93c94_o.jpg	2	(25, 32)	m	101	56	740	740	-90	0	75
30601258@n03	10204739113_0e2ae11708_o.jpg	6	(25, 32)	m	336	640	841	842	-85	0	94
30601258@n03	10204739113_0e2ae11708_o.jpg	1	(25, 32)	f	693	247	720	720	-85	30	132
30601258@n03	11518638385_cac7193c86_o.jpg	2	(25, 32)	m	87	20	728	728	-95	0	79
30601258@n03	11241841104_2bc04809e0_o.jpg	1	(25, 32)	f	1039	1432	624	625	185	30	120
30601258@n03	11431644464_5510e0b7e9_o.jpg	2	(25, 32)	m	223	58	780	781	-85	0	40
30601258@n03	11562657036_5fe2235hed_o.jpg	5	(25, 32)	f	518	234	444	444	-15	0	78
30601258@n03	11438175534_c13ee0375c_o.jpg	2	(25, 32)	m	890	229	746	746	-105	30	132
30601258@n03	11438175534_c13ee0375c_o.jpg	1	(25, 32)	f	996	1222	733	733	-85	0	109

Figure 3: txt file included with the Adience dataset

In order to perform preliminary processing on the data, considering that the Adience dataset is made up of images that have not been filtered or cropped in any way, which simulates images taken from the real world as evident by Figure 4.

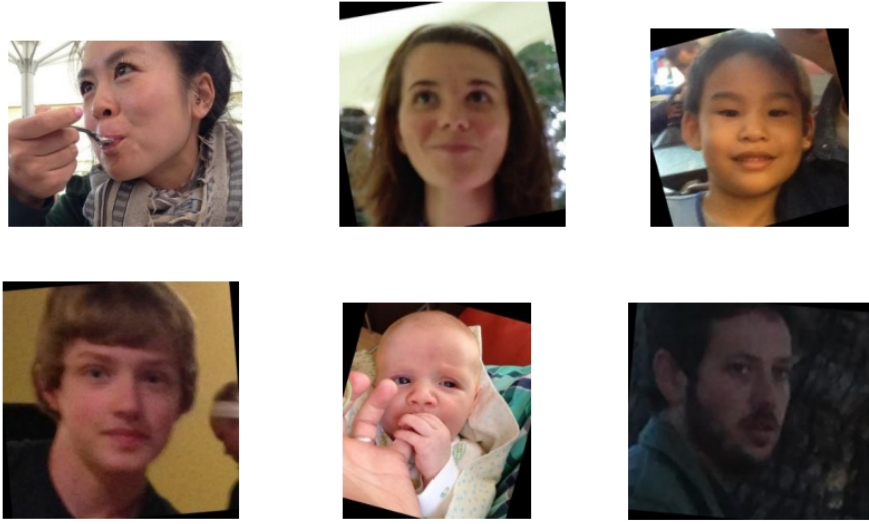


Figure 4: Sample images from the Adience dataset

The Images from left to right, then top to bottom, have the following dimensions:

1st Image: Width 752 pixels, Height 600 pixels

2nd Image: Width 440 pixels, Height 439 pixels

3rd Image: Width 488 pixels, Height 588 pixels

4th Image: Width 383 pixels, Height 383 pixels

5th Image: Width 284 pixels, Height 326 pixels

6th Image: Width 242 pixels, Height 211 pixels

Consequently, it is crucial to resize the images to a standardized size of 224 by 224 or any other specified dimensions. Subsequently, the RGB (Red, Green, Blue) values and Z score values need to be normalized, and lastly, the ground truth responses must undergo min-max normalization using the algorithm below.

$$\tilde{y}_i = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}}.$$

Figure 5: Scaling Based on the Minimum and Maximum Values.

Now this cropped image would be passed through YOLOv8(Kuprashevich and Tolstykh, n.d.) This will assign all the of the faces or bodies so that our model can make effective use of them., Since we want to improve the capabilities of lightweight models, we will be using pre-trained architectures that have a high level of accuracy for face recognition, and our approach will be one that goes from coarse to fine.

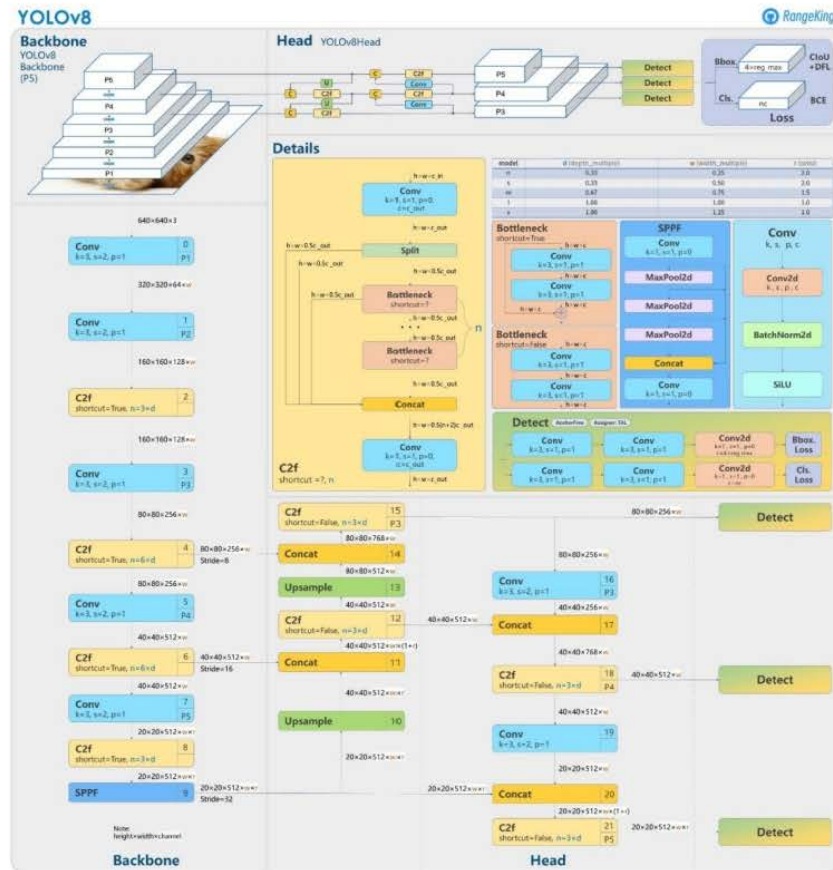


Figure 6: YOLOv8 Architecture (RangeKing (Range King), 2023)

In the first stage of model building, a simple CNN network would be trained sequentially using various pre trained architectures to bucket the probable age of the person from a sample image taken from the Adience dataset (Hassner et al., 2015) After preprocessing, the images are subjected to training with an optimal learning rate and suitable optimizer, such as ADAM, for

a limited number of epochs, Since the age outputs are in bucketed form during the initial stage, the SoftMax activation function is preferred. As the process progresses, the number of epochs will be increased, and the categorical cross-entropy loss function will be optimized using the Sharpness Aware Minimization (SAM) technique (Foret et al., 2020), This pretraining should assist the network to be more robust, resilient and accurate in capturing the general or basic pattern of a human face, and for second stage the model will be fine-tuned to better recognize the specific features, The subsequent stage would be to select the models with appropriate outcomes in terms of accuracy to choose which one are to be preferred and further used, to reduce the risk of overfitting in our model we will also add dropout layers along with data augmentation and regularization techniques wherever they are applicable or needed.

At this stage the heads of the model will be frozen, in order to further improve the performance of this model, more complex classifiers such as graph convolutional networks or transformers or such, would be utilized and the model's performance would be evaluated.

In next stage more complex state of the art models would be selected based on their compatibility with our approach and their feasibility to undergo ablation, once selected these models will be evaluated in terms of their performance and after that ablation would be performed on them to reduce their complexity while at the same time, they would be fine-tuned for better performance in terms of accuracy using the learning from light-weight models.

8. Required Resources

The ideal resources required for this study would include the following:

- **Adience dataset:** The dataset, published in 2014, comprises 26,580 photos distributed among 2,284 subjects. Each photo is labelled with binary gender information and assigned one label from eight distinct age groups. The dataset is divided into five separate splits. Its main objective is to capture images under realistic conditions, encompassing variations in appearance, pose, lighting conditions, and image quality.
- **Computational resources:** Sufficient computing power would be required to train and evaluate the models; This study will utilize a workstation with the following specifications:
 - **CPU:** Intel i9-10850K,
 - **GPU:** Nvidia RTX 3080 with 10GB VRAM,

- **Memory/RAM:** 32GB,
- **Storage:** SSD
- **Operating System:** Windows 11 64Bit.
- **TensorFlow Version 2.13.0 or later,**
- **Anaconda version 23.1.0 or later,**
- **CUDA Toolkit 12.2 or later,**
- **Keras 3.7.2 or later,**
- **OpenCV 4.8.0 or later.**

9. Research Plan

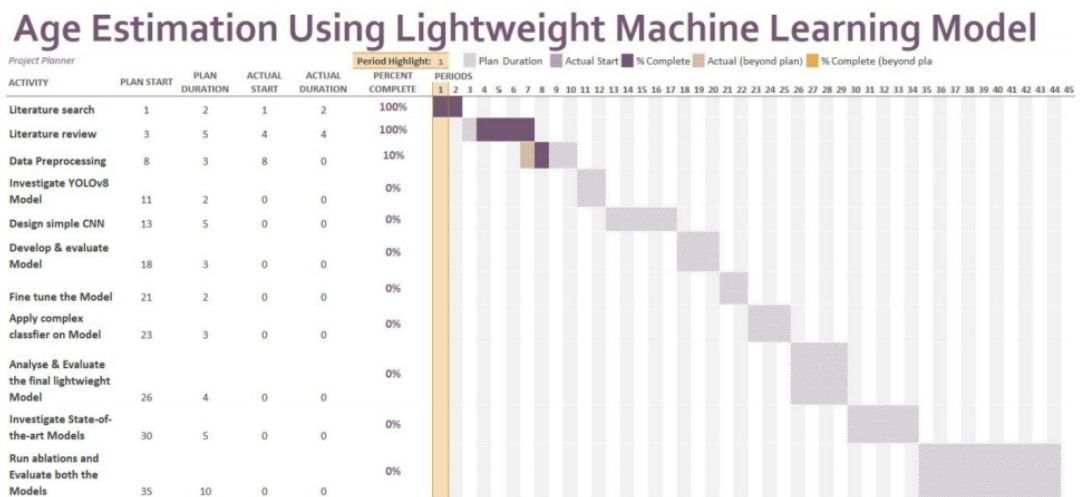


Table 2: Research Plan in Gantt Chart

References

- Anon (2023) *IMDB-WIKI - 500k+ face images with age and gender labels*. [online] Available at: <https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/> [Accessed 14 Jul. 2023].
- Anon (2023) *RangeKing (Range King)*. [online] Available at: <https://github.com/RangeKing> [Accessed 14 Jul. 2023].
- Anon (2023) *UTKFace / Large Scale Face Dataset*. [online] Available at: <https://susanqq.github.io/UTKFace/> [Accessed 14 Jul. 2023].
- Cao, W., Mirjalili, V. and Raschka, S., (2020) Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140, pp.325–331.

Demochkina, P. and Savchenko, A. V., (2021) MobileEmotiFace: Efficient Facial Image Representations in Video-Based Emotion Recognition on Mobile Devices. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, [online] 12665 LNCS, pp.266–274. Available at: https://link.springer.com/chapter/10.1007/978-3-030-68821-9_25 [Accessed 14 Jul. 2023].

Escalera, S., Gonzalez, J., Baro, X., Pardo, P., Fabian, J., Oliu, M., Escalante, H.J., Huerta, I. and Guyon, I., (2015) ChaLearn looking at people 2015 new competitions: Age estimation and cultural event recognition. *Proceedings of the International Joint Conference on Neural Networks*, 2015-September.

Fernández, C., Huerta, I. and Prati, A., (2015) A comparative evaluation of regression learning algorithms for facial age estimation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, [online] 8912, pp.133–144. Available at: https://link.springer.com/chapter/10.1007/978-3-319-13737-7_12 [Accessed 13 Jul. 2023].

Foret, P., Kleiner, A., Mobahi, H. and Neyshabur, B., (2020) Sharpness-Aware Minimization for Efficiently Improving Generalization. *ICLR 2021 - 9th International Conference on Learning Representations*. [online] Available at: <https://arxiv.org/abs/2010.01412v3> [Accessed 13 Jul. 2023].

Guo, G., (2012) Human age estimation and sex classification. *Studies in Computational Intelligence*, [online] 409, pp.101–131. Available at: https://link.springer.com/chapter/10.1007/978-3-642-28598-1_4 [Accessed 14 Jul. 2023].

Hassner, T., Harel, S., Paz, E. and Enbar, R., (2015) Effective face frontalization in unconstrained images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June, pp.4295–4304.

Karkkainen, K. and Joo, J., (2021) FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021*, pp.1547–1557.

Kuang, Z., Huang, C. and Zhang, W., (2016) Deeply Learned Rich Coding for Cross-Dataset Facial Age Estimation. *Proceedings of the IEEE International Conference on Computer Vision*, 2015-February, pp.338–343.

Kuprashevich, M. and Tolstykh, I., (n.d.) MiVOLO: Multi-input Transformer for Age and Gender Estimation. [online] Available at: <https://github.com>. [Accessed 13 Jul. 2023].

Liu, X., Li, S., Kan, M., Zhang, J., Wu, S., Liu, W., Han, H., Shan, S. and Chen, X., (2016) AgeNet: Deeply Learned Regressor and Classifier for Robust Apparent Age Estimation. *Proceedings of the IEEE International Conference on Computer Vision*, 2015-February, pp.258–266.

Malli, R.C., Aygun, M. and Ekenel, H.K., (2016) Apparent Age Estimation Using Ensemble of Deep Learning Models. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp.714–721.

Ou, F.Z., Chen, X., Zhang, R., Huang, Y., Li, S., Li, J., Li, Y., Cao, L. and Wang, Y.G., (2021) SDD-FIQA: Unsupervised Face Image Quality Assessment with Similarity Distribution Distance. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.7666–7675.

Savchenko, A. V., (2021) Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. *SISY 2021 - IEEE 19th International Symposium on Intelligent Systems and Informatics, Proceedings*, pp.119–124.

Saxena, A., Singh, P. and Singh, S.N., (2021) Gender and age detection using deep learning. *Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering*, pp.719–724.

Sharma, A.K. and Foroosh, H., (2020) Slim-CNN: A Light-Weight CNN for Face Attribute Prediction. *Proceedings - 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020*, pp.329–335.

Zhang, K., Gao, C., Guo, L., Sun, M., Yuan, X., Han, T.X., Zhao, Z. and Li, B., (2017) Age Group and Gender Estimation in the Wild with Deep RoR Architecture. *IEEE Access*, 5, pp.22492–22503.

Zhu, Y., Li, Y., Mu, G. and Guo, G., (2016) A Study on Apparent Age Estimation. *Proceedings of the IEEE International Conference on Computer Vision*, 2015-Febru, pp.267–273.