


Article

Addressing Demographic Bias in Age Estimation Models through Optimized Dataset Composition

Nenad Panić *, Marina Marjanović and Timea Bezdán 

Faculty of Technical Sciences, Singidunum University, Danijelova 32, 11000 Belgrade, Serbia; mmarjanovic@singidunum.ac.rs (M.M.); tbezdán@singidunum.ac.rs (T.B.)

* Correspondence: nenad.panic.22@singimail.rs

Abstract: Bias in facial recognition systems often results in unequal performance across demographic groups. This study addresses this by investigating how dataset composition affects the performance and bias of age estimation models across ethnicities. We fine-tuned pre-trained Convolutional Neural Networks (CNNs) like VGG19 on the diverse UTKFace dataset (23,705 samples: 10,078 White, 4526 Black, 3434 Asian) and APPA-REAL (7691 samples: 6686 White, 231 Black, 674 Asian). Our approach involved adjusting dataset compositions by oversampling minority groups or reducing samples from overrepresented groups to mitigate bias. We conducted experiments to identify the optimal dataset composition that minimizes performance disparities among ethnic groups. The primary performance metric was Mean Absolute Error (MAE), measuring the average magnitude of prediction errors. We also analyzed the standard deviation of MAE across ethnic groups to assess performance consistency and equity. Our findings reveal that simple oversampling of minority groups does not ensure equitable performance. Instead, systematic adjustments, including reducing samples from overrepresented groups, led to more balanced performance and lower MAE standard deviations across ethnicities. These insights highlight the importance of tailored dataset adjustments and suggest exploring advanced data processing methods and algorithmic tweaks to enhance fairness and accuracy in facial recognition technologies.

Keywords: facial recognition; age estimation; convolutional neural network (CNN); ethnicity bias

MSC: 68T45



Citation: Panić, N.; Marjanović, M.; Bezdán, T. Addressing Demographic Bias in Age Estimation Models through Optimized Dataset Composition. *Mathematics* **2024**, *12*, 2358. <https://doi.org/10.3390/math12152358>

Academic Editors: Silvia Liberata Ullo and Li Zhang

Received: 23 June 2024

Revised: 26 July 2024

Accepted: 26 July 2024

Published: 28 July 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Bias in facial recognition systems is a critical issue, impacting fairness, transparency, and accuracy. This bias can manifest in various forms, including age, gender, and ethnicity disparities, often resulting from inherent assumptions and decision-making processes embedded within the model architecture. For example, studies have analyzed the impact of age [1,2], demonstrating worse performance on children's faces. Other studies compare face recognition performance between males and females [3,4], showing that face recognition systems perform worse for females, partly because women's faces are generally more covered due to longer hair.

Addressing bias in facial recognition systems is essential to ensure these technologies are equitable and just, respecting the rights and dignity of all individuals while complying with various regulations aimed at preventing discrimination. These include the Universal Declaration of Human Rights, the European Convention on Human Rights, and the General Data Protection Regulation (GDPR) [5,6].

Public training datasets often exacerbate the issue of bias by being heavily skewed toward certain ethnic groups, particularly White/Caucasian faces. This lack of diversity can result in less accurate recognition of individuals from underrepresented ethnic groups. Models trained on such biased datasets fail to generalize well across different demographic groups, leading to systematic inaccuracies and unfair outcomes.

As automatic age estimation becomes increasingly used in applications like forensics [7] and surveillance, this facial recognition sub-task has garnered significant research attention. This study focuses on improving the fairness of age estimation models, specifically addressing racial bias.

Utilizing the UTKFace and APPA-REAL datasets, chosen for their demographic diversity and inclusion of labels such as real age and ethnicity necessary for this research, we investigate the impact of unbalanced training data on model performance and bias. The primary research questions are as follows: (1) How does adjusting dataset composition affect the performance of age estimation models? (2) Can rebalancing datasets mitigate bias across different ethnic groups? These questions are crucial for understanding how to develop more equitable and accurate facial recognition systems. Rather than focusing on outperforming the state-of-the-art on these two datasets, we aim to:

- Analyze the relationship between dataset composition and both overall and ethnicity-specific model performance;
- Quantify the extent to which dataset rebalancing can mitigate bias in age estimation models;
- Determine whether dataset rebalancing alone is sufficient or if it should be combined with other bias mitigation techniques.

By addressing these objectives, this study aims to contribute to the development of fairer and more accurate age estimation models, ultimately enhancing the reliability and equity of facial recognition technology.

Age estimation can be tackled through various methods, including manual feature extraction techniques and deep learning models like Convolutional Neural Networks (CNNs). Researchers have extensively studied factors that influence facial aging, which can be intrinsic (genetic) or extrinsic (environmental). Several methods for image representation and age modeling have been explored. These include anthropometric models, active shape models (ASMs), active appearance models (AAMs), aging pattern subspace (AGES), age manifolds, appearance models, and hybrid models. Additionally, feature extraction techniques such as Gabor filters, linear discriminant analysis (LDA), local binary patterns (LBPs), local directional patterns (LDPs), local ternary patterns (LTPs), gray-level co-occurrence matrix (GLCM), spatially flexible patches (SFPs), Grassmann manifolds, and biologically inspired features (BIFs) have also been investigated [8].

Angulu et al. (2018) surveyed various age estimation techniques, summarizing the Mean Absolute Error (MAE) and Cumulative Score (CS) of different age or age-group estimation models. They found that hybrid approaches, combining classification and regression, generally outperform using either method alone. Furthermore, deep learning methods, particularly CNNs, have demonstrated promising results, often surpassing traditional methods [8].

ELKarazle et al. (2022) supported these findings by providing a comprehensive overview of machine learning techniques for estimating age from facial images. They highlighted the challenges in this task, such as variations in aging patterns among individuals due to genetics, lifestyle, health conditions, and environmental factors. Additionally, they noted the limited availability of diverse and high-quality facial image datasets covering a wide range of ages, genders, and ethnicities, as well as the variations in lighting, pose, and facial expressions that influence overall age estimation accuracy and performance across different ethnic groups. ELKarazle et al. concluded that deep learning models, especially those based on transfer learning, generally outperform handcrafted models due to their ability to learn complex features automatically [9].

This conclusion is further supported by another study where manual feature extraction techniques were applied to the facial-age dataset and the UTK Face dataset. The resulting filtered images were converted to scalars and fed to a Random Forest classifier and a Support Vector classifier. When compared with a CNN, the CNN outperformed traditional machine learning techniques for age classification by up to 40% [10]. This proven track record of CNNs is why we have chosen to use them for our research purposes.

As mentioned earlier, the age estimation problem has garnered significant attention from researchers. An in-depth comparison of studies on age estimation, as well as dataset imbalance mitigation, can be seen in Table 1.

Table 1. Comparison of previous work.

Paper	Methodologies	Advantages	Disadvantages	Main Findings
Analysis of Race and Gender Bias in Deep Age Estimation Models [11]	WideResNet (UTKFace, IMDB-WIKI) and FaceNet (IMDB-WIKI)	Utilizes pre-trained models; comprehensive evaluation using MAE and RMSE	Performance inconsistency across datasets; possible impact of non-racial factors like image quality and pose	Male subjects have more accurate age estimation; inconsistent race bias across datasets; suggests makeup as a factor for gender bias
FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age [12]	ResNet-34, evaluated on FairFace, UTKFace, LFWA+, CelebA	Balanced dataset across seven racial groups; improved cross-dataset performance	Potential biases in external datasets; limited evaluation metrics	FairFace model shows consistent performance with less than 1% accuracy discrepancy between male/female and White/non-White classifications
Age and Gender Prediction From Face Images Using Attentional Convolutional Network [13]	Residual Attention Network (RAN) and ResNet	Attention mechanisms for important facial regions; ensemble model improves performance	Requires complex model training; potential overfitting	Ensemble model outperforms individual models in age and gender prediction accuracy
An Intelligent Human Age Prediction from Face Image Framework Based on Deep Learning Algorithms [14]	Deep Convolutional Neural Network (DCNN) with Cuckoo Search (CS)	High accuracy and computational efficiency; novel use of CS for optimization	Complexity in preprocessing; requires extensive computational resources	DCNN-CS model outperforms other methods like CNN, DNN, LSTM, and SVR in accuracy and efficiency
Age Estimation on Human Face Image Using Support Vector Regression and Texture-Based Features [15]	SVR with LBP, LPQ, and BSIF for feature extraction	Effective feature extraction methods; combines multiple texture features	Limited to texture-based features; high computational cost	Combination of BSIF and LPQ features with a PCA dimension of 70 achieves the best MAE
Age estimation via face images: a survey [8]	Various validation strategies; multi-manifold metric learning	Comprehensive overview; emphasizes validation strategies	Does not provide new experimental results; limited practical applications	Highlights the importance of validation strategies to avoid overfitting and enhance generalization
From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation [16]	CNN with bias correction on predictions	Addresses bias correction; improves real-age estimation	Limited to specific biases; requires an extensive dataset	Using apparent labels for training improves real-age estimation; bias correction enhances performance
Diagnosing deep learning models for high-accuracy age estimation from a single image [17]	Systematic diagnosis with deep learning models; multi-task learning architectures	Comprehensive evaluation of training procedures; multi-task approach	High complexity; requires large datasets	Regression-based approach with MAE loss favored; multi-task learning architecture outperforms other models

Table 1. Cont.

Paper	Methodologies	Advantages	Disadvantages	Main Findings
On the effect of age perception biases for real age regression [18]	VGG16 with face attributes integration	Incorporates human perception biases; improves age estimation accuracy	Requires complex model adaptation; extensive training	Incorporating face attributes enhances age estimation; attribute-based analysis reveals influence of gender, race, happiness, and makeup
A survey of Methods for Managing the Classification and Solution of Data Imbalance Problem [19]	Oversampling (SMOTE, ADASYN), Undersampling (RUS, T-Link), SVM, KNN, Naïve Bayes, Decision Tree, Bagging, Boosting (AdaBoost, RUSBoost, SMOTEBoost)	Highlights methods that improve minority class prediction and reduce class imbalance impacts; hybrid methods often yield better accuracy	Increased computational cost; potential information loss during undersampling; challenges with high-dimensional data	Data imbalance remains a critical challenge impacting classifier performance; hybrid and ensemble methods are effective but computationally intensive
Addressing the Class Imbalance Problem in Medical Datasets [20]	SMOTE oversampling, cluster-based undersampling, decision tree, Fuzzy Unordered Rule Induction Algorithm	Balances data; improves minority class prediction; effective for medical datasets	Increased computational cost; complexity in implementation	Modified cluster-based undersampling method outperforms traditional methods; SMOTE shows good classification outcome
Imbalanced Dataset Classification and Solutions: A Review [21]	Data-level techniques (SMOTE, ADASYN), Algorithm-level (cost-sensitive learning), Ensemble methods (Bagging, Boosting)	Comprehensive review of techniques; applicable across various domains	Complexity of implementation; lack of empirical validation for some techniques	Techniques like SMOTE and cost-sensitive learning are effective; method choice should be dataset-specific for optimal results
Handling imbalanced datasets: A review [22]	Various re-sampling techniques (random oversampling, undersampling), cost-sensitive learning, ensemble methods (boosting, bagging), feature selection	Comprehensive overview; highlights multiple approaches; practical recommendations	No single best method; some techniques can lead to overfitting or loss of useful data	Emphasizes the need for a tailored approach based on dataset characteristics; recommends ensemble methods and cost-sensitive learning

Previous works, such as the analysis by Puc et al. (2020) [11], investigate the performance of pre-trained age estimation models on datasets like UTKFace and APPA-REAL across different race and gender groups. They find that models tend to be more accurate for males than females, suggesting gender bias. Additionally, performance differences across races show inconsistent variations between datasets, indicating that factors like image quality and pose may also impact accuracy. While Puc et al. (2020) [11] acknowledge potential dataset imbalance, they do not actively manipulate the dataset to mitigate bias. Our study specifically investigates the effects of rebalancing datasets to achieve a more equitable representation of different racial groups, providing concrete evidence of rebalancing as a bias mitigation strategy.

Karkkainen and Joo (2019) [12] introduce the FairFace dataset to mitigate racial bias in facial attribute datasets. However, they do not explore the impact of unbalanced training data on model performance and bias as we have. Their use of age ranges rather than exact ages complicates direct comparison with our work, which focuses on precise age regression.

Despite achieving relatively comparable performance across racial groups, their accuracy remains around 60%.

Abdolrashidi et al. (2020) [13] focus on improving the accuracy of age and gender prediction using an ensemble of attentional and residual convolutional neural networks. Although they utilize the demographically diverse UTKFace dataset, they do not analyze the impact of dataset imbalance on model performance across ethnicities or genders. Their study highlights challenges in age and gender prediction due to intra-class variations but does not investigate bias mitigation strategies. Our work extends their findings by specifically addressing dataset imbalance and exploring techniques to mitigate bias.

Sathyavathi and Baskaran (2023) [14] focus on improving age prediction accuracy using a deep learning framework combining a Deep Convolutional Neural Network (DCNN) with a Cuckoo Search (CS) algorithm. While their datasets (UTKFace, FGNET, CACD) are diverse, they do not analyze the effect of dataset composition on performance for specific ethnic groups or explore bias mitigation strategies. Their study mentions potential issues with capturing relevant features for accurate age estimation but does not delve deeper into these issues or investigate their causes.

Amelia and Wahyono (2022) [15] aim to improve age estimation accuracy using texture-based features and Support Vector Regression (SVR). They acknowledge the limitations of their dataset, which primarily consists of images from Western countries and may not perform well on Asian images due to underrepresentation. This suggests potential bias, but this has not been deeply investigated in their study.

The survey on age estimation by Angulu et al. (2018) [8] discusses dataset challenges but does not deeply explore imbalances related to ethnicity. Our research targets the critical issue of dataset imbalance, ensuring fairness and accuracy across diverse ethnic groups.

The paper by Xing et al. (2019) [17] explores model architectures but does not extensively evaluate performance across diverse ethnic groups. Our research meticulously assesses accuracy degradation concerning ethnicity, providing valuable insights into algorithmic biases in facial recognition systems.

The study by Clapes et al. (2018) [16] analyzes biases in apparent age estimation and leverages this to improve real age estimation. They identify target bias (gender, ethnicity, makeup, facial expression) and guess bias (biases introduced by guessers). Although they do not explicitly investigate the impact of unbalanced training data, their work touches on the issue of bias in age estimation models.

Jacques et al. (2019) [18] focus on improving real-age estimation by incorporating apparent age and facial attributes into an end-to-end deep learning model. They highlight the importance of considering biases related to facial attributes in age estimation but do not address dataset imbalance or rebalancing techniques directly. Their work identifies several issues in existing methods, such as bias in age perception influenced by various factors, including gender, race, facial expression, and makeup, affecting the accuracy of age estimation models.

Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem (2020) by Hasib et al. [19] provides an extensive review of methods addressing class imbalance in datasets. This paper categorizes methodologies into data-level methods, algorithm-level methods, ensemble methods, and hybrid methods. Data-level methods include oversampling techniques like SMOTE and ADASYN, which generate synthetic data to balance class distributions, and undersampling techniques like RUS and T-Link, which reduce the size of the majority class to balance the dataset. Algorithm-level methods focus on enhancing classifiers like SVM, KNN, Naïve Bayes, and Decision Trees to handle imbalanced data.

Ensemble methods such as Bagging and Boosting combine multiple algorithms to improve classification performance. The paper also highlights hybrid methods that integrate data sampling and algorithm boosting to address class imbalance effectively. The study emphasizes the computational challenges associated with these methods, particularly in handling high-dimensional data and ensuring minimal information loss during sampling.

Overall, the paper concludes that while hybrid and ensemble methods provide significant improvements, they come with increased computational costs and complexity.

The paper “Addressing the Class Imbalance Problem in Medical Datasets” by Rahman and Davis (2013) [20] investigates the performance of oversampling and undersampling techniques to balance cardiovascular data. The methodologies used include the SMOTE oversampling technique and a cluster-based undersampling technique. The advantages of this approach are that it balances the data, improves minority class prediction, and is effective for medical datasets. However, it has disadvantages, such as increased computational cost and complexity in implementation. The main finding of the paper is that the modified cluster-based undersampling method outperforms traditional methods, and SMOTE shows a good classification outcome.

In “Imbalanced Dataset Classification and Solutions: A Review” (2014) by Dr. D. Ramyachitra and P. Manikandan [21], the authors present a comprehensive review of the challenges and solutions associated with imbalanced datasets. They explore various methodologies, including data-level techniques like SMOTE, algorithmic-level methods such as cost-sensitive learning, and ensemble methods like Bagging and Boosting. The advantages of their approach include a detailed categorization of methods and their effectiveness in different scenarios. However, the paper also highlights the complexity of implementing these methods and the lack of empirical validation for some techniques. Their main findings suggest that while techniques like SMOTE and cost-sensitive learning are effective, the choice of method should be tailored to the specific characteristics of the dataset to achieve optimal results.

Handling Imbalanced Datasets: A Review by Kotsiantis, Kanellopoulos, and Pintelas (2006) [22] investigates various techniques for handling imbalanced datasets, which are common in real-world classification problems where one class significantly outnumbers the other. The authors categorize solutions into data-level and algorithmic-level methods. At the data level, solutions include various forms of re-sampling, such as random oversampling, random undersampling, directed oversampling, and directed undersampling, along with feature selection tailored for imbalanced datasets. Algorithmic-level solutions involve adjusting misclassification costs and decision thresholds and employing one-class learning methods. The review highlights the use of ensemble methods, such as boosting and bagging, to improve classification outcomes by combining multiple models. The paper emphasizes that while re-sampling methods can address imbalance, they may introduce other issues like overfitting or loss of useful data. The review also discusses evaluation metrics specific to imbalanced datasets, including ROC curves and precision–recall metrics. The main finding is that no single method universally outperforms others, and the choice of technique often depends on the specific characteristics of the dataset and the problem at hand.

To address dataset imbalance, we explored various strategies to mitigate bias in age estimation models by adjusting the representation of different racial groups. Unlike previous studies that employed advanced techniques and combinations such as SMOTE, ADASYN, and ensemble methods, our approach systematically reduced the sample size of specific racial groups while keeping others constant to observe the impact on model performance.

From our experimentation, we observed that dataset variations significantly influence the Mean Absolute Error (MAE) and standard deviation across different ethnic groups. For instance, reducing the sample size of the Asian group by 90% resulted in an overall MAE of 5.4808 and a standard deviation of 0.3015, indicating improved performance consistency compared to the original dataset (Overall MAE: 4.8925, Standard Deviation: 0.7401). However, reducing samples from other groups did not always yield better results, highlighting the complexity of achieving balanced performance across all demographics.

In some cases, balancing datasets by reducing samples from overrepresented groups led to lower standard deviations and more equitable performance across different racial groups. For example, balancing the White group to 20% and 40% yielded standard deviations of 0.0440 and 0.2451, respectively, indicating more stable performance.

Our findings suggest that while advanced techniques like SMOTE and ensemble methods have their merits, a systematic and simpler approach to adjusting group representation can also provide valuable insights and potentially enhance fairness in model performance. This novel approach of selectively reducing samples from specific groups may complement existing techniques and offer new perspectives in future research.

In conclusion, our study demonstrates that dataset rebalancing can effectively mitigate bias and improve the equity of age estimation models. By systematically analyzing the impact of various dataset compositions, we provide a comprehensive understanding of how different approaches affect model performance and fairness. Our findings contribute to the ongoing efforts to develop fairer and more accurate facial recognition technologies, ultimately advancing the field toward more equitable and unbiased solutions.

2. Materials and Methods

This study utilizes two publicly available datasets for experiments: the UTKFace dataset [23] and the APPA-REAL dataset [24]. These datasets were chosen for their demographic diversity and labels such as real age, ethnicity, and gender. Detailed composition of both datasets is provided in Table 2.

Table 2. Overview of the two datasets.

Name	Total Number of Samples	Gender			Race				Age Range
		Male	Female	White	Black	Asian	Indian	Others	
UTKFace	23,705	12,391	11,314	10,078	4526	3434	3975	1692	1–116
APPA-REAL	7591	3818	3773	6686	231	674			1–100

The UTKFace dataset contains 23,705 samples, while the APPA-REAL dataset contains 7591 samples. We split the UTKFace dataset into training and test sets, whereas the APPA-REAL dataset comes pre-split into training, test, and validation sets. As shown in Table 1, both datasets have an equal number of male and female samples, and they cover a wide age range from 1 to 116 years.

However, there are disparities in racial representation. Both datasets are heavily weighted towards the White race. Although the UTKFace dataset does not have an equal number of samples across different race groups, it has more samples of Black and Asian groups compared to the APPA-REAL dataset. The UTKFace dataset also includes images of the Indian group and others representing ethnicities such as Hispanic, Latino, and Middle Eastern.

Since the APPA-REAL dataset contains only three ethnic groups, we combined the Black and Indian groups from the UTKFace dataset into one and discarded the “Others” group. This decision was made to facilitate comparison, as the “Others” group contains a mix of ethnicities.

The only preprocessing steps applied to these images involved scaling them to a size of $224 \times 224 \times 3$, as required by the VGG19 model [25], and using its `preprocess_input` function, which centers the color channels at zero and converts the images from RGB to BGR. Figure 1 shows a few example images from both the UTKFace and APPA-REAL datasets.

Due to the proven track record of CNNs and their superior performance compared to manual feature extraction techniques, we chose to utilize them in our experiments. CNNs are deep learning models specifically designed to process and analyze visual data by automatically detecting features such as edges, textures, and shapes from raw pixel data through layers of convolutional filters. This enables CNNs to effectively recognize patterns and objects, making them particularly suited for tasks such as image classification, object detection, and age estimation from facial images.

We opted to employ CNNs with transfer learning for several advantages. Transfer learning harnesses knowledge acquired from a pre-trained model on a large dataset for a related task, significantly reducing training time and data requirements. Pre-trained models

have already gleaned valuable features from extensive datasets, enhancing performance on new tasks, particularly with smaller datasets. Additionally, transfer learning enables the utilization of sophisticated models without the need for extensive computational resources to train from scratch. Moreover, earlier layers of transfer learning models, designed for extracting generic features, can be fine-tuned for specific applications.



Figure 1. Example images from both datasets.

We employed widely recognized CNN models extensively used in scientific studies: VGG16 [25], VGG19 [25], ResNet50 [26], and MobileNetV2 [27], all pre-trained on the ImageNet [28] dataset. These models were selected for their proven effectiveness in feature extraction from complex visual data, including facial images, which is crucial for accurate age estimation. Specifically:

- VGG19 and ResNet50 are known for their deep architectures, allowing them to capture intricate features through multiple layers of convolutions. This depth can be advantageous in learning hierarchical representations of facial features relevant to age;
- MobileNetV2 is chosen for its efficiency and suitability for mobile and embedded applications, offering a balance between computational efficiency and performance, which is valuable for practical deployment scenarios;
- VGG16 offers a simpler architecture compared to VGG19 but still maintains strong performance in various computer vision tasks, making it a reliable benchmark in our comparative analysis.

It is important to note that intra-class performance variation was similar between these different models. VGG19 had the best performance on the original and equally oversampled dataset versions; however, all models needed to be opened for training to achieve better performance. The specific layers opened for training varied among the models, with optimal results achieved by progressively fine-tuning layers. Opening all layers for training worsened the performance of all models, underscoring the importance of selective fine-tuning.

In conclusion, while VGG19 emerged as the best-performing model in our specific experimental setup, other models like ResNet50 and MobileNetV2 also showed competitive performance. Future research could further explore the potential of ensemble methods and other advanced CNN architectures, such as EfficientNet, to handle demographic variability better and potentially enhance model fairness and accuracy.

Our primary objective was to optimize the models for predicting real age using the UTKFace and APPA-REAL datasets.

To achieve this, we conducted a grid search, varying hyperparameters within empirically justified ranges. Grid search allowed us to systematically explore combinations of hyperparameters to identify the set that resulted in the best model performance. Specifically, we tested different learning rates, batch sizes, and numbers of epochs within specified ranges. The grid search involved training multiple instances of each model with different

combinations of these hyperparameters and selecting the combinations that yielded the lowest Mean Absolute Error (MAE) on the validation set.

The learning rate was varied within the range of 0.1 to 0.000001. This wide range was chosen to balance between convergence speed and fine-grained adjustments. Higher learning rates can lead to faster convergence but might overshoot optimal values, resulting in poor performance. Lower learning rates allow for finer adjustments, improving the model's ability to converge to a more optimal solution. This range allows us to explore different levels of model weight updates, ensuring we do not miss any optimal points that might occur at different scales. The optimal learning rate for both datasets was 0.0001, which minimized the MAE on the validation set.

Batch sizes were explored within the range of 16 to 128. This range was selected to balance computational efficiency and model generalization. Smaller batch sizes, while computationally more expensive, did not lead to better generalization compared to a batch size of 64. Larger batch sizes utilized GPU capabilities more efficiently but sometimes led to poorer generalization. The optimal batch size of 64 provided the best accuracy, balancing computational efficiency and model generalization, and minimized the MAE observed during the grid search.

The number of epochs was varied within the range of 30 to 100. This range was chosen to find the point where the model achieves optimal performance without overfitting. More epochs help the model learn more complex patterns in the data but also increase the risk of overfitting, especially with limited data. This range is generally sufficient to allow the model to learn complex patterns while giving us the flexibility to stop training once the validation error stabilizes. Training for 60 epochs was sufficient for convergence on the APPA-REAL dataset, while 50 epochs were optimal for the UTKFace dataset. Beyond these points, additional epochs did not significantly reduce the validation error and sometimes led to overfitting.

Additionally, we evaluated different optimizers, such as Adam and SGD, assessing their impact on training dynamics and convergence speed. The configuration of fully connected layers at the model's top was also varied to optimize age estimation performance. Throughout our experiments, we progressively opened layers one by one and, eventually, the entire base model for training to evaluate and optimize performance comprehensively.

Both models used the Adam optimizer and benefitted from fine-tuning of the base model. However, the specific layers unlocked for training differed: fine-tuning commenced from layer 17 for the APPA-REAL dataset, whereas optimal results for the UTKFace dataset were achieved by fine-tuning from layer 7 onwards. To mitigate overfitting due to the relatively small dataset sizes, we applied data augmentation techniques, including image rotation (with a range of 40 degrees), width and height shift (0.2), zoom (0.2), and horizontal flip. Additionally, a random state of 42 was set for both oversampling and data augmentation.

By systematically adjusting these parameters through grid search, we aimed to find the best combination that minimized the MAE on the validation set, ensuring robust and generalizable model performance. The insights gained into the relationship between parameter settings and model performance are critical for optimizing deep learning models for age estimation tasks.

Our primary focus was not solely on achieving the highest possible accuracy but on understanding the relationship between dataset composition and model performance across different groups. Using grid search, we obtained well-performing hyperparameters that adequately meet our research objectives. While grid search provides a systematic approach to hyperparameter tuning, it is computationally expensive. Future work could incorporate advanced techniques such as Random Search or Bayesian Optimization, which sample hyperparameters from specified distributions, potentially offering better performance with reduced computational cost. Although our goal was not to achieve state-of-the-art overall performance, the insights gained into group performance relationships are robust with the

current tuning. Exploring these advanced tuning methods in future studies could further enhance both overall accuracy and fairness.

Following the identification of the best-performing model from our initial evaluation of the original dataset composition, all subsequent tests were conducted exclusively on this model. Its architecture is illustrated in Figure 2.

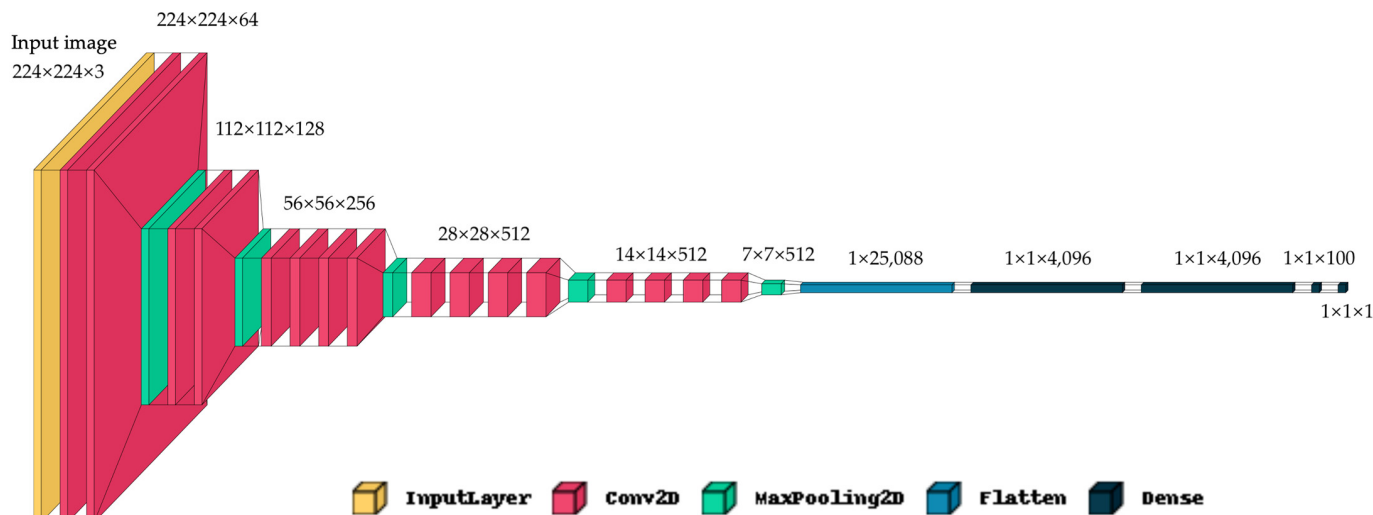


Figure 2. Representation of the VGG-19 architecture used in this research.

As illustrated in Figure 2, our model takes an input size of $224 \times 224 \times 3$. The architecture starts with a block containing two convolutional layers followed by a max pooling layer. Max pooling is a downsampling technique used to reduce the spatial dimensions of the input representation, which helps to lower computational load, control overfitting, and enhance the network’s robustness. This step is essential in CNN architectures to simplify the representation while preserving important features.

Next, we have another block with two convolutional layers and a max pooling layer. This is followed by a third block with four convolutional layers and a max pooling layer. The fourth and fifth blocks each consist of four convolutional layers and a max pooling layer. This totals 16 convolutional layers in the network.

After the convolutional and pooling layers, a flattening layer converts the multi-dimensional tensor into a one-dimensional vector. This transformation is crucial because it allows the output from the convolutional and pooling layers to be fed into fully connected (dense) layers for final classification or regression tasks, effectively bridging the two parts of the network.

Following the flattening layer are two dense layers with 4,096 nodes each, then another dense layer with 100 nodes, and finally, the output layer with a single node. The hidden convolutional and dense layers utilize the ReLU (Rectified Linear Unit) activation function, which is widely used in CNNs. ReLU introduces non-linearity into the model, enabling it to learn complex patterns and functions. It is computationally efficient, involving a simple operation of setting negative values to zero while keeping positive values unchanged, and it also helps mitigate the vanishing gradient problem. Since we are dealing with a regression task, the activation function used in the output layer is the linear activation function. The linear activation function does not transform its input, meaning the output of the neuron is directly equal to its input.

Now that we have fine-tuned our base model architectures and hyperparameters, we proceeded with the experiment. To thoroughly assess the influence of dataset composition on model performance—both overall and for specific demographic groups—we oversampled both datasets until the number of samples in the Black and Asian groups matched that of the dominant White group. This oversampled dataset serves as our baseline alongside the original dataset composition results.

The purpose of oversampling the minority groups to match the dominant group was to facilitate controlled reduction experiments, where we systematically reduced each group’s sample size from 10% to 100%. This allowed us to measure performance differences for the reduced group and observe how other groups were affected. We compared these results against the original dataset composition, the equally oversampled dataset composition, and all variations in between.

This method enabled us to identify which variations in dataset composition minimized performance variance between groups.

3. Results

In our analysis, we used Mean Absolute Error (MAE), Standard Deviation (SD), Disparate Impact (DI), and Equality of Opportunity (EoO) to evaluate the performance of our VGG19 model across different dataset compositions and ethnic groups. The formula for Mean Absolute Error (MAE) is given by:

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

where y_i is the actual age, \hat{y}_i is the predicted age, and n is the total number of samples.

To assess the variability in model performance between the ethnic groups (White, Black, and Asian), we calculated the standard deviation of the MAEs for these groups. The MAE for each group was computed separately and then used to determine the standard deviation.

The standard deviation (SD) of the MAEs is given by:

$$SD = \sqrt{\frac{1}{3} \sum_{j=1}^3 (MAE_j - \overline{MAE})^2}$$

1. MAE_j represents the MAE for the j -th group (White, Black, or Asian).
2. \overline{MAE} is the mean MAE across the three groups.

Calculating the SD of the MAEs helps in understanding the extent of performance variability among different groups, highlighting any potential biases or inconsistencies in the model’s predictions.

Additionally, incorporated fairness metrics, Disparate Impact and Equality of Opportunity, further assess the fairness of our model across different ethnic groups. These metrics are defined as follows:

Disparate Impact:

$$DI = \frac{MAE_{group}}{MAE_{White}}$$

DI measures the ratio of the MAE for a specific group to that of the White group. A DI value of 1 indicates perfect parity, suggesting no disparate impact. Values less than 1 indicate a potential disadvantage for the group in question.

Equality of Opportunity:

$$EoO = 1 - \left| \frac{MAE_{group} - MAE_{White}}{MAE_{White}} \right|$$

EoO assesses how close the MAE of a specific group is to that of the White group, adjusted for the scale of the White group’s MAE. An EoO value close to 1 suggests that both groups experience similar error rates, indicating fairness in terms of model performance.

By employing these metrics, we can better understand and mitigate any potential biases our model may exhibit, ensuring that it performs equitably across different demographic groups.

The APPA-REAL dataset, which is highly unbalanced with the Black and Asian groups together accounting for less than 10% of the available data, surprisingly shows

relatively good results in terms of MAE variations across these groups, with a standard deviation of 0.19 on a model trained on the dataset’s original form. It might be assumed that oversampling to equalize group sizes would result in equal performance across these groups or at least reduce the MAE deviation. However, our findings suggest otherwise.

The model trained with 20% fewer samples from the White group exhibited the smallest standard deviation of 0.04, with the White and Black groups performing almost equally well, having MAEs of 7.1676 and 7.1663, respectively, and the Asian group slightly better at 7.0736. The original dataset composition resulted in only the 7th smallest standard deviation among groups, with the oversampled equal dataset ranking 8th.

This indicates that six other dataset compositions performed better in terms of reducing MAE variation across groups, with the 20% reduced White group dataset achieving a standard deviation reduction of 78.94% compared to the original dataset and an 80.95% reduction compared to the equally oversampled dataset. The overall best MAE was achieved with the original dataset composition (6.45), whereas the most equal model had an overall MAE of 7.16, suggesting that achieving equal performance across groups comes at a small cost to overall performance.

The worst-performing model was trained with the dataset composition where the White group was completely omitted, resulting in an overall MAE of 8.89 and a standard deviation of 0.94. In comparison, the most equal model had a 95.74% smaller variation. Table 3 presents these results, sorted by the least standard deviation.

Table 3. Comparison of all variations on the APPA-REAL dataset sorted by lowest standard deviation between groups performance.

Dataset Variations	Overall MAE	White Group MAE	Black Group MAE	Asian Group MAE	Standard Deviation	DI Black Group	EoO Black Group	DI Asian Group	EoO Asian Group
White 20%	7.1603	7.1676	7.1663	7.0736	0.0440	0.9998	0.9998	0.9869	0.9869
Black 60%	6.7788	6.764	6.7813	6.9478	0.0828	1.0026	0.9974	1.0272	0.9728
Asian 90%	6.9338	6.956	6.9148	6.6874	0.1181	0.9941	0.9941	0.9613	0.9613
Black 70%	6.7632	6.7431	7.1447	6.8198	0.1740	1.0595	0.9405	1.0114	0.9886
Black 30%	6.7834	6.7537	7.2087	6.9303	0.1872	1.0673	0.9327	1.0261	0.9739
White 10%	6.8679	6.8687	7.1783	6.7177	0.1917	1.0451	0.9549	0.978	0.978
Original	6.4588	6.4241	6.4529	6.8593	0.1987	1.0045	0.9955	1.0677	0.9323
Equal	6.7129	6.4547	6.9672	6.5739	0.2189	1.0794	0.9206	1.0185	0.9815
White 40%	7.1286	7.153	7.3405	6.7528	0.2451	1.0262	0.9739	0.944	0.944
Asian 50%	6.6626	6.6547	7.1479	6.5305	0.2666	1.0741	0.9259	0.9813	0.9813
White 70%	7.3275	7.3464	7.625	6.9741	0.2666	1.0379	0.9621	0.9493	0.9493
Asian 100%	6.8969	6.8269	7.2377	7.5441	0.2938	1.0602	0.9398	1.1049	0.8951
Black 10%	6.7724	6.77	7.2787	6.5685	0.2988	1.0751	0.9249	0.9703	0.9703
Asian 70%	6.8689	6.9326	6.7004	6.2145	0.2991	0.9665	0.9665	0.8963	0.8963
Black 40%	7.1176	7.0839	7.7754	7.2036	0.3017	1.0977	0.9023	1.0169	0.9831
Asian 10%	6.6531	6.6385	7.2428	6.5518	0.3073	1.0910	0.909	0.9869	0.9869
Asian 40%	6.7736	6.731	7.5072	6.927	0.3295	1.1152	0.8848	1.0291	0.9709
White 30%	7.1779	7.2144	7.49	6.6156	0.3650	1.0382	0.9621	0.9169	0.9169
Black 90%	6.8216	6.8529	7.2218	6.2795	0.3877	1.0538	0.9462	0.9162	0.9162
Asian 30%	6.8468	6.8523	7.4678	6.4999	0.3999	1.0898	0.9102	0.949	0.949
Asian 60%	7.1463	7.1156	8.0009	7.1076	0.4192	1.1244	0.8756	0.9989	0.9989
Asian 20%	6.7844	6.7327	7.733	6.9448	0.4303	1.1486	0.8514	1.0315	0.9685
White 50%	7.224	7.2357	7.9872	6.7413	0.5122	1.1039	0.8961	0.9317	0.9317
White 60%	7.6724	7.7254	8.1431	6.8498	0.5389	1.0541	0.9459	0.8866	0.8866
Black 80%	6.7672	6.762	7.7404	6.3828	0.5719	1.1447	0.8553	0.9439	0.9439
Black 50%	6.8294	6.845	7.7236	6.2425	0.6081	1.1285	0.8715	0.9119	0.9119
White 90%	8.2641	8.3888	8.039	6.9359	0.6191	0.9583	0.9583	0.8267	0.8267
Black 100%	6.8285	6.7755	8.1289	6.8429	0.6227	1.2000	0.8000	1.0100	0.9900
White 80%	7.6813	7.7504	8.1856	6.6583	0.6424	1.0562	0.9438	0.8589	0.8589
Black 20%	6.8741	6.8916	7.8781	6.2144	0.6831	1.1431	0.8569	0.9017	0.9017

Table 3. Cont.

Dataset Variations	Overall MAE	White Group MAE	Black Group MAE	Asian Group MAE	Standard Deviation	DI Black Group	EoO Black Group	DI Asian Group	EoO Asian Group
Asian 80%	6.7975	6.7792	8.0498	6.4349	0.6944	1.1873	0.8127	0.9491	0.9491
White 100%	8.8956	9.1258	7.3192	6.9751	0.9432	0.8022	0.8022	0.7646	0.7646

For better clarity of the changes in model performance, we have included Figure 3 which shows the overall MAE and standard deviation for different dataset variations.

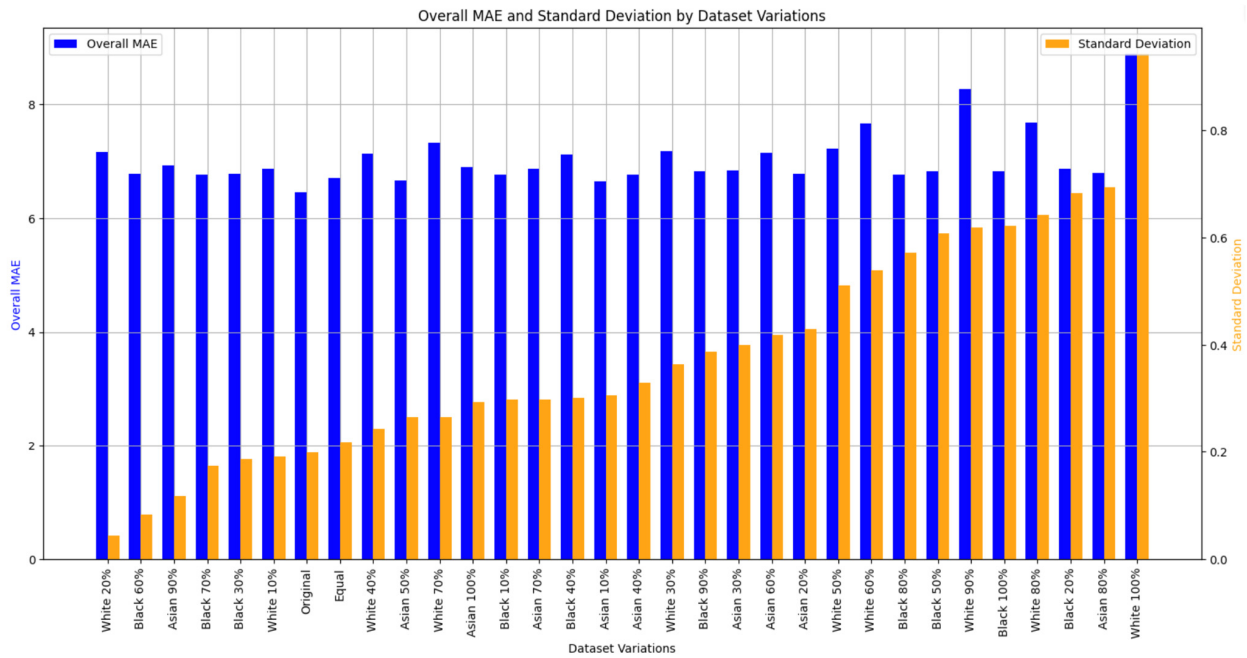


Figure 3. Overall MAE and Standard Deviation by variations of the APPA-REAL dataset.

Figure 4 provides a detailed view of MAE and standard deviation trends for different groups and overall.

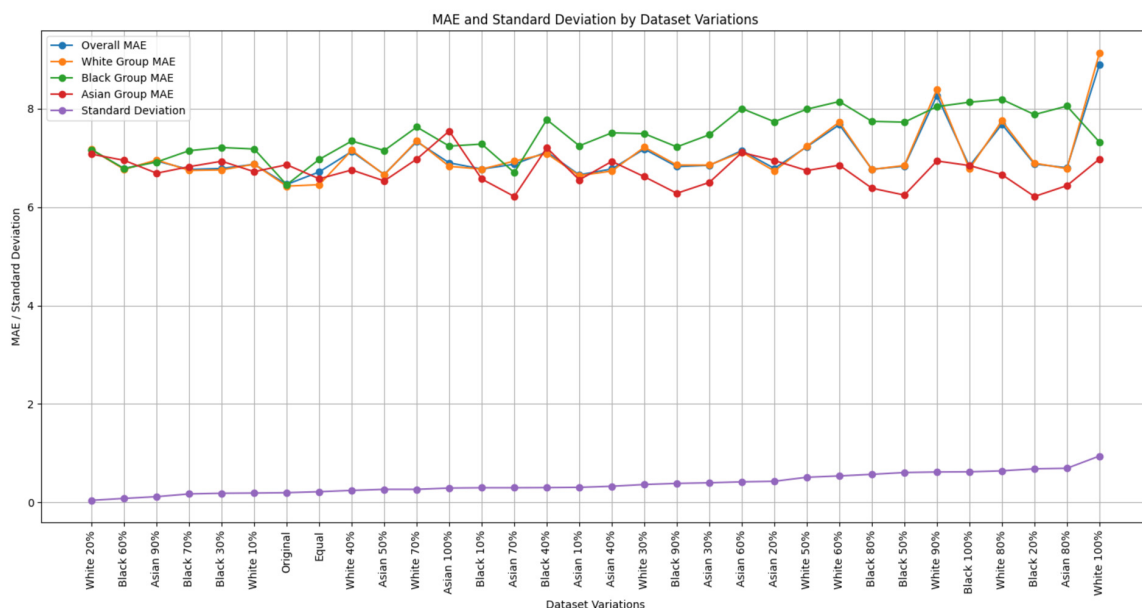


Figure 4. MAE and Standard Deviation by variations of the APPA-REAL dataset.

To further understand the fairness of these models, we analyze disparate impact (DI) and equality of opportunity (EoO) metrics. These metrics help us assess how different demographic groups are affected by the model’s performance.

Disparate Impact (DI) is a measure that indicates whether one group is adversely affected compared to another. A DI value close to 1 indicates that the model treats different groups equitably, while a value significantly different from 1 indicates potential bias.

Figure 5 shows the Disparate Impact for the Black and Asian groups across various dataset compositions.

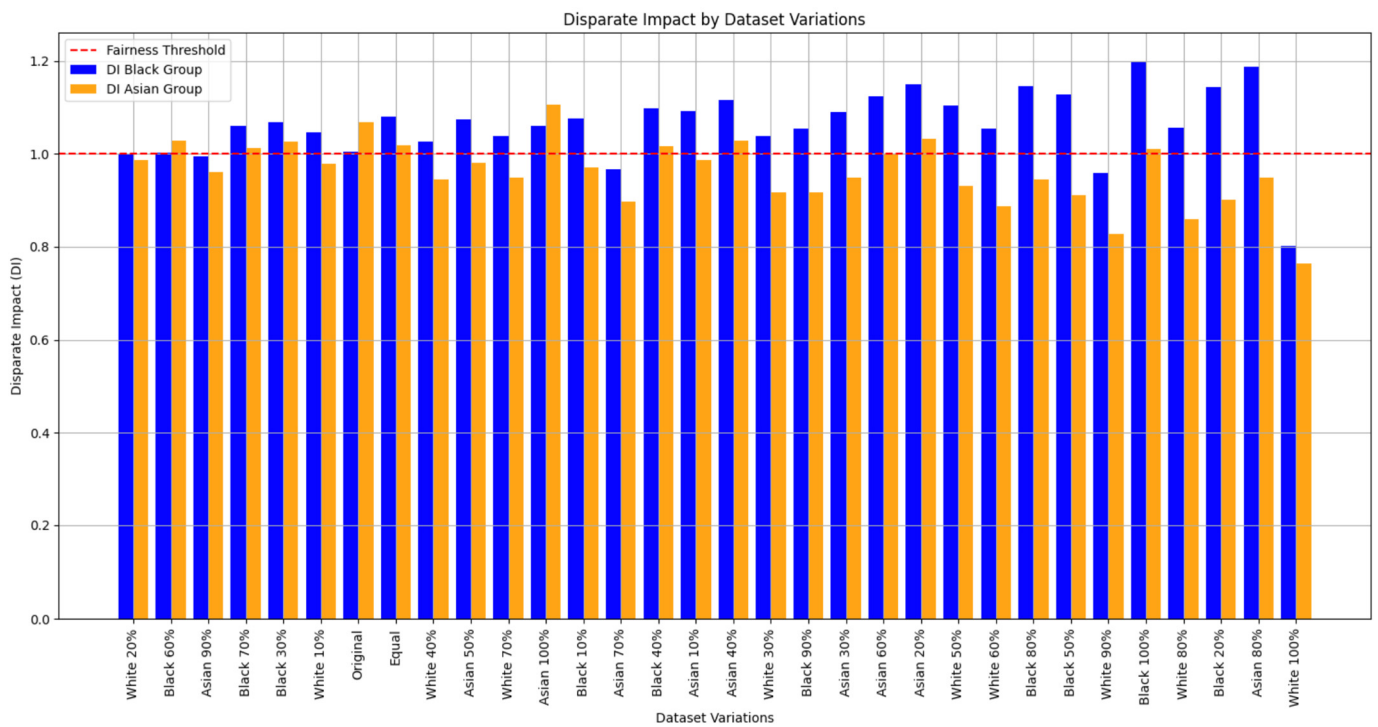


Figure 5. Disparate Impact by variations of the APPA-REAL dataset.

For the White 20% variation, the model shows DI values of 0.9998 for both Black and Asian groups, suggesting that this variation achieves almost perfect fairness across these groups. In the Black 60% variation, the DI values are 1.0026 for the Black group and 1.0272 for the Asian group. These values are close to 1, indicating minimal disparate impact and suggesting that this variation also promotes fairness. The Asian 90% variation has DI values of 0.9941 for both groups, maintaining a balance in treatment across demographic groups and ensuring minimal bias.

Interestingly, despite equalizing the sample sizes, the DI values for the Equal Dataset variation are 1.0794 for the Black group and 1.0185 for the Asian group. These values suggest a slight bias, indicating that mere equalization of dataset proportions does not necessarily eliminate disparate impact.

Equality of Opportunity (EoO) measures whether different groups have the same chances of achieving a favorable outcome. Like DI, an EoO value close to 1 indicates equitable performance across groups.

Figure 6 shows Equality of Opportunity values for the Black and Asian groups across various dataset compositions.

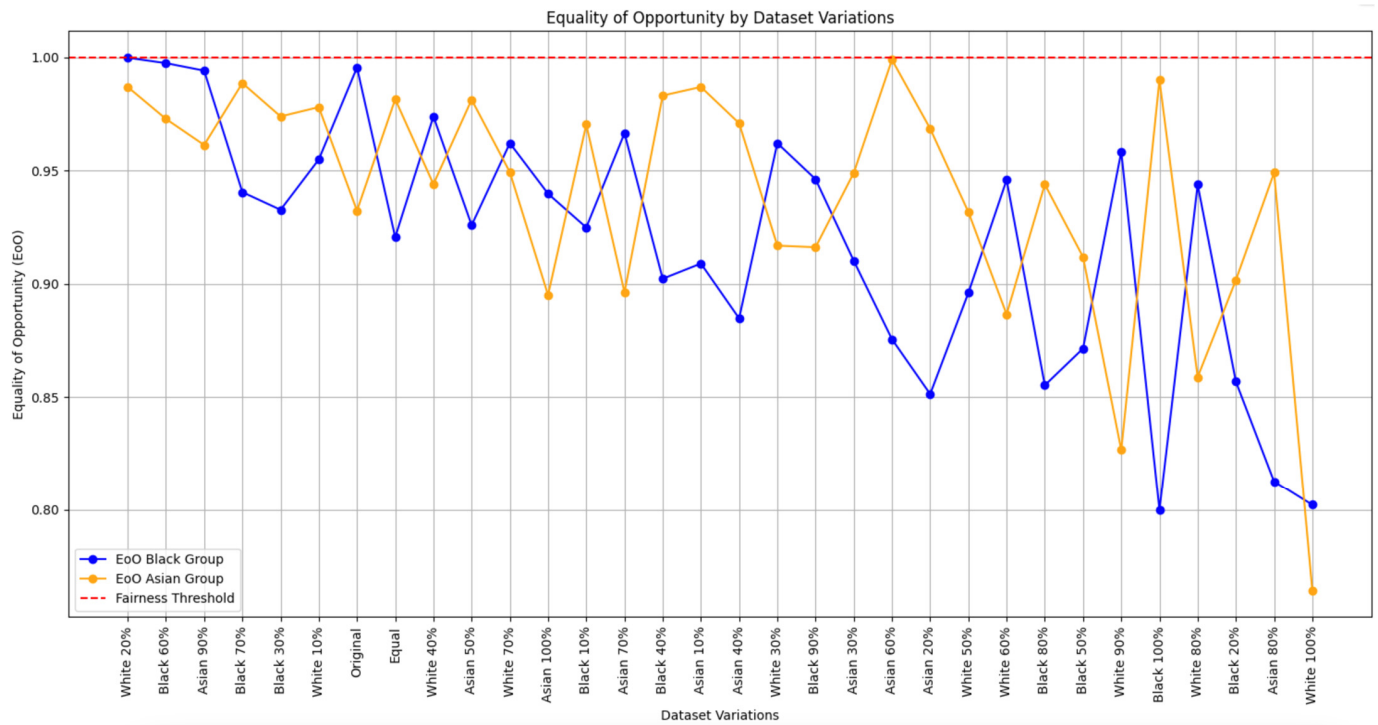


Figure 6. Equality of Opportunity by variations of the APPA-REAL dataset.

In the White 20% variation, EoO values for this variation are 0.9998 for both the Black and Asian groups, demonstrating that this model provides equal opportunity to these groups. For the Black 60% variation, the EoO values are 0.9974 for the Black group and 0.9728 for the Asian group, indicating good equality of opportunity with slight variations. The Asian 90% variation shows EoO values of 0.9941 for both groups, suggesting that the model performs equitably across these demographics.

Conversely, the EoO values for the Equal Dataset variation are 0.9206 for the Black group and 0.9815 for the Asian group. These values reflect some inequity, highlighting that equal dataset proportions do not automatically translate to equal opportunity.

Examining the lower-performing models, particularly those at the bottom of the table, reveals additional insights. For instance, the White 100% variation, with the highest standard deviation (0.9432), has DI values of 0.8022 for both Black and Asian groups, indicating a significant disparate impact. The EoO values for this variation are 0.7646, showing considerable inequality of opportunity. This suggests that completely excluding certain demographic groups can lead to substantial biases and unfair outcomes.

Another example is the Black 100% variation, which has a standard deviation of 0.6227. The DI values are 1.2000 for the Black group and 1.0100 for the Asian group, indicating a higher disparate impact, especially for the Black group. The EoO values are 0.8000 for the Black group and 0.9900 for the Asian group, further demonstrating significant inequalities in opportunity.

These examples illustrate that dataset compositions with extreme imbalances or complete exclusions of certain groups tend to exhibit higher disparities in both impact and opportunity. Therefore, it is crucial to consider both performance metrics and fairness measures when evaluating and selecting dataset variations for training models. The goal is to achieve a balance that minimizes bias and promotes equitable outcomes across all demographic groups.

Diving deeper into separate groups, we can observe in Table 4 how the MAE of a group is responding to reductions in samples; we will first discuss the White group. The performance of the White group alone is like that of the other groups, the best with the original dataset composition with an MAE of 6.42, and the second best performing is the

equal dataset, where, again, all the samples are still present. As can be seen in Table 4, there is a definite worsening of performance the more samples we remove. The increase in MAE is not linear. A reaction to a 10% decrease in samples can vary from a 0.20% increase in MAE for the group to 8.79%. It is clear, however, that decreasing the number of samples by 10% does not automatically equate to a 10% decrease in performance. Most of the time, the worsening rate is far below 10%, and it never even reaches such an immediate decrease. The standard deviation increase also closely follows our pattern of more and more samples being removed during training. The performance of the White group, when not being used for training at all, equates to a 42.06% higher MAE during testing compared to the original dataset and 41.38% compared to the equally oversampled dataset. Clearly, the presence of these samples does matter.

Table 4. Performance changes upon APPA-REAL White group reduction.

Dataset Variations	Overall MAE	White Group MAE	Black Group MAE	Asian Group MAE	Standard Deviation	MAE Change Compared to Previous	Compared to Equal	Compared to Original	Black Group Compared to Original	Black Group Compared to Equal	Asian Group Compared to Original	Asian Group Compared to Equal
Original	6.4588	6.4241	6.4529	6.8593	0.1987							
Equal	6.7129	6.4547	6.9672	6.5739	0.2189	0.48%		0.48%	7.97%		−4.16%	
10%	6.8679	6.8687	7.1783	6.7177	0.1917	6.41%	6.41%	6.92%	11.24%	3.03%	−2.06%	2.19%
20%	7.1603	7.1676	7.1663	7.0736	0.0440	4.35%	11.04%	11.57%	11.06%	2.86%	3.12%	7.60%
30%	7.1779	7.2144	7.49	6.6156	0.3650	0.65%	11.77%	12.30%	16.07%	7.50%	−3.55%	0.63%
40%	7.1286	7.153	7.3405	6.7528	0.2451	−0.85%	10.82%	11.35%	13.76%	5.36%	−1.55%	2.72%
50%	7.224	7.2357	7.9872	6.7413	0.5122	1.16%	12.10%	12.63%	23.78%	14.64%	−1.72%	2.55%
60%	7.6724	7.7254	8.1431	6.8498	0.5389	6.77%	19.69%	20.26%	26.19%	16.88%	−0.14%	4.20%
70%	7.3275	7.3464	7.625	6.9741	0.2666	−4.91%	13.81%	14.36%	18.16%	9.44%	1.67%	6.09%
80%	7.6813	7.7504	8.1856	6.6583	0.6424	5.50%	20.07%	20.65%	26.85%	17.49%	−2.93%	1.28%
90%	8.2641	8.3888	8.039	6.9359	0.6191	8.24%	29.96%	30.58%	24.58%	15.38%	1.12%	5.51%
100%	8.8956	9.1258	7.3192	6.9751	0.9432	8.79%	41.38%	42.06%	13.42%	5.05%	1.69%	6.10%

What is also very interesting to see is that the Black group’s performance also worsens as more and more samples of the White group are removed. With the highest MAE increase for that group at 26.85% compared to the performance on the original dataset, when 80% of samples from the White group are removed. Interestingly, performance starts to get better again as 90% or 100% of the White group samples are removed, with the MAE increase for the Black group compared to the original performance being 13.42% when 100% of White group samples are removed. The Black group is less reactive compared to its performance on the equally oversampled dataset, with variations in performance never reaching a 20% increase but usually being somewhere around 5%. When 100% of White samples are removed, the performance for the Black group compared to the Equal dataset is worse by only 5.05%.

The Asian group performance stays relatively the same regardless of the number of White group samples removed, with the increase in its MAE varying from −3.55% to 3.12% compared to the original dataset performance. The Asian group seems to be a bit more reactive compared with the performance on the Equal dataset, with MAE increases of around 6%. Clearly, the White and Black groups are more correlated.

Examining the performance of the Black group, as shown in Table 5, reveals a noticeable degradation as a percentage of its samples are removed. Surprisingly, there is not a linear correlation between increased sample removal and worsening performance. For instance, removing 10% of samples results in a 12.80% increase in MAE, while removing 20% leads to a 22.09% increase compared to the original dataset performance. However, beyond this point, the MAE increase fluctuates between 5.09% and 20.49%, suggesting that performance stabilizes or even improves when more samples are removed rather than just 10% or 20%. Notably, only when the Black group is entirely excluded from training do we observe the highest MAE increase, reaching 25.97% compared to the original dataset. This

increase is significantly lower than that observed for the White group, indicating greater resilience to underrepresentation among Black group samples.

Table 5. Performance changes upon APPA-REAL Black group reduction.

Dataset Variations	Overall MAE	White Group MAE	Black Group MAE	Asian Group MAE	Standard Deviation	MAE Change Compared to Previous	Compared to Equal	Compared to Original	White Group Compared to Original	White Group Compared to Equal	Asian Group Compared to Original	Asian Group Compared to Equal
Original	6.4588	6.4241	6.4529	6.8593	0.1987							
Equal	6.7129	6.4547	6.9672	6.5739	0.2189	7.97%		7.97%	0.48%		−4.16%	
10%	6.7724	6.77	7.2787	6.5685	0.2988	4.47%	4.47%	12.80%	5.38%	4.88%	−4.24%	−0.08%
20%	6.8741	6.8916	7.8781	6.2144	0.6831	8.23%	13.07%	22.09%	7.28%	6.77%	−9.40%	−5.47%
30%	6.7834	6.7537	7.2087	6.9303	0.1872	−8.50%	3.47%	11.71%	5.13%	4.63%	1.04%	5.42%
40%	7.1176	7.0839	7.7754	7.2036	0.3017	7.86%	11.60%	20.49%	10.27%	9.75%	5.02%	9.58%
50%	6.8294	6.845	7.7236	6.2425	0.6081	−0.67%	10.86%	19.69%	6.55%	6.05%	−8.99%	−5.04%
60%	6.7788	6.764	6.7813	6.9478	0.0828	−12.20%	−2.67%	5.09%	5.29%	4.79%	1.29%	5.69%
70%	6.7632	6.7431	7.1447	6.8198	0.1740	5.36%	2.55%	10.72%	4.97%	4.47%	−0.58%	3.74%
80%	6.7672	6.762	7.7404	6.3828	0.5719	8.34%	11.10%	19.95%	5.26%	4.76%	−6.95%	−2.91%
90%	6.8216	6.8529	7.2218	6.2795	0.3877	−6.70%	3.65%	11.92%	6.67%	6.17%	−8.45%	−4.48%
100%	6.8285	6.7755	8.1289	6.8429	0.6227	12.56%	16.67%	25.97%	5.47%	4.97%	−0.24%	4.09%

Interestingly, the Black group’s performance appears more resilient compared to the Equal dataset, with MAE increases averaging around 10% across different tests, peaking at a 16.97% increase when all samples are removed during testing. Regarding how other groups react to reductions in Black group samples, there is not a clear correlation. The White group’s MAE increase ranges from 4.97% to 10.27% and does not consistently show higher increases when more Black group samples are removed. The Asian group shows a slightly more reactive response to the removal of Black group samples compared to White group samples, albeit not significantly. Removing Black group samples actually improves Asian group performance, with MAE decreases ranging from −0.58% to −9.40% compared to the original dataset when 20% of Black group samples are removed.

Regarding the Asian group, as reflected in Table 6, it appears to be the least reactive to sample reductions. Interestingly, removing 70% of its samples results in a decrease of 9.40% in MAE compared to the original dataset performance, suggesting improved accuracy in some cases. Only when all samples are removed does the MAE increase by 9.98% compared to the original dataset and by 14.76% compared to the equal model. In contrast, the White group shows little reaction to reductions in Asian group samples, with MAE increases ranging consistently between 5% and 10%, regardless of the number of Asian samples removed.

Table 6. Performance changes upon APPA-REAL Asian group reduction.

Dataset Variations	Overall MAE	White Group MAE	Black Group MAE	Asian Group MAE	Standard Deviation	MAE Change Compared to Previous	Compared to Equal	Compared to Original	White Group Compared to Original	White Group Compared to Equal	Black Group Compared to Original	Black Group Compared to Equal
Original	6.4588	6.4241	6.4529	6.8593	0.1987							
Equal	6.7129	6.4547	6.9672	6.5739	0.2189	−4.16%		−4.16%	0.48%		7.97%	
10%	6.6531	6.6385	7.2428	6.5518	0.3073	−0.34%	−0.34%	−4.48%	3.34%	2.85%	12.24%	3.96%
20%	6.7844	6.7327	7.733	6.9448	0.4303	6.00%	5.64%	1.25%	4.80%	4.31%	19.84%	10.99%
30%	6.8468	6.8523	7.4678	6.4999	0.3999	−6.41%	−1.13%	−5.24%	6.67%	6.16%	15.73%	7.19%
40%	6.7736	6.731	7.5072	6.927	0.3295	6.57%	5.37%	0.99%	4.78%	4.28%	16.34%	7.75%
50%	6.6626	6.6547	7.1479	6.5305	0.2666	−5.72%	−0.66%	−4.79%	3.59%	3.10%	10.77%	2.59%
60%	7.1463	7.1156	8.0009	7.1076	0.4192	8.84%	8.12%	3.62%	10.76%	10.24%	23.99%	14.84%
70%	6.8689	6.9326	6.7004	6.2145	0.2991	−12.57%	−5.47%	−9.40%	7.92%	7.40%	3.84%	−3.83%
80%	6.7975	6.7792	8.0498	6.4349	0.6944	3.55%	−2.11%	−6.19%	5.53%	5.03%	24.75%	15.54%
90%	6.9338	6.956	6.9148	6.6874	0.1181	3.92%	1.73%	−2.51%	8.28%	7.77%	7.16%	−0.75%
100%	6.8969	6.8269	7.2377	7.5441	0.2938	12.81%	14.76%	9.98%	6.27%	5.77%	12.16%	3.88%

Conversely, the Black group exhibits more noticeable reactions, experiencing MAE increases ranging from 3.84% to 24.75% when 80% of Asian group samples are removed compared to the original dataset performance. Interestingly, the Black group’s performance remains more stable compared to the equal oversampled dataset, suggesting that oversampling generally helps stabilize performance.

The UTK Face dataset, while more balanced initially compared to the APPA REAL dataset with a higher representation of Black and Asian groups, surprisingly does not achieve equitable performance across groups as effectively as models trained on the APPA REAL dataset. Overall, MAE is lower on the UTK dataset due to the larger number of samples available for training. The top-performing model remains the one trained on the original dataset, achieving an overall MAE of 4.89, followed closely by the equally oversampled dataset at an MAE of 4.98.

Notably, the standard deviation across groups in the original dataset is 0.74. Contrary to expectation, the equally oversampled dataset shows increased variation with a standard deviation of 0.85, ranking 18th in terms of variation compared to the original dataset’s 5th place. This underscores that mere equalization of dataset proportions is insufficient for achieving balanced performance across demographic groups.

The model demonstrating the least deviation across groups involves removing 90% of Asian group samples during training, resulting in a standard deviation of 0.30. Although this model exhibits a higher overall MAE of 5.48 compared to the original (a 12.06% increase) and the equally oversampled (a 10.04% increase) models, it significantly reduces performance variation across groups by 59.45% compared to the original and by 64.70% compared to the equally oversampled model. Clearly, the benefits of reduced group-wise variation outweigh the slight increase in overall MAE.

Interestingly, the top four models in terms of variation reduction all involve some reduction of Asian group samples during training, demonstrating that models perform better as more Asian samples are removed. This approach also proved effective on the APPA REAL dataset, ranking third in terms of group variation with a standard deviation of 0.11.

Conversely, models that removed all samples from the White group during training exhibited the highest variation (1.36), mirroring findings in the APPA REAL dataset. The lowest five ranks also shared similarities with APPA REAL, with the complete removal of Black group samples resulting in the fourth worst variation on the UTK dataset, akin to its fifth place in APPA REAL.

Throughout our tests on the UTK dataset, a consistent trend emerged: the White group typically exhibited the highest MAE (~5), closely followed by the Black group (~5), slightly lower than the White group. Surprisingly, despite its initially lower sample count, the Asian group consistently performed the best with an MAE of around 3. However, it was the performance degradation of this group that ultimately led to more equitable performance across all groups.

For a detailed overview of the experiment results on the UTK Face dataset sorted by standard deviation, refer to Table 7.

Table 7. Comparison of all variations on the UTKFace dataset sorted by the lowest standard deviation between groups’ performance.

Dataset Variations	Overall MAE	White Group MAE	Black Group MAE	Asian Group MAE	Standard Deviation	DI Black Group	EoO Black Group	DI Asian Group	EoO Asian Group
Asian 90%	5.4808	5.7351	5.4583	5.0036	0.3015	0.9517	0.9518	0.8724	0.8725
Asian 100%	5.3499	5.6983	5.2564	4.7976	0.3677	0.9225	0.9226	0.8417	0.8418
Asian 80%	5.1197	5.5972	5.2153	4.0006	0.6807	0.9318	0.9319	0.7147	0.7148
Asian 70%	5.178	5.5635	5.2437	3.908	0.7170	0.9425	0.9426	0.7025	0.7026
Original	4.8925	5.3772	5.0445	3.6674	0.7401	0.9381	0.9382	0.6820	0.6821
Asian 60%	5.1361	5.7023	5.2071	3.877	0.7707	0.9132	0.9133	0.6799	0.6800

Table 7. Cont.

Dataset Variations	Overall MAE	White Group MAE	Black Group MAE	Asian Group MAE	Standard Deviation	DI Black Group	EoO Black Group	DI Asian Group	EoO Asian Group
Black 10%	4.9504	5.4966	5.072	3.6505	0.7894	0.9227	0.9228	0.6643	0.6644
White 10%	5.028	5.5884	5.1332	3.7261	0.7926	0.9185	0.9186	0.6670	0.6671
Asian 40%	5.0608	5.6095	5.1833	3.7543	0.7934	0.9240	0.9241	0.6683	0.6684
White 20%	5.0549	5.6	5.1896	3.7361	0.7996	0.9267	0.9268	0.6672	0.6673
Asian 50%	5.2064	5.8149	5.2504	3.9063	0.8005	0.9030	0.9031	0.6717	0.6718
Black 50%	5.1155	5.5867	5.3529	3.7797	0.8024	0.9581	0.9582	0.6764	0.6765
Asian 20%	4.9849	5.5227	5.1505	3.6308	0.8183	0.9325	0.9326	0.6575	0.6576
Black 30%	5.2779	5.821	5.4611	3.8847	0.8408	0.9381	0.9382	0.6678	0.6679
Asian 30%	5.088	5.6537	5.2456	3.6904	0.8458	0.9277	0.9278	0.6528	0.6529
Black 60%	5.1542	5.6783	5.3728	3.7417	0.8501	0.9462	0.9463	0.6591	0.6592
White 40%	5.1123	5.7115	5.2279	3.7152	0.8503	0.9154	0.9155	0.6505	0.6506
Equal	4.9879	5.5465	5.1663	3.5711	0.8557	0.9315	0.9316	0.6439	0.6440
White 50%	5.152	5.7823	5.2317	3.7499	0.8582	0.9047	0.9048	0.6484	0.6485
Asian 10%	5.0262	5.6409	5.1295	3.6179	0.8588	0.9094	0.9095	0.6417	0.6418
White 30%	5.0891	5.6989	5.2045	3.6709	0.8634	0.9132	0.9133	0.6444	0.6445
Black 70%	5.1646	5.5988	5.5342	3.6896	0.8851	0.9885	0.9886	0.6581	0.6582
Black 20%	5.0512	5.6157	5.2732	3.5516	0.9031	0.9389	0.9390	0.6323	0.6324
Black 90%	5.2389	5.6447	5.6612	3.7357	0.9038	1.0029	1.0030	0.6618	0.6619
Black 80%	5.3513	5.8564	5.6689	3.8171	0.9203	0.9679	0.9680	0.6528	0.6529
White 60%	5.1887	5.8597	5.2998	3.6538	0.9362	0.9043	0.9044	0.6234	0.6235
White 70%	5.3421	6.0802	5.3782	3.7933	0.9565	0.8846	0.8847	0.6239	0.6240
Black 40%	5.3027	5.92	5.535	3.6802	0.9778	0.9350	0.9351	0.6215	0.6216
Black 100%	5.3681	5.7606	5.9055	3.7052	1.0048	1.0252	1.0253	0.6432	0.6433
White 80%	5.3377	6.1819	5.2886	3.7123	1.0209	0.8555	0.8556	0.6005	0.6006
White 90%	5.5117	6.4859	5.3614	3.7876	1.1066	0.8266	0.8267	0.5841	0.5842
White 100%	5.8152	7.1516	5.3912	3.803	1.3676	0.7536	0.7537	0.5316	0.5317

For better clarity of the changes in model performance, we have included Figure 7 which shows the overall MAE and standard deviation for different dataset variations.

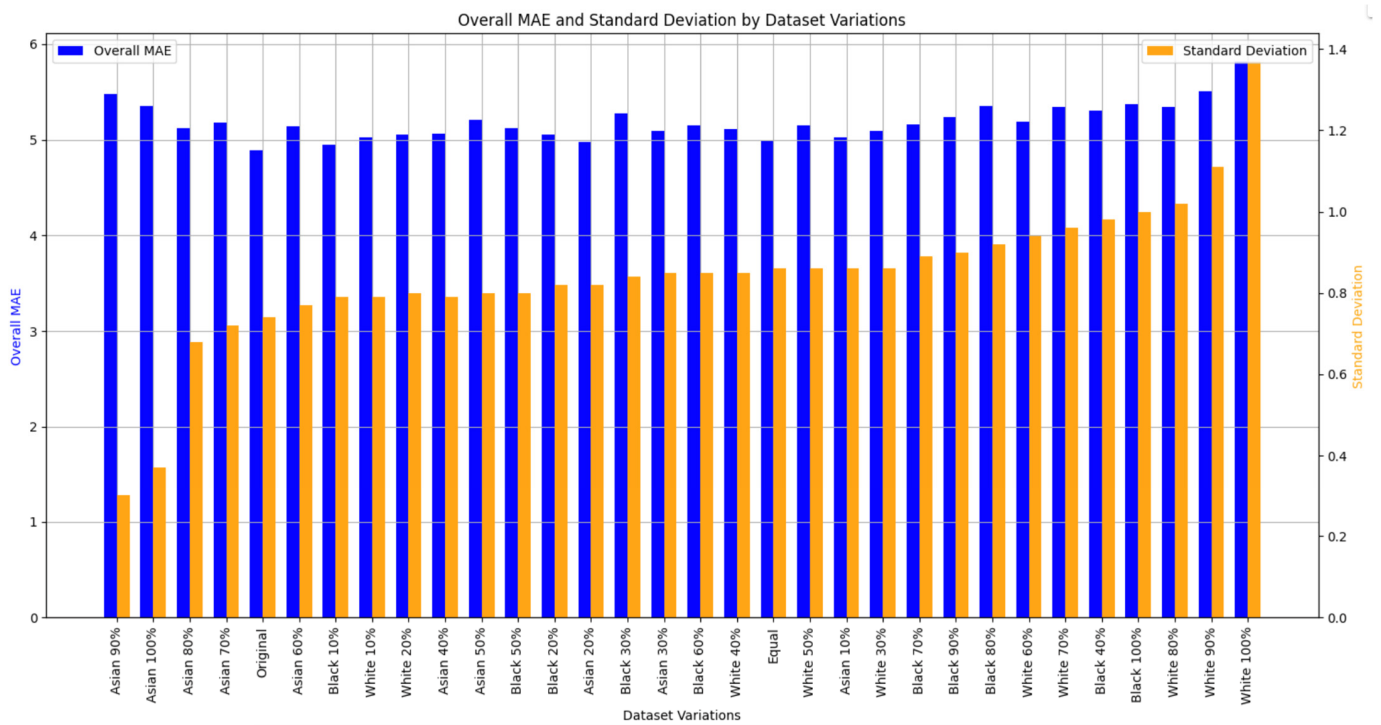


Figure 7. Overall MAE and Standard Deviation by variations of the UTK Face dataset.

Figure 8 provides a detailed view of MAE and standard deviation trends for different groups and overall.

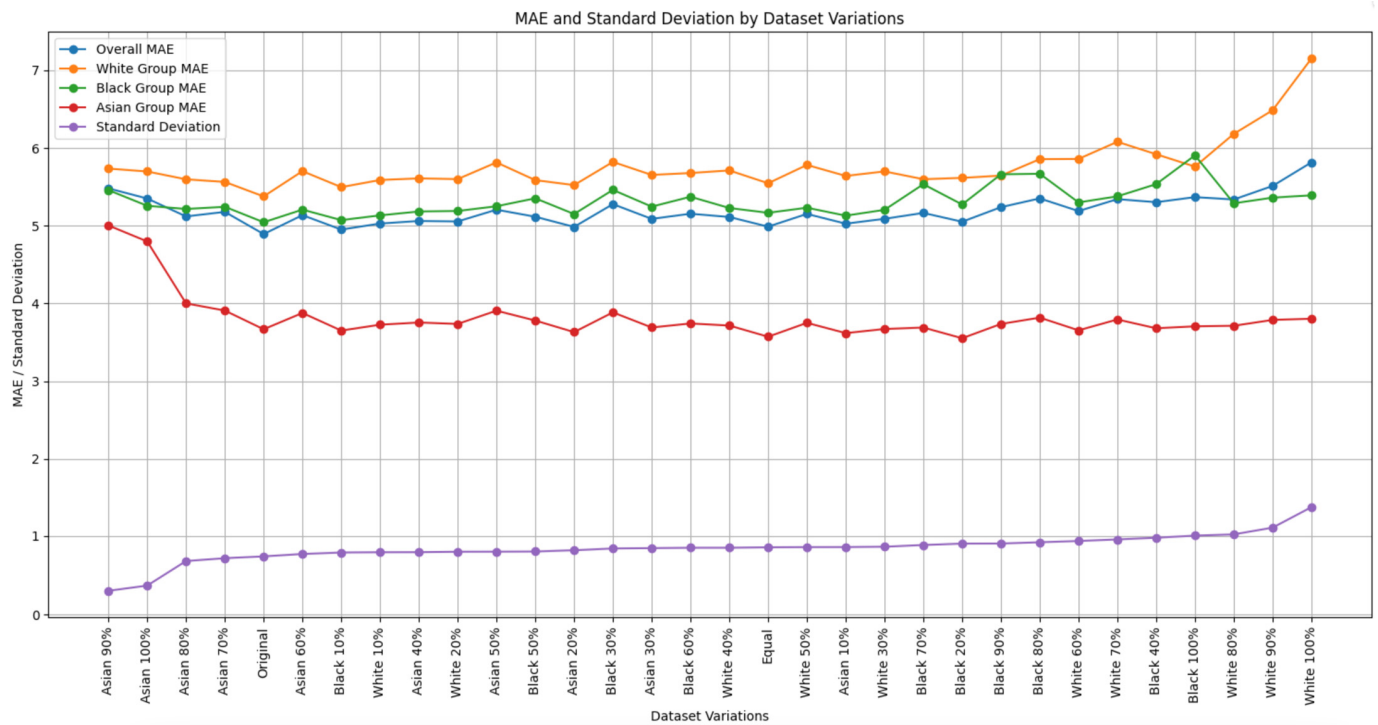


Figure 8. MAE and Standard Deviation by variations of the UTK Face dataset.

For the UTK Face dataset, the Asian 90% variation shows DI values of 0.9517 for the Black group and 0.8724 for the Asian group, suggesting that this variation achieves nearly fair treatment across these groups. Similarly, the Asian 100% variation presents DI values of 0.9225 for the Black group and 0.8417 for the Asian group, indicating minimal disparate impact and promoting fairness.

In contrast, the Equal Dataset Variation, despite equalizing the sample sizes, has DI values of 0.9315 for the Black group and 0.6439 for the Asian group, reflecting some bias. The worst performing model in terms of DI is the White 100% variation, with DI values of 0.7536 for the Black group and 0.5316 for the Asian group, indicating significant bias. Figure 9 shows the Disparate Impact for the Black and Asian groups across various dataset variations.

Equality of Opportunity (EoO) measures whether different groups have the same chances of achieving a favorable outcome. Like DI, an EoO value close to 1 indicates equitable performance across groups.

The Asian 90% variation shows EoO values of 0.9518 for the Black group and 0.8725 for the Asian group, demonstrating that this model provides nearly equal opportunities to these groups. The Asian 100% variation presents EoO values of 0.9226 for the Black group and 0.8418 for the Asian group, also indicating good equality of opportunity.

Conversely, the Equal Dataset Variation shows EoO values of 0.9316 for the Black group and 0.6440 for the Asian group, reflecting some inequity. The White 100% variation has the worst EoO values at 0.7537 for the Black group and 0.5317 for the Asian group, highlighting substantial inequity. Figure 10 shows Equality of Opportunity values for the Black and Asian groups across dataset compositions.

Turning our attention to the performance of each group individually, we begin with the White group. As depicted in Table 8, a noticeable trend emerges where a 10% decrease in sample size does not necessarily translate to a 10% decline in group performance. Instead, we observe performance changes typically ranging between 1% and 3%. The most significant increase in MAE, approximately 10.26%, occurs only when reducing White group samples from a 90% cut to complete omission.

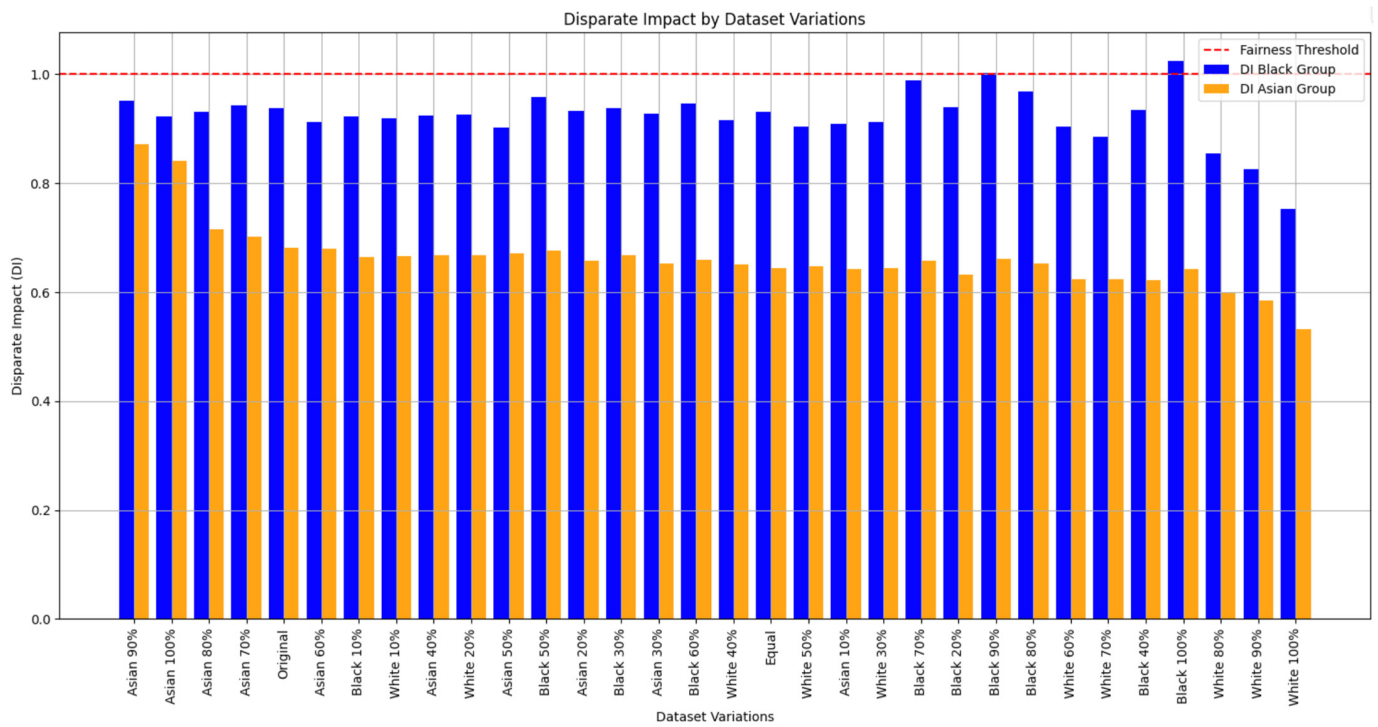


Figure 9. Disparate Impact by variations of the UTK Face dataset.

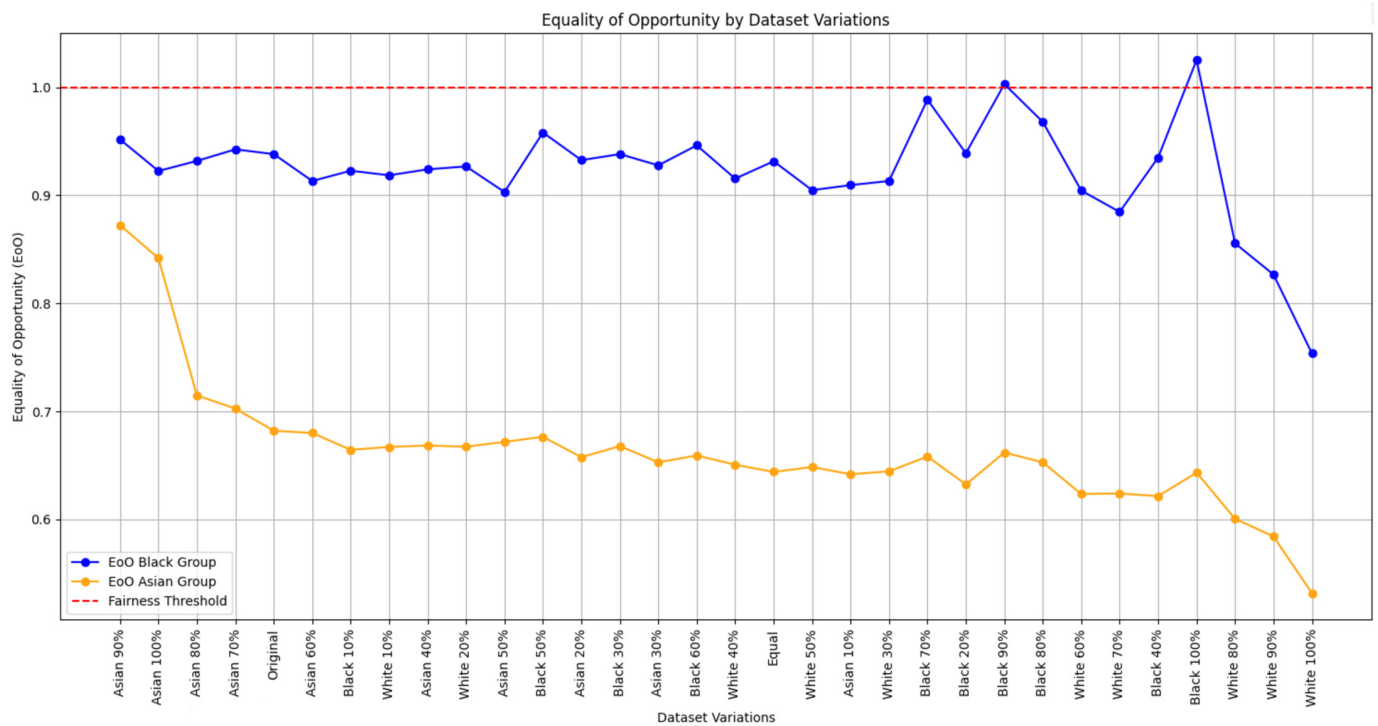


Figure 10. Equality of Opportunity by variations of the UTK Face dataset.

Comparing these results with the equal and original datasets, similar to the APPA REAL dataset findings, the White group exhibits a more subdued reaction to sample reductions in the equal dataset compared to the original. The highest MAE increase, compared to the equal dataset, reaches 28.94%, and compared to the original dataset, it reaches 33%.

Table 8. Performance changes upon UTKFace White group reduction.

Dataset Variations	Overall MAE	White Group MAE	Black Group MAE	Asian Group MAE	Standard Deviation	MAE Change Compared to Previous	Compared to Equal	Compared to Original	Black Group Compared to Original	Black Group Compared to Equal	Asian Group Compared to Original	Asian Group Compared to Equal
Original	4.8925	5.3772	5.0445	3.6674	0.7401							
Equal	4.9879	5.5465	5.1663	3.5711	0.8557	3.15%		3.15%	2.41%		−2.63%	
10%	5.028	5.5884	5.1332	3.7261	0.7926	0.76%	0.76%	3.93%	1.76%	−0.64%	1.60%	4.34%
20%	5.0549	5.6	5.1896	3.7361	0.7996	0.21%	0.96%	4.14%	2.88%	0.45%	1.87%	4.62%
30%	5.0891	5.6989	5.2045	3.6709	0.8634	1.77%	2.75%	5.98%	3.17%	0.74%	0.10%	2.79%
40%	5.1123	5.7115	5.2279	3.7152	0.8503	0.22%	2.97%	6.22%	3.64%	1.19%	1.30%	4.04%
50%	5.152	5.7823	5.2317	3.7499	0.8582	1.24%	4.25%	7.53%	3.71%	1.27%	2.25%	5.01%
60%	5.1887	5.8597	5.2998	3.6538	0.9362	1.34%	5.65%	8.97%	5.06%	2.58%	−0.37%	2.32%
70%	5.3421	6.0802	5.3782	3.7933	0.9565	3.76%	9.62%	13.07%	6.62%	4.10%	3.43%	6.22%
80%	5.3377	6.1819	5.2886	3.7123	1.0209	1.67%	11.46%	14.97%	4.84%	2.37%	1.22%	3.95%
90%	5.5117	6.4859	5.3614	3.7876	1.1066	4.92%	16.94%	20.62%	6.28%	3.78%	3.28%	6.06%
100%	5.8152	7.1516	5.3912	3.803	1.3676	10.26%	28.94%	33.00%	6.87%	4.35%	3.70%	6.49%

Interestingly, akin to observations from the APPA REAL dataset, the White group shows notable sensitivity to reductions in its sample size during training. Conversely, other groups exhibit less pronounced reactions to changes in the White group’s sample composition. The Black group’s performance worsens by approximately 1% to 4%, irrespective of the percentage of White samples removed. Similarly, the Asian group shows a consistent performance decline of around 6%, regardless of whether 20% or 90% of White group samples are removed. This pattern suggests, similar to findings in the APPA REAL dataset, that the Asian group remains relatively unaffected by such changes.

However, in contrast to the APPA-REAL dataset, we observe a weaker correlation in performance between the White and Black groups in this dataset analysis.

Examining the performance of the Black group, detailed in Table 9, reveals a clear trend where increasing sample removal correlates with higher MAE. The highest MAE increase occurs when all Black group samples are excluded during training, resulting in a 17.07% increase compared to the original performance. Interestingly, this increase is slightly lower at 14.31% when compared to the equal model performance, indicating that oversampling benefits the Black group by stabilizing its performance.

Table 9. Performance changes upon UTKFace Black group reduction.

Dataset Variations	Overall MAE	White Group MAE	Black Group MAE	Asian Group MAE	Standard Deviation	MAE Change Compared to Previous	Compared to Equal	Compared to Original	White Group Compared to Original	White Group Compared To Equal	Asian Group Compared to Original	Asian Group Compared to Equal
Original	4.8925	5.3772	5.0445	3.6674	0.7401							
Equal	4.9879	5.5465	5.1663	3.5711	0.8557	2.41%		2.41%	3.15%		−2.63%	
10%	4.9504	5.4966	5.072	3.6505	0.7894	−1.83%	−1.83%	0.55%	2.22%	−0.90%	−0.46%	2.22%
20%	5.0512	5.6157	5.2732	3.5516	0.9031	3.97%	2.07%	4.53%	4.44%	1.25%	−3.16%	−0.55%
30%	5.2779	5.821	5.4611	3.8847	0.8408	3.56%	5.71%	8.26%	8.25%	4.95%	5.93%	8.78%
40%	5.3027	5.92	5.535	3.6802	0.9778	1.35%	7.14%	9.72%	10.09%	6.73%	0.35%	3.06%
50%	5.1155	5.5867	5.3529	3.7797	0.8024	−3.29%	3.61%	6.11%	3.90%	0.72%	3.06%	5.84%
60%	5.1542	5.6783	5.3728	3.7417	0.8501	0.37%	4.00%	6.51%	5.60%	2.38%	2.03%	4.78%
70%	5.1646	5.5988	5.5342	3.6896	0.8851	3.00%	7.12%	9.71%	4.12%	0.94%	0.61%	3.32%
80%	5.3513	5.8564	5.6689	3.8171	0.9203	2.43%	9.73%	12.38%	8.91%	5.59%	4.08%	6.89%
90%	5.2389	5.6447	5.6612	3.7357	0.9038	−0.14%	9.58%	12.23%	4.97%	1.77%	1.86%	4.61%
100%	5.3681	5.7606	5.9055	3.7052	1.0048	4.32%	14.31%	17.07%	7.13%	3.86%	1.03%	3.76%

Similar to observations with the White group, a 10% reduction in sample size does not linearly equate to a 10% increase in MAE for the Black group. Instead, we observe a high single-digit increase, with a 14.31% increase when moving from a 90% reduction to complete removal of Black group samples.

In contrast, other groups show minimal reaction to changes in the Black group’s sample size. The White group’s MAE increases by approximately 1% to 6%, regardless of the percentage of Black group samples removed. Similarly, the Asian group exhibits varying MAE increases of 2% to 8%, showing a consistent pattern with the findings from the APPA REAL dataset tests for the Black group.

Turning to the Asian group, as detailed in Table 10, we observe a pattern similar to that seen in the results from the APPA-REAL dataset. Reductions in sample size generally result in minor fluctuations in MAE, typically ranging from a slight decrease to an increase of 1–4%. An exception occurs when reducing samples from 80% to 90%, where we see a significant jump in MAE by 25.07%. Surprisingly, the performance change between 90% of samples removed and 100% is actually a 4.12% improvement in performance, indicating a complex relationship between sample size and performance for the Asian group.

Table 10. Performance changes upon UTKFace Asian group reduction.

Dataset Variations	Overall MAE	White Group MAE	Black Group MAE	Asian Group MAE	Standard Deviation	MAE Change Compared to Previous	Compared to Equal	Compared to Original	White Group Compared to Original	White Group Compared to Equal	Black Group Compared to Original	Black Group Compared to Equal
Original	4.8925	5.3772	5.0445	3.6674	0.7401							
Equal	4.9879	5.5465	5.1663	3.5711	0.8557	−2.63%		−2.63%	3.15%		2.41%	
10%	5.0262	5.6409	5.1295	3.6179	0.8588	1.31%	1.31%	−1.35%	4.90%	1.70%	1.69%	−0.71%
20%	4.9849	5.5227	5.1505	3.6308	0.8183	0.36%	1.67%	−1.00%	2.71%	−0.43%	2.10%	−0.31%
30%	5.088	5.6537	5.2456	3.6904	0.8458	1.64%	3.34%	0.63%	5.14%	1.93%	3.99%	1.53%
40%	5.0608	5.6095	5.1833	3.7543	0.7934	1.73%	5.13%	2.37%	4.32%	1.14%	2.75%	0.33%
50%	5.2064	5.8149	5.2504	3.9063	0.8005	4.05%	9.39%	6.51%	8.14%	4.84%	4.08%	1.63%
60%	5.1361	5.7023	5.2071	3.877	0.7707	−0.75%	8.57%	5.72%	6.05%	2.81%	3.22%	0.79%
70%	5.178	5.5635	5.2437	3.908	0.7170	0.80%	9.43%	6.56%	3.46%	0.31%	3.95%	1.50%
80%	5.1197	5.5972	5.2153	4.0006	0.6807	2.37%	12.03%	9.09%	4.09%	0.91%	3.39%	0.95%
90%	5.4808	5.7351	5.4583	5.0036	0.3015	25.07%	40.11%	36.43%	6.66%	3.40%	8.20%	5.65%
100%	5.3499	5.6983	5.2564	4.7976	0.3677	−4.12%	34.35%	30.82%	5.97%	2.74%	4.20%	1.74%

Comparing these outcomes with the original and equal dataset performances, the Asian group shows a response pattern akin to the White group. Removing 100% of Asian group samples results in a 34.35% increase in MAE compared to the equal dataset and a 30.82% increase compared to the original dataset. This reaction is notably more pronounced compared to findings from the APPA REAL dataset.

Regarding the reactions of other groups to reductions in Asian group sample size, we observe consistent increases in MAE ranging from 1% to 8%, regardless of the extent of Asian sample reductions. This contrasts somewhat with findings from the APPA REAL dataset, where the Black group exhibited more significant MAE increases that were not strictly correlated with reductions in Asian sample size. The most substantial reaction in the APPA REAL dataset saw a 24.75% increase in MAE for the Black group.

To provide a comprehensive evaluation of our model’s performance, we compare our results with those from other related works, focusing on overall MAE, race-specific accuracies, and standard deviation. While previous studies on age estimation have reported different accuracies for various ethnicities, none have explicitly addressed the disparities in performance across different demographic groups. Table 11 below presents a summary of our findings alongside related work.

As observed, our model achieves a significantly lower overall MAE and standard deviation on both the UTKFace and APPA-REAL datasets compared to previous studies. This indicates a more balanced performance across different ethnic groups, emphasizing the effectiveness of our approach in addressing performance disparities.

Table 11. Comparison of our results with related work.

Paper	Dataset	Overall MAE	Race			Standard Deviation
			White	Black	Asian	
[11]	UTKFace		9.79	7.71	9.56	0.931
[11]	APPA-REAL		7.79	8.23	7.85	0.1948
[18]	APPA-REAL	7.356	7.40	7.73	6.59	0.4789
[16]	APPA-REAL	13.5774	13.6386	14.1607	12.4729	0.7055
Our results	UTKFace	5.4808	5.7351	5.4583	5.0036	0.3015
Our results	APPA-REAL	7.1603	7.1676	7.1663	7.0735	0.044

4. Discussion

In this study, we investigated the impact of dataset composition on the performance of age estimation models, focusing on mitigating bias across different ethnic groups. We employed a transfer learning approach, utilizing pre-trained CNN models (VGG16, VGG19, ResNet50, and MobileNetV2) and fine-tuning them on the UTKFace and APPA-REAL datasets, chosen for their demographic diversity and inclusion of relevant labels, such as real age and ethnicity.

Our methodology involved systematically manipulating the dataset composition by oversampling minority groups to match the majority group and then gradually reducing the sample size of each group. This allowed us to analyze the relationship between dataset composition and model performance, both overall and for specific ethnic groups. We used Mean Absolute Error (MAE) and standard deviation as our primary evaluation metrics.

Our findings reveal that simply balancing the dataset by oversampling minority groups does not necessarily lead to equitable performance across ethnicities. This aligns with the observations of Puc et al. (2020) [11], who found performance discrepancies across different racial groups in age estimation models but did not actively manipulate the datasets to mitigate these biases. In contrast, our research demonstrated that varying the number of samples from the majority group (White) can lead to mixed results. In some cases, it led to a more balanced performance across ethnic groups, as indicated by a lower standard deviation of MAE, while in other cases, it did not significantly improve fairness or overall performance. This suggests that oversampling and undersampling may not always be effective strategies on their own for mitigating bias, and a more nuanced approach is needed.

For the UTKFace dataset, the model trained on the original dataset composition achieved an overall MAE of 4.89 with a standard deviation of 0.74. Reducing the sample size of the Asian group by 90% resulted in the smallest standard deviation (0.30) and an overall MAE of 5.48. This reduction improved performance consistency compared to the original dataset. Conversely, reducing samples from other groups did not always yield better results, underscoring the complexity of achieving balanced performance across all demographics.

The APPA-REAL dataset, on the other hand, exhibited different trends. The original dataset composition led to an overall MAE of 6.45 and a standard deviation of 0.20. Reducing the White group to 20% achieved the most balanced performance, with a standard deviation of 0.04 and an overall MAE of 7.16. This indicates that reducing the representation of the majority group can enhance fairness across different ethnic groups. The worst-performing model omitted the White group entirely, resulting in an overall MAE of 8.89 and a standard deviation of 0.94, demonstrating that complete exclusion of a dominant group adversely affects model performance.

These findings highlight the ability of our approach to systematically uncover and address biases in age estimation models by manipulating dataset composition. Our method provides a clear framework for analyzing and improving model fairness.

We were intrigued by the distinct performance of the Asian group within the UTK Face dataset compared to other groups, consistently showing a smaller MAE by approximately

2 points. To investigate this further, we conducted an analysis using our trained models, evaluating example images from all groups within both the UTK dataset and the APPA REAL dataset. Our findings, illustrated in Figure 11, reveal an interesting observation.

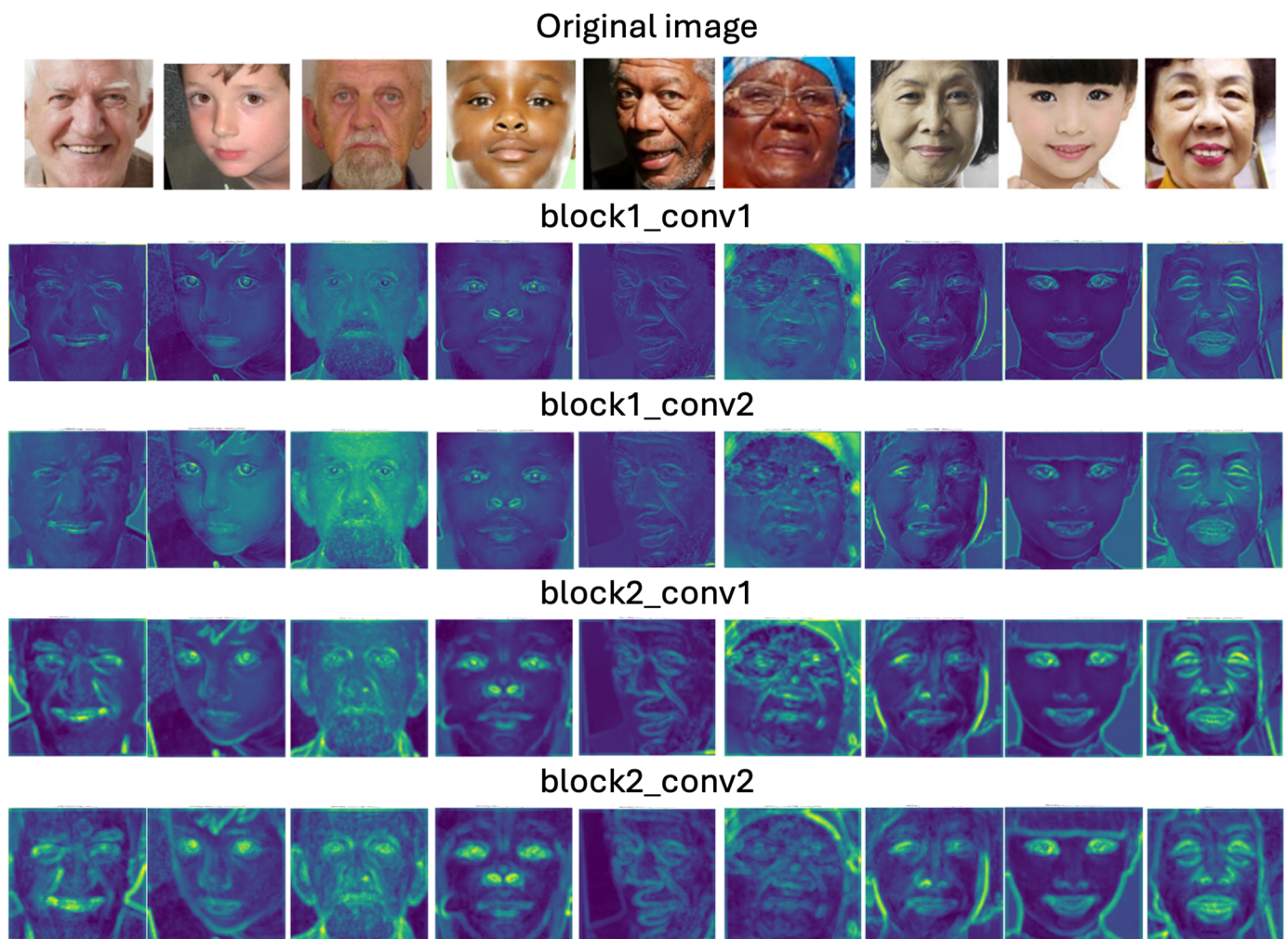


Figure 11. Example images from the UTKFace dataset and their feature maps through layers.

When examining examples from the UTK Face dataset, the model demonstrates a preference for certain facial features across different groups. For the White and Black groups, there is a notable focus on the eyes, nose, and mouth, with stronger activations within these regions. Additionally, some emphasis is placed on outlining the edges of the face. In contrast, activations for the Asian group predominantly concentrate on the edges of the face, particularly around the shape of the cheeks, in addition to the mouth, nose, and eyes. While the model analyzes similar facial features across all groups, the intensity and distribution of these activations vary.

This distinction in activation patterns may help explain why the Asian group shows different performance characteristics compared to the White and Black groups within the UTK dataset. In contrast, the performance of the White and Black groups appears more aligned across various tests.

Now, let us explore why we do not observe the same relationship in the APPA REAL results. This can be explained by examining the feature maps. As shown in Figure 12, the feature maps of the APPA REAL dataset reveal that the model exhibits a consistent pattern of activations across all ethnic groups. Significant activations are observed within the face, particularly highlighting the cheeks, smile, eyes, and nose, along with some background details. This uniform pattern of feature extraction across different ethnicities

likely contributes to the relatively equal performance results observed across these groups in the APPA REAL dataset.

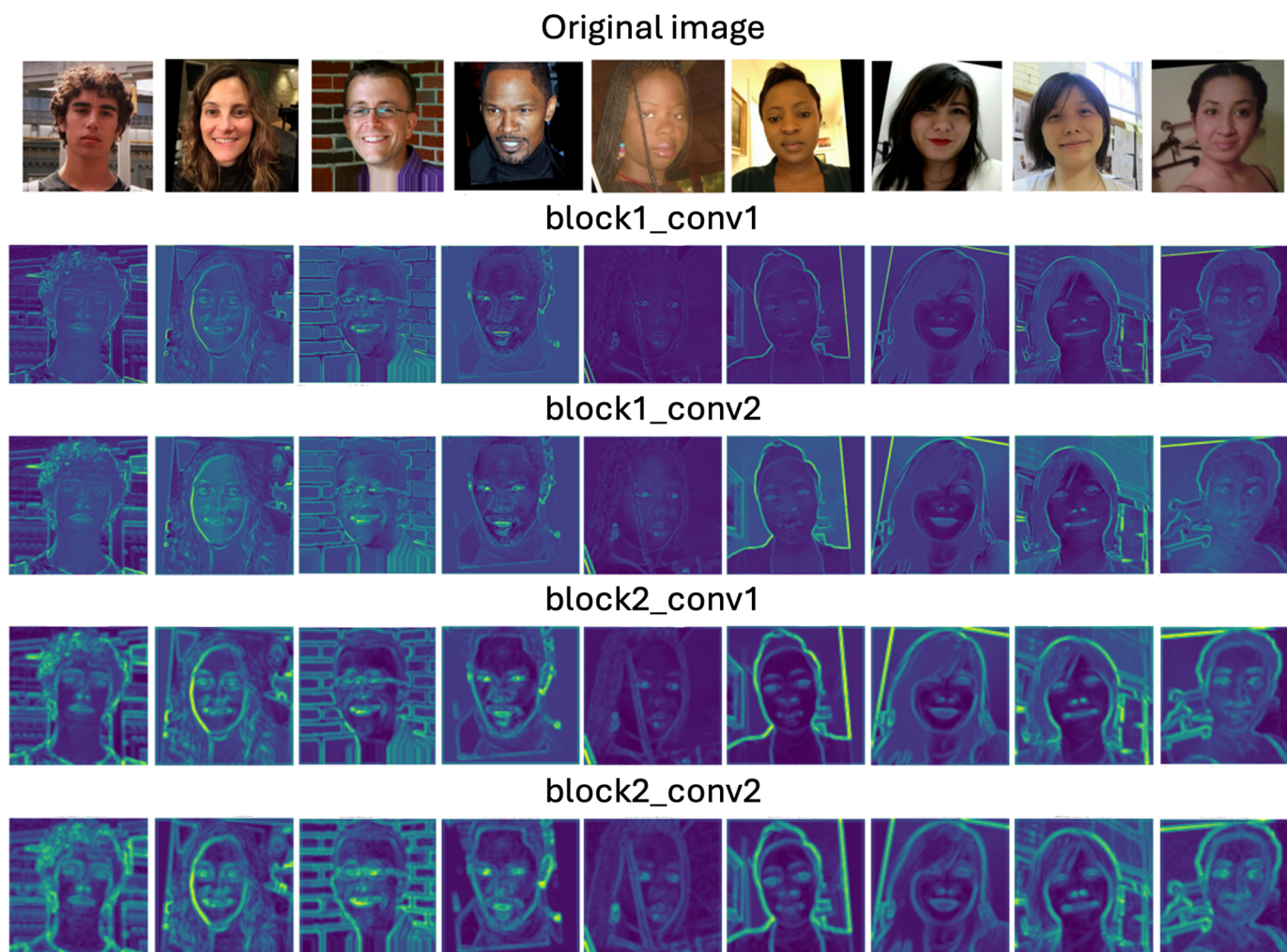


Figure 12. Example images from the APPA-REAL dataset and their feature maps through layers.

This difference in feature detection and model performance can be attributed to several factors. One possibility is that the more visible background in APPA REAL images leads the model to follow a different path in recognizing images, thereby treating them more uniformly. By analyzing the feature maps, we gain valuable insights into how the model processes images from different datasets and why there might be variations in performance across different ethnic groups.

The analysis of feature maps provided further insights into the model's behavior. For the UTKFace dataset, we observed distinct activation patterns for the Asian group compared to the White and Black groups, which could explain the performance differences. This finding aligns with Abdolrashidi et al. (2020) [13], who highlighted challenges in age prediction due to intra-class variations. In contrast, the APPA-REAL dataset showed consistent activation patterns across all ethnic groups, contributing to the more balanced performance observed in this dataset.

These results highlight the complexity of achieving equal performance among different demographic groups. It is clear that equal performance among groups or classes is a very complex problem without a simple solution. Oversampling the dataset to make distributions even should not be the only step taken; it is just an introduction to more granular processing that needs to be performed. Although we have seen some patterns shared between the two datasets and how performance among groups varies in different

scenarios, the reactions are not identical. One dataset composition that works for one dataset may not work for another due to various factors, such as lighting and contrast.

Our findings suggest that a balanced approach, incorporating both undersampling of the majority class and oversampling of minority groups, may be more effective in mitigating bias than solely relying on oversampling. The minority classes in our datasets have significantly fewer samples compared to the majority group, necessitating oversampling to balance the datasets. This disparity presents challenges, and despite using data augmentation to reduce overfitting, having larger datasets with equal group sizes from the start would be far more beneficial. This would eliminate the need to generate synthetic data, thereby avoiding the introduction of potential noise and enhancing the model's reliability. While we demonstrated that equally sampled datasets do not automatically result in equal performance, working with unique samples from a balanced dataset would improve both model fairness and reliability.

The drawback to our method, as well as some other proposed methods, is that it is computationally expensive. It requires extensive testing of different combinations. A possible improvement could involve using a representative sample of the dataset to find the right combination of group ratios. Even though we cannot expect high accuracy at this point, the relationship between groups and how dataset adjustment affects equity among them should still be visible while reducing testing time since less data are involved. Following this, we could continue training with ensemble models on the tailored group ratios but now with a full dataset.

Our findings resonate with previous research, such as the extensive review by Hasib et al. (2020) [19], which addressed class imbalance in datasets. They categorized methodologies into data-level methods, algorithm-level methods, ensemble methods, and hybrid methods. Techniques like SMOTE and ADASYN, which generate synthetic data to balance class distributions, and undersampling techniques like RUS and T-Link were highlighted for their effectiveness. However, the increased computational cost and potential information loss during undersampling were noted as significant challenges. Our approach of systematically reducing samples from overrepresented groups and observing the impact on model performance revealed that while dataset rebalancing can reduce standard deviations and improve performance consistency across ethnicities, it does not always guarantee better overall performance. This observation is consistent with the conclusions of Ramyachitra and Manikandan (2014) [21], who noted that while techniques like SMOTE and cost-sensitive learning are effective, the choice of method should be tailored to the specific characteristics of the dataset.

Similarly, Rahman and Davis (2013) [20] investigated the performance of oversampling and undersampling techniques to balance cardiovascular data. Their findings emphasized that while SMOTE showed good classification outcomes, the modified cluster-based undersampling method outperformed traditional methods. This highlights the effectiveness of hybrid methods, which integrate data sampling and algorithm boosting to address class imbalance. Our findings support this approach, demonstrating that a combination of techniques can provide a more balanced performance across different ethnic groups.

Kotsiantis, Kanellopoulos, and Pintelas (2006) [22] also emphasized the need for a tailored approach based on dataset characteristics. They discussed the use of ensemble methods, such as boosting and bagging, to improve classification outcomes by combining multiple models. Our study supports this by demonstrating that a systematic approach to adjusting group representation can complement existing techniques and offer new perspectives in achieving fairness.

To apply the findings of our experiment to practical facial recognition systems, several key steps can be taken. Firstly, our approach of balancing datasets through a combination of oversampling minority groups and undersampling the majority class should be implemented. This ensures more equitable performance across different ethnic groups. Moreover, the insights gained from our feature map analysis, which identified varying

activation patterns across ethnic groups, can be used to fine-tune model architectures and improve their sensitivity to diverse facial features.

Real-world applications should prioritize creating and utilizing larger, more balanced datasets from the start to avoid the need for synthetic data and reduce potential noise. This will enhance model reliability and fairness. Additionally, employing representative samples of datasets to determine optimal group ratios can help streamline the process, reducing computational costs and testing times. Combining these data-level adjustments with advanced algorithmic techniques, such as ensemble methods, can further enhance model performance and ensure a more equitable application of facial recognition technology.

Possible improvements include combining our dataset rebalancing technique with advanced bias mitigation methods such as adversarial training or fairness constraints. Additionally, exploring more diverse and comprehensive datasets could further enhance the generalizability and fairness of the models. Future research should also consider the integration of socio-demographic factors to develop more nuanced models that can better account for the complexity of human faces.

In conclusion, our study demonstrates that while simple oversampling and under-sampling techniques can reduce bias to some extent, achieving true fairness requires a combination of nuanced dataset adjustments and sophisticated algorithmic methods. By systematically analyzing the impact of various dataset compositions, we provide valuable insights into developing fairer and more accurate facial recognition models.

Author Contributions: Conceptualization, N.P. and M.M.; Data curation, N.P.; Formal analysis, N.P.; Investigation, N.P.; Methodology, N.P. and M.M.; Project administration, M.M.; Resources, N.P. and M.M.; Software, N.P.; Supervision, M.M.; Validation, M.M. and T.B.; Visualization, N.P.; Writing—original draft, N.P.; Writing—review and editing, N.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Science Fund of the Republic of Serbia, Grant No. 7502, Intelligent Multi-Agent Control and Optimization applied to Green Buildings and Environmental Monitoring Drone Swarms—ECOSwarm.

Data Availability Statement: The data were derived from the following resources available in the public domain: <https://susanqq.github.io/UTKFace/> (accessed on 19 June 2024) and <https://chalearnlap.cvc.uab.cat/dataset/26/description/> (accessed on 19 June 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Michalski, D.; Yiu, S.Y.; Malec, C. The Impact of Age and Threshold Variation on Facial Recognition Algorithm Performance Using Images of Children. In Proceedings of the International Conference on Biometrics (ICB), Gold Coast, Australia, 20–23 February 2018; pp. 217–224.
2. Srinivas, N.; Ricanek, K.; Michalski, D.; Bolme, D.S.; King, M. Face recognition algorithm bias: Performance differences on images of children and adults. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA, 16–20 June 2019.
3. Albiero, V.; Bowyer, K.W. Is face recognition sexist? no, gendered hairstyles and biology are. *arXiv* **2020**, arXiv:2008.06989.
4. Albiero, V.; Zhang, K.; Bowyer, K.W. How does gender balance in training data affect face recognition accuracy. In Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), Houston, TX, USA, 28 September–1 October 2020; pp. 1–10.
5. Terhörst, P.; Kolf, J.N.; Huber, M.; Kirchbuchner, F.; Damer, N.; Moreno, A.M.; Fierrez, J.; Kuijper, A. A Comprehensive Study on Face Recognition Biases Beyond Demographics. *IEEE Trans. Technol. Soc.* **2022**, *3*, 16–30. [[CrossRef](#)]
6. Voigt, P.; Bussche, A.V.D. *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 1st ed.; Springer: Cham, Switzerland, 2017; pp. 141–187.
7. Albert, A.M.; Ricanek, K.; Patterson, E. A review of the literature on the aging adult skull and face: Implications for forensic science research and applications. *Forensic Sci. Int.* **2007**, *172*, 1–9. [[CrossRef](#)] [[PubMed](#)]
8. Angulu, R.; Tapamo, J.R.; Adewumi, A.O. Age estimation via face images: A survey. *J. Image Video Proc.* **2018**, *2018*, 42. [[CrossRef](#)]
9. Khaled, E.K.; Valliappan, R.; Patrick, T. Facial Age Estimation Using Machine Learning Techniques: An Overview. *Big Data Cogn. Comput.* **2022**, *6*, 128.

10. Age Detection Using Facial Images: Traditional Machine Learning vs. Deep Learning, towardsdatascience.com. Available online: <https://towardsdatascience.com/age-detection-using-facial-images-traditional-machine-learning-vs-deep-learning-2437b2feeab2> (accessed on 18 June 2024).
11. Andraz, P.; Vitomir, S.; Klemen, G. Analysis of Race and Gender Bias in Deep Age Estimation Models. In Proceedings of the 8th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–22 January 2021.
12. Kimmo, K.; Jungseock, J. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv* **2019**, arXiv:1908.04913.
13. Abdolrashidi, A.; Minaei, M.; Azimi, E. Age and Gender Prediction From Face Images Using Attentional Convolutional Network. *arXiv* **2020**, arXiv:2010.03791.
14. Sathyavathi, S.; Baskaran, K.R. An Intelligent Human Age Prediction from Face Image Framework Based on Deep Learning Algorithms. *Inf. Technol. Control* **2023**, *52*, 245–257. [[CrossRef](#)]
15. Amelia, J.S.; Wahyono. Age Estimation on Human Face Image Using Support Vector Regression and Text-Based Features. *Int. J. Adv. Comput. Sci. Appl.* **2022**. [[CrossRef](#)]
16. Clapes, A.; Bilici, O.; Temirova, D.; Avots, E.; Anbarjafari, G.; Escalera, S. From Apparent to Real Age: Gender, Age, Ethnic, Makeup, and Expression Bias Analysis in Real Age Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2373–2382.
17. Xing, J.; Li, K.; Hu, W.; Yuan, C.; Ling, H. Diagnosing deep learning models for high accuracy age estimation from a single image. *Pattern Recognit.* **2017**, *66*, 106–116. [[CrossRef](#)]
18. Jacques, J.C.S.; Ozcinar, C.; Marjanovic, M.; Baró, X.; Anbarjafari, G.; Escalera, S. On the effect of age perception biases for real age regression. In Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition, Lille, France, 14–18 May 2019; pp. 1–8.
19. Khan, H.; Iqbal, S.; Shah, F.M.; Mahmud, J.A.; Popel, M.H.; Showrow, I.H.; Ahmed, S.; Rahman, O. A Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem. *arXiv* **2020**, arXiv:2012.11870. [[CrossRef](#)]
20. Rahman, M.M.; Davis, D.N. Addressing the class imbalance problem in medical datasets. *Int. J. Mach. Learn. Comput.* **2013**, *3*, 224. [[CrossRef](#)]
21. Ramyachitra, D.; Manikandan, P. Imbalanced dataset classification and solutions: A review. *Int. J. Comput. Bus. Res.* **2014**, *5*, 1–29.
22. Kotsiantis, S.; Kanellopoulos, D.; Pintelas, P. Handling imbalanced datasets: A review. *GESTS Int. Trans. Comput. Sci. Eng.* **2006**, *30*, 25–36.
23. UTKFace, github.io. Available online: <https://susanqq.github.io/UTKFace/> (accessed on 19 June 2024).
24. APPA-REAL, chlearnlap.cvc.uab.cat. Available online: <https://chlearnlap.cvc.uab.cat/dataset/26/description/> (accessed on 19 June 2024).
25. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. [[CrossRef](#)]
26. Kaiming, H.; Xiangyu, Z.; Shaoqing, R.; Jian, S. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385. [[CrossRef](#)]
27. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Hartwig, A. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. [[CrossRef](#)]
28. ImageNet. Available online: <https://www.image-net.org> (accessed on 19 June 2024).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.