THE IMPACT OF DEMOGRAPHIC COMPOSITION IN UTKFACE AND APPA-REAL DATASETS ON FAIRNESS IN AGE ESTIMATION MODELS

Nenad PANIĆ*

[1]*Faculty of Technical Sciences, Singidunum University, Belgrade, Serbia, nenad.panic.22@singimail.rs*

Abstract: *This paper analyzes the impact of demographic composition on fairness in facial age estimation models trained on the UTKFace and APPA-REAL datasets. Building on previously published empirical results, the study provides a theoretical and analytical interpretation of how dataset imbalance affects model bias. Through comparative evaluation of group-wise performance metrics including Mean Absolute Error, Standard Deviation, Disparate Impact, and Equality of Opportunity, the paper introduces the concept of a Distributional Fairness Baseline (DFB) as a diagnostic framework for separating dataset-driven bias from model-induced bias. The analysis reveals that fairness is primarily a function of the representativeness and internal structure of training data, rather than model architecture. Contrary to common assumptions, full data equalization through oversampling does not necessarily enhance equity and may even amplify disparities due to overfitting and redundancy. Instead, moderate redistribution, particularly controlled undersampling of dominant groups often achieves an optimal balance between accuracy and fairness. These findings emphasize that equitable model performance depends on both quantitative and qualitative diversity within datasets, establishing data design as the central determinant of fairness in automated age estimation systems.*

Keywords: *fairness, dataset bias, age estimation, demographic imbalance, distributional fairness baseline*

Original scientific paper
Received: 28.10.2025
Accepted: 29.11.2025
Available online: 07.12.2025
DOI: 10.5937/jpmnt13-62440

1. Introduction

Fairness and absence of bias in face processing systems have become key requirements with the rise of artificial intelligence applications in domains such as forensics, security, and commercial applications (Terhörst et al., 2021; Voigt & Bussche, 2017). Research has shown that computer vision models can exhibit significant differences in performance across demographic groups, leading to unequal treatment of users of different ages, genders, or ethnicities (Michalski et al., 2018; Srinivas et al., 2019). This bias problem is widely documented in the literature, for example, face recognition algorithms often underperform certain subcategories of the population such as women or young people, due to the unbalanced data they were trained on (Albiero & Bowyer, 2020; Albiero et al., 2020; Puc et al., 2021).

* Corresponding author

In a historical case of analysis of commercial gender classification systems, Buolamwini and Gebru (2018) showed that models trained on ethnically unbalanced sets classify dark-skinned female faces significantly worse than Caucasian male faces, precisely because of a skewed data distribution in favor of one group. Such examples clearly show that the cause of bias lies not only in the architecture of the model, but above all in the data on which the model was trained (Clapés et al., 2018; Jacques et al., 2019). In other words, even the most sophisticated neural networks will exhibit bias if they are trained on data sets that do not include representative and balanced samples of all relevant groups.

In the context of facial age estimation, model fairness is particularly important as such models are increasingly used in applications that directly affect people, from age verification in digital services to demographic analysis and personalized marketing (Angulu et al., 2018). Bias in age estimation can lead to systematic errors, such as overestimating or underestimating the age of certain ethnic groups. The consequences of such deviations may include the erosion of user confidence, reduced usability of the system in practice, and potential discriminatory outcomes. Therefore, understanding the sources of bias is critical to building fair and reliable models.

Earlier research in the field of age estimation from face images has mostly focused on improving algorithmic approaches, including optimization of convolutional neural network architectures and the introduction of mechanisms such as fairness regulation (Khan et al., 2020; Ramyachitra & Manikandan, 2014; Xing et al., 2017;). However, in practice, it turns out that the improvement of algorithms has a limited effect if the dataset itself is unevenly composed. This paper starts from the assumption that the data, or more precisely, their demographic composition, is the primary source of model bias, and that understanding the distributional structure of the datasets is a key factor in evaluating the fairness of the model.

The paper builds on the research of Panić et al. (2024), published in the journal Mathematics, which empirically analyzed the impact of adjusting the training set on the performance of age estimation models. The results of that research in this paper serve as an empirical basis for the theoretical interpretation of the relationship between data structure and model bias. Unlike the previous work, which had an experimental character and dealt with the quantification of the effects of different dataset compositions, here the focus shifts to theoretical and statistical interpretation, that is, understanding the mechanisms that connect the distribution of data and the measure of fairness of the model, without repeating specific experiments.

The aim of the paper is to analyze analytically and theoretically the components of bias in models for age estimation, that is, to separate the effects that come from the uneven representation of data (dataset bias) from the effects that come from the behavior of the model (model bias). In particular, the impact of ethnic and age structure in the popular public datasets UTKFace and APPA-REAL on the variability and stability of model performance across demographic groups is considered.

In this context, the concept of "distributional fairness baseline" (DFB) is introduced. A baseline level of fairness based solely on data distribution. DFB is a theoretical benchmark that describes the expected level of fairness of a model when its performance is proportional to the representation of certain groups in the data. By comparing the actual results of the model with this baseline, it is possible to quantify the additional bias that does not arise from the distribution, but from the very way in which the model internalizes the distribution of the data. In this way, the paper not only confirms the importance of dataset composition but also proposes an analytical framework for quantitative understanding and measurement of bias in age estimation models.

2. Literature review

Bias in facial image processing algorithms has attracted much attention from researchers, regulators, and the general public in recent years. Initial work mainly highlighted gender and ethnic biases in facial recognition systems. One notable example is the Gender Shades study (Buolamwini and Gebru, 2018), which showed that commercial facial gender classification systems have up to 34% higher error for dark-skinned female faces than for Caucasian faces, due to unbalanced training data and lack of diversity. These findings have raised awareness that training models on unrepresentative sets can lead to systematic errors to the detriment of minority groups.

Similarly, research by Albiero and Bowyer (2020) analyzed the performance of algorithms for face recognition by gender and concluded that apparent gender bias actually stems from factors such as hairstyles and other external characteristics, again implying that collecting balanced data (e.g., women with different hair styles) is key to fair performance.

In the domain of age estimation, bias is somewhat less researched, but there are papers that point to similar patterns. For example, Clapés et al. (2018) analyzed biases on the APPA-REAL database and observed that models can produce different errors depending on gender and ethnicity, even when the actual age is the same. Jacques et al. (2019) similarly showed that human age-perception bias embedded in annotations can propagate to model predictions. Their work represents one of the early attempts to quantify differences in the performance of age estimation models for different demographic groups.

Following on from this, (Puc et al. 2021) examine the influence of gender and race (including their combinations) in deep models for age estimation on the UTKFace and APPA-REAL sets. The authors find that the models are on average more accurate for men than for women, while differences by race are inconsistent between sets, suggesting that error is significantly influenced by data characteristics and quality (e.g., pose, recording conditions), rather than just nominal race. These results support an approach that considers, in addition to overall group proportions, intragroup diversity and uniform data quality as key fairness factors.

Torralba and Efros (2011) were among the first to point out that models implicitly "learn" the bias present in the data. Their work Unbiased look at dataset bias showed that models trained on one dataset generalize poorly to others due to differences in distribution. In the face domain, Kärkkäinen and Joo (2019) developed the FairFace dataset as a balanced set by race, gender and age, showing that by smoothing the data distribution, model bias can be significantly reduced.
Data balancing techniques such as oversampling and undersampling are still the most common approaches to mitigating these problems. Oversampling increases the representation of minority classes, while undersampling reduces the number of samples of the majority group (Ramyachitra and Manikandan, 2014). However, excessive oversampling can cause overfitting, while undersampling leads to loss of information and a drop in accuracy.

The work of Panić et al. (2024) noted that simply equalizing the number of samples by group did not result in the fairest model performance. On the contrary, moderate fitting of the distribution has been shown to be more effective in achieving error balance across groups, implying the existence of an optimal point between overall accuracy and fairness.

Recent research has broadened beyond traditional CNN-based methods toward more advanced architectures, including foundation models, transformers, and graph-based approaches. Hassanpour et al. (2024) showed that large language models such as ChatGPT, despite not being explicitly trained for biometrics can recognize identities, classify gender, and estimate age with notable accuracy when prompted appropriately, highlighting both the potential and the ethical concerns of foundation models in facial analysis.

Transformer-based multi-task systems represent another significant development. Narayan et al. (2025) introduced FaceXFormer, a unified architecture capable of handling ten facial analysis tasks, including age, gender, and race estimation. By treating tasks as learnable tokens, the model achieves competitive or state-of-the-art performance while operating in real time, illustrating the growing shift toward integrated multi-task solutions.

Studies focused on demographic-specific performance further emphasize the need for diverse data. Oladipo et al. (2024) found that Local Binary Patterns outperform other feature-extraction methods for age estimation on Black faces, underscoring the "other-race effect" and demonstrating that techniques optimized on predominantly Caucasian datasets may generalize poorly across racial groups.

Efforts to standardize evaluation practices have revealed additional sources of inconsistency. Paplham and Franc (2024) showed that many reported differences between age-estimation models stem from variations in preprocessing, alignment, or dataset setup rather than intrinsic architectural advantages. Their unified benchmark based on FaRL encourages more reliable comparison across studies.

Further improvements continue to emerge. Dey et al. (2024) demonstrated that robust preprocessing and augmentation can substantially improve CNN-based age and gender prediction on real-world images. Shou et al. (2024) proposed the MCGRL framework, which integrates CNN-derived semantic features with masked graph convolution and contrastive learning to achieve strong performance on multiple datasets. In summary literature indicates that: (i) model fairness is strongly determined by the diversity, representativeness, and quality of training data; (ii) fairness can be evaluated through group-level error metrics such as error variance, disparate impact, and equality of opportunity; (iii) balancing datasets helps but is insufficient without high-quality, intra-group-diverse samples; and (iv) emerging transformer, foundation-model, and graph-based architectures provide promising improvements but also introduce new considerations for fairness assessment.

3. Overview of UTKFace and APPA-REAL datasets

For the theoretical analysis of the influence of data distribution on the fairness of the model, as stated, this work will make use of the experimental results published in the research of Panić et al. (2024). In that work, two widely known datasets for facial age estimation, UTKFace and APPA-REAL, were used. Chosen for their availability, size and existence of demographic meta-data such as gender, age and ethnicity. Their different ethnic composition makes them suitable for studying the influence of demographic representation on model performance.

UTKFace is a large set of face images collected in the wild, annotated with age, gender and ethnicity (Panic et al., 2024). In its original form, it contains 23,705 images of faces of people aged 1 to 116 years, divided into five categories of ethnicity: White, Black, Asian, Indian, and "Other" (Panic et al., 2024).

As part of the research by Panić et al. (2024), this set was pre-filtered to be comparable to the APPA-REAL database, which contains only three ethnic labels (White, Black, Asian). In doing so, the samples marked as "other" were omitted, while the samples marked as "Indian" were added to the category of the black race, in accordance with the experimental framework of that work.

After filtering, UTKFace contained a total of 22,013 images, distributed by group, namely White Race contained 10,078 samples (45.8%), Black Race (including samples previously labeled as Indian) had 8,501 samples (38.6%) and Asian Race had 3,434 samples (15.6%).

Such a structure ensured a relatively balanced distribution by group and allowed to analyze the influence of slight demographic unevenness on the equity metrics.

APPA-REAL is a dataset developed as part of the ChaLearn Looking at People Challenge 2018 (Agustsson et al., 2017). It consists of 7,591 face images with two labels for each face, real age and apparent age. In the research by Panić et al. (2024), real age was used as the target variable.

In contrast to the UTKFace set, APPA-REAL is strongly unbalanced by ethnicity: more than 88% of the samples are made up of Caucasian faces, while Asian and Black populations are represented by about 9% and 3%, respectively (Clapés et al., 2018). Specifically, out of a total of 7,591 images, 6,686 are white, 231 black and 674 Asian (Panić et al., 2024). This distribution of about 30:1 between the majority and the smallest group represents an extreme example of imbalance, which makes it suitable for analyzing the effects of imbalance on model error.

In the aforementioned research, model training and evaluation were conducted using the VGG19 architecture, adapted for each dataset by a fine-tuning procedure. The results showed that the choice of architecture (e.g. VGG19, ResNet50, MobileNetV2) had no decisive influence on the bias patterns, while the data distribution had a dominant effect.

4. Analysis methodology

In this study, we analyze and reinterpret results from a previously published study (Panić et al., 2024), which included detailed performance of the VGG19 model trained and evaluated on the UTKFace and APPA-REAL datasets, with different variations in their ethnic composition.

Within that research, all the metric indicators such as Mean Absolute Error, Standard Deviation, Disparate Impact and Equality of Opportunity (MAE, SD, DI, EoO) have already been quantitatively calculated, and in this paper, they are used again for analytical purposes, that is, to interpret the relationship between the data distribution and the expressed bias of the model. Therefore, we do not perform any recalculation of the values but apply the existing metric results within the new theoretical framework, with the aim of defining and illustrating the concept of distributional fairness baseline.

Complete numerical results have already been published in Panić et al. (2024) and are not reproduced here; below, we refer to them and synthesize the patterns relevant to our analytical framework. Detailed tables for all variations (e.g. "Asian 90%", "White 20%", "Equal") are available in the cited publication (UTKFace - Table 8; APPA-REAL - Table 4).

The methodological approach therefore has three components:

- Performance and equity metrics
- Linking performance to distribution
- Distributional Fairness Baseline (DFB)

4.1. Performance and equity metrics

4.1.1. Basic performance indicator – MAE

As a basic measure of model accuracy, we use the mean absolute error (MAE) expressed in years, which represents the average absolute deviation of the predicted age from the subject's actual age. In the original research, the MAE is calculated overall for the entire test set and separately by ethnic group (White, Black, Asian):

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \mid y_i - \widehat{y_i} \mid \quad (1)$$

where y_i denotes the actual age, and $\widehat{y_i}$ predicted age, while N represents the number of samples.

4.1.2. Standard deviation of errors between groups – SD

To quantify fairness between groups, we rely on the standard deviation of the MAE between ethnic groups, which represents the degree of dispersion of errors and thus indirectly measures the unevenness of model performance.

Formally, if MAE_B, MAE_C and MAE_Avalues for White, Black, and Asian groups, then:

$$SD = \sqrt{\frac{(MAE_B - \bar{MAE})^2 + (MAE_C - \bar{MAE})^2 + (MAE_A - \bar{MAE})^2}{3}} \quad (2)$$

$\bar{MAE}$ is the average value over all three groups.

Smaller values of SD indicate more uniform performance and greater fairness of the model (ideally $SD = 0$ would mean complete error parity).

4.1.3. Additional fairness indicators – Disparate Impact (DI) and Equality of Opportunity (EoO)

In the original research (Panić et al., 2024), the metrics Disparate Impact and Equality of Opportunity were adapted to the regression context, and we use them again in this paper as indicators of relative equality of performance among groups.

Disparate impact (DI) measures the ratio of the error of a particular group to the reference (White) group:

$$DI_{grp} = \frac{MAE_{grp}}{MAE_{White}} \quad (3)$$

$DI = 1$ means complete parity; $DI > 1$ indicates greater errors (less favorable outcome) for the given group.

Equality of Opportunity (EoO) normalizes the error difference to the range 0 to 1:

$$EoO_{grp} = 1 - \left| \frac{MAE_{grp} - MAE_{White}}{MAE_{White}} \right| \quad (4)$$

where $EoO = 1$ means complete equality, while lower values represent a deviation from fair parity.

DI and EoO are interrelated, both measures express the same phenomenon of disparity, but through different scales. In further analysis, we use the standard deviation (SD) as an aggregate indicator of global equity, while we report DI and EoO when it is necessary to identify specific disadvantaged groups.

4.2. Linking performance to distribution

After defining the metrics, the next step of the methodology is to relate the performance of the model to the distribution of the training data.

From the available tables (Panić et al., 2024), the following scenarios were selected that include three basic configurations:

- Original set - model trained on the original, unmodified versions of UTKFace and APPA-REAL with their natural distributions;
- Balanced set - model trained on an artificially balanced version of datasets by oversampling minority groups to equal proportions;
- Intermediate configurations - models trained on dataset variants that gradually change the proportion of groups (e.g. "White 20%" means that the number of samples of the White group is reduced by 20%, while the other groups are kept intact).

  Such variants (more than 30 per dataset) provide a continuum from extremely unbalanced to perfectly balanced distributions, allowing analysis of the effect of representation on model error.

The analytical approach is based on monitoring trends: how the MAE of a certain group changes in relation to the change in its representation in the training set, whether the overall accuracy changes, and where the optimal point of compromise between accuracy and fairness is located. Particular attention was paid to identifying the configurations that yield the minimum standard deviation across groups (the fairest performance) and comparing them to the configurations with the maximum overall accuracy.

4.3. Distributional Fairness Baseline (DFB)

As a central part of the methodological framework, we introduce the concept of distributional fairness baseline (DFB). A theoretical reference point that represents the expected level of fair model performance conditioned solely on the data distribution.

The essence of the concept is that differences in performance between groups can be expected even without explicit model bias: a group that makes up a smaller part of the set will naturally have a larger error due to the smaller number of samples available. This "natural" disparity constitutes the baseline level of fairness, that is, what can be attributed to the distribution of the data.

Intuitively, if a group makes up 10% of the training set, its predictive error will almost certainly be larger than the error of a group that makes up 70%. If the difference in errors is smaller than the distribution suggests, the model has achieved above-average generalization; if it is larger, there is an additional source of bias (e.g., image quality, facial features, overfitting).

Quantitatively, DFB can be approximated by assuming that MAE is inversely proportional to the amount of available data. If groups A and B have $n_A$and $n_B$samples, the expected ratio of their errors can be heuristically expressed as $\sqrt{n_B/n_A}$. Although we do not calculate explicit baseline values in this paper, we use this principle as a qualitative basis for interpretation.

5. Analysis of results

In this chapter, we present the results and analysis of the relationship between data distribution and model performance. We observe configurations such as the original version of the UTKFace and APPA-REAL sets, and fully balanced versions obtained by oversampling minority groups as well as intermediate configurations, that is, scenarios in which, after the initial balancing of the datasets, the number of samples of individual groups is systematically reduced (e.g. "Black 50%" means that 50% of the samples of the Black group are removed from the already balanced set). Systematic reduction of samples was done from balanced datasets to enable an isolated observation of the effect of controlled reduction of individual groups without additional imbalances from the original distributions. Each group went through a gradual reduction of samples from 10% to 100%, for each of the configurations all the metrics listed in the previous chapter were stored.

The aim of the analysis is to answer three questions:

- Does a higher group representation in the training data really lead to a lower error (MAE)?
- How does extreme imbalance affect model fairness?
- Is there an optimal trade-off point between overall accuracy and error equality across groups?

5.1. Performance on original (unbalanced) sets

We start with the analysis of the models trained on the original versions of the UTKFace and APPA-REAL datasets, because they reflect the real scenario of the collected data, that is, the biases that would appear in the usual use of these databases without special corrections. Table 1 shows for both sets the representation of each ethnic group in percentages and the obtained MAE on the test set for that group, as well as the total MAE.

Table 1. Group representation and model error (MAE) by group for the original versions of the UTKFace and APPA-REAL datasets

Dataset	Group	Representation percentage	MAE of the model (years)
UTKFace	White	45.8%	5.38
	Black	38.6%	5.04
	Asian	15.6%	3.67
Total	-	100%	4.89
APPA-REAL	White	88%	6.42
	Black	3%	6.45
	Asian	9%	6.86
Total	-	100%	6.46

Source: Author

From Table 1, several observations immediately stand out, let's start with UTKFace, which is an almost balanced set. Overall, the model on it achieves an MAE of 4.89 years on the test, which is expectedly better than on APPA-REAL which has less data. However, the distribution of errors among groups is far from uniform. The White group has the largest error (5.38), then the Black group has a slightly smaller error (5.04), while the Asian group has a significantly smaller error (3.67). This difference is surprising because the Asian group is also the least represented with only 15.6%. Intuition would suggest that less data leads to more error, but here we see the opposite trend, the smallest group has the best result. The standard deviation of MAE between groups for UTKFace is 0.74, which is quite a large deviation. Fairness viewed through the DI metric says that the Black group has a DI of 0.94 (meaning ~6% less error than White) and the Asian DI of 0.68 (Asian error is 32% less than White group error). The EoO for the Asian group is only 0.68, which is well below 1, signaling a significant difference. These numbers indicate a bias in favor of the Asian group, or alternatively, a bias against the White (and somewhat less for Black) group in UTKFace. Such an outcome could be due to several factors that we will discuss (e.g. whether Asian images are more homogeneous or the age range is different), but for now we note that more data did not guarantee a smaller error in the UTKFace case.

Next, let's look at APPA-REAL, which is a highly unbalanced set. The total MAE is 6.46 years, which is expectedly weaker than UTKFace. What is interesting is that the performance between the groups is relatively close: White group 6.42, Black 6.45, Asian 6.86. The standard deviation between these errors is about 0.20, which is quite small in absolute terms. Looking diagonally, the Black group, which is extremely small at 3%, has an almost identical MAE to White, while the Asian group, with 9% of the data, has a slightly larger error (0.44 higher MAE than whites). The DI for the Black group is ~1.00 (virtually no disparity), for the Asian ~1.07 (7% higher error than whites). This is quite unexpected given the drastic imbalance. The model somehow managed to maintain almost fair performance between White and Black groups, and only slightly worse for Asians. This outcome is sometimes attributed in the literature to the phenomenon that, when the model has very little data about a group, it actually fails to learn the special characteristics of that group but implicitly associates the patterns of the majority with it, which can result in a similar error rate. In other words, with only 231 images from the Black group, it is possible that the model did not even develop an excessive bias towards black faces, it treated them similarly to other faces, which resulted in a similar MAE here. The Asian group, with slightly more samples (674), may be represented enough for the model to learn some specifics for it, but still not enough for the accuracy to be equal to the White group, so the MAE is slightly higher. In any case, the APPA-REAL original model shows surprisingly little error variance across groups given the extreme dominance of one group in the data. This result with a standard deviation of 0.19 can be seen as a reference point. Despite the unfair distribution, the model achieved a relatively fair performance (at least between the White and Black groups). Is that the maximum of fairness? Not necessarily, we will see that with further interventions even greater equality can be achieved, but often at the cost of overall precision.

To illustrate these differences, Figure 1 shows the model errors by group for the original sets, indicating the percentage representation of the groups.

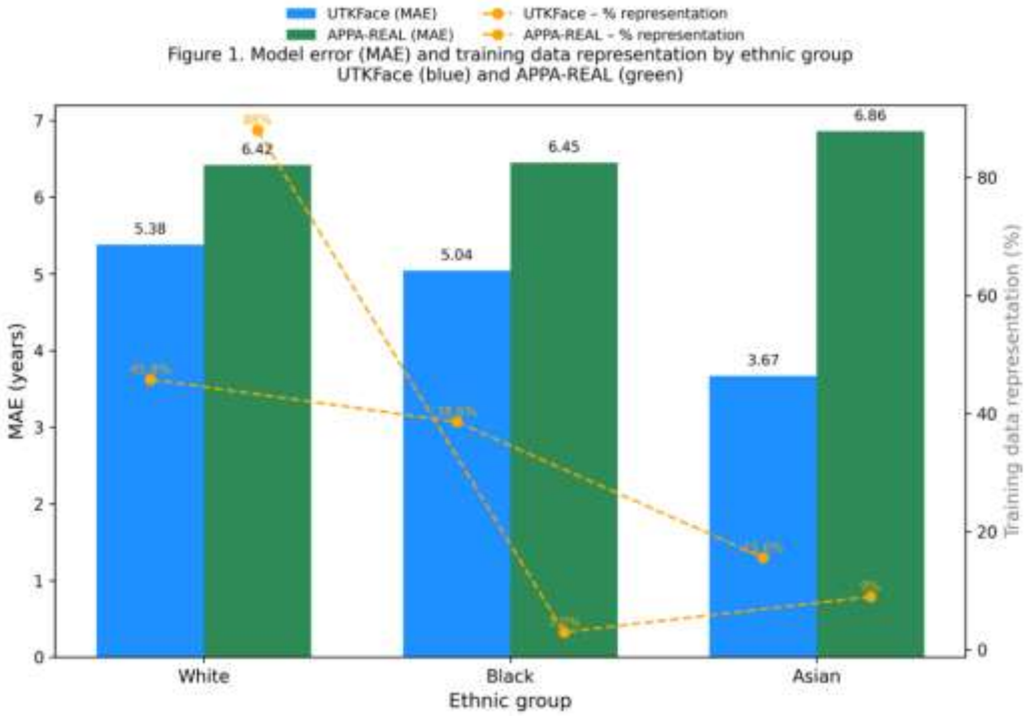

Figure 1. Model error (MAE) by ethnic group on the original versions of the UTKFace and APPA-REAL sets

Source: Author

Analyzing Figure 1 and Table 1 together, we can see the following: In APPA-REAL, there is an expected trend that the group with extremely low representation (black, 3%) has slightly worse (or barely worse) accuracy than the majority, which is consistent with the assumption that more data helps. Also, the group with medium representation (Asian, 9%) has a slightly larger error, which again makes sense. In contrast, in the UTKFace case we see the opposite trend: the group with the least data (15.6%) has the best result. This inverse disparity suggests that the distribution of the data is not the only factor determining the error. In UTKFace, Asian images may be easier for the model, possibly are more homogeneous in terms of age characteristics, or there is a bias in the labels. Alternatively, it may be that in the presence of many white and black faces, the model overfits to some extent on them and that the generalization for Asians turns out to be better.

This is where the notion of distributional fairness baseline is introduced: If the error were solely a function of the number of samples, we would expect the MAE to decrease monotonically with increasing representation. APPA-REAL partially follows that logic (Black vs White negligible difference, Asian higher error with more samples than Black group, which is a slight deviation), while UTKFace clearly violates - the Asian group out-performed what its 15.6% share would predict. We will consider this phenomenon in more detail below, but intuitively we can say: the UTKFace model had a somewhat "easier" path for Asian examples in the feature space, probably due to the way the data is structured.

5.2. Impact of data balancing and cross-configuration

Next, we look at how modifying the distribution of the training data affects the performance of the model. We focus on two specific cases: (i) Equally balanced set - a scenario where there are equal numbers of White, Black and Asian faces in the training set, achieved by oversampling minority groups to the number of majority samples; (ii) Optimal configuration for fairness - the scenario (among the variants tested) that gave the lowest standard deviation of MAE across groups, that is, the most uniform errors.

For APPA-REAL, as we stated, the original standard deviation was 0.1987 (7th lowest among all variants), while the equally balanced variant had a standard deviation of 0.2189. Surprisingly, oversampling to full parity did not improve fairness, but slightly worsened it. The standard deviation increased from ~0.20 to ~0.22. This implies that increasing the minority samples alone was not enough to even out the errors; moreover, the model appears to have begun to produce more variation across groups.

For example, in the fully equal version of APPA-REAL, the MAE of the White group is 6.45, the Black group is 6.97, and the Asian group is 6.57, which means that the error for the Black group has increased compared to the original (where it was 6.45). It is likely that repeating a small number of Black groups images caused the model to overfit those samples, negatively impacting generalization. Paradoxically, the model is worse on Black groups faces despite having seen them repeatedly during training. On the other hand, the fairest configuration for APPA-REAL turned out to be the one labeled “White 20%”. Thus, a scenario where the number of whites is reduced by 20% (leaving 80% of the samples of the White group), while the other groups remain intact. In that case, the resulting standard deviation was only 0.04, practically negligible. The errors by group were then: MAE_White = 7.17, MAE_Black = 7.17, MAE_Asian = 7.07. Thus, the White and Black groups had almost identical MAE, the Asian group barely slightly better. That scenario indeed reached almost perfect parity (DI≈1.00 for both minority groups; EoO≈0.99).

However, this came at a price, the total MAE then is 7.16, which is noticeably worse than the original's MAE=6.46 (about 0.7 years more error). So, to make the model maximally fair on APPA-REAL, we paid a drop in overall accuracy of ~11%. The question is: is that 11% deterioration worth the significant gain in equity? It depends on the context of application, somewhere it might be if fairness is critically important, e.g. legal or ethical implications, somewhere not if accuracy is a priority, e.g. in medical diagnostics where a mistake means wrong therapy. Interestingly, completely removing the White group (“White 100%” scenario) was the worst possible scenario, then the model was trained only on Black and Asian groups; the overall MAE jumped to 8.89 and the standard deviation to 0.94 as the White group error (which appears in the test, though not in the training) exploded to over 9 years. As expected, eliminating the largest group is not a smart strategy, but it is useful to see that extreme: extreme "balancing" (in this case leaving only minorities) dramatically harms both accuracy and fairness.

Let's summarize the findings for APPA-REAL:

- There is a trade-off between accuracy and fairness: The fairest model has slightly worse overall accuracy.
- Oversampling to equal proportions did not give the best results, even worse than the original in fairness, which is counterintuitive but explainable through potential overfitting and noise.

- Mild undersampling of the majority group provided optimal fairness, with minimal disparity.
- Excessive undersampling degrades both accuracy and fairness (because then the other group becomes the new majority and the model breaks down).
- These findings support the idea that the optimal distribution is not trivially "33-33-33" for everyone, but depends on the specific data and its characteristics.

For UTKFace, the situation is even more interesting because the starting point was already somewhat balanced but with unequal errors. As a reminder, the original UTKFace had a standard deviation of 0.74, and the fully equalized variant obtained by oversampling the Black and Asian groups to the number of Whites gave an even worse standard deviation of 0.856. So, as with APPA-REAL, equal oversampling increased the error variance, the fairness worsened, and it ranked only 18th in fairness, while the original was ranked 5th. This is further proof that oversampling a minority does not in itself guarantee a fair outcome.

The fairest UTKFace configuration, interestingly, was the one labeled "Asian 90%" - that is, removing 90% of the Asian samples leaving only 10% of the Asian part. That sounds counterintuitive but remember the Asian group originally had the advantage with a much smaller MAE. By removing most of the Asian training data, the model practically worsened its performance for Asians to a level closer to the others, causing the variance to drop. In that scenario, the standard deviation is only 0.30, which is a huge drop, the fairness is improved by ~60% compared to the original. However, the overall MAE then rises to 5.48 (~12% higher than the original 4.89). So again, a trade-off, the fair model is less accurate. In fact, all four top-fair models for UTKFace included significant reduction of Asian samples (removing 80–100% Asian). Clearly, it "fixes" equity by equalizing performance downwards, a downgrade for Asians leads to equalization with others.

On the other hand, the variants that removed the White group had the highest variation, which is analogous to the APPA-REAL situation, while removing the Black group completely is also among the worst (4th worst variant). These extremes confirm again: excessive "balancing" by removing most of the data of one group leads to bad fairness. Interestingly, both removing the White and removing the Black group leads to high disparity, implying that the variance increases whenever one of the groups is missing. It was best to partially suppress that group that had a disproportionately low error, in UTK it is Asian.

Looking at the overall picture, the results indicate that balance in terms of the number of samples is not a magic solution. The UTKFace example is instructive, although the dataset is relatively balanced, the model was biased towards the Asian group. The sample balance didn't solve that; it even made it worse. The reason may be in the characteristics of the data itself: it is possible that the age distribution of the Asian subset is different, or that the visual aging patterns of Asians in the given data allow easier estimation of the model (which would be an interesting biometric implication). In contrast, APPA-REAL shows that even extremely unbalanced data does not inevitably lead to massive bias, the original model was relatively fair, perhaps because the minority data were too small to "distort" the model.

5.3. Correlation between representation and error

As we indicated, we will try to more quantitatively shed light on the relationship between the percentage of representation of a certain group in the training set and the achieved MAE for that group. Based on the results from the previous sections, we can construct a series of shape points (percentage, MAE) for each group in different scenarios. However, due to the complexity of the interactions, the linear correlation may not be high. Nevertheless, certain tendencies are visible.

Globally, groups that are very few in number are prone to higher error, e.g. the Asian group in APPA-REAL with 9% representation has a higher MAE than the White group which is 88% of the dataset. Also, in the extreme APPA-REAL "White 0%" scenario where there are no whites, whites have a huge error in testing, confirming that 0% representation is a catastrophic error. On the other hand, when some group is dominant in the set, the model often optimizes its total error, but this does not guarantee the minimum error among all, we saw that the White group samples in UTKFace did not have the smallest MAE despite the most data; however, whites in APPA-REAL did have the lowest MAE, albeit minimally compared to blacks.

When looking at the mean values, it can be observed that changes in the balance of the representation of ethnic groups in the training set directly affect the differences in model errors, but not in a linear way. In general, approaching a more even distribution most often leads to a reduction of differences between groups, but there are also cases in which a partial imbalance has a similar effect, if the model thereby loses dominant or "too easy" subsets that previously violated global equity.

Using the example of the APPA-REAL set, the original distribution in which Whites made up about 88% of the data, Blacks about 3%, and Asians 9%, resulted in average absolute errors of 6.42 years for the white, 6.45 for the black, and 6.86 for the Asian group. When the "White 20%" scenario was run, in which the proportion of whites was reduced to approximately 28% and the proportion of blacks and Asians increased to about 36%, the errors almost leveled off at 7.17 for the white, 7.17 for the black, and 7.07 for the Asian group. Although the overall MAE of the model in this regime increased by about eleven percent, the standard deviation of the group errors decreased by more than eighty percent. In other words, the model became significantly fairer, as the variation between groups almost disappeared with minimal deterioration in overall accuracy. A similar pattern was observed in the UTKFace set, but through the opposite mechanism. In the original distribution, where whites were about 46%, blacks 39%, and Asians 16%, there was a marked difference, especially between the Asian and other groups (MAE 3.67 vs. 5.38 and 5.04). However, in the "Asian 90%" scenario, where the share of the Asian group was reduced to less than 5%, while White and Black groups rose to about 48% each, the differences narrowed: the errors were 5.73 for the white group and about 5.00 for the Asian group. Although a greater imbalance in the number of samples was achieved here, by removing the anomalously "easy" group, a better balance of performance was achieved among all of the groups.

These findings show that the increase in fairness is not a simple function of numerical balance, but rather the result of optimizing the intergroup structure of the data. In practice, this means that equity can also be improved by reducing dominant classes or removing specific subsets that do not represent the wider population. Thus, the same principle can be seen in both sets: a moderate redistribution of data either towards greater balance (as in APPA-REAL) or towards greater homogeneity (as in UTKFace) leads to a reduction of extreme differences in errors with a minimal drop in overall accuracy.

We can show the correlation between representation and error with graphs that give a deeper insight. Figure 2 shows the group-specific values of the average absolute error (MAE) in relation to the percentage of the remaining representation of each ethnic group in the training set.

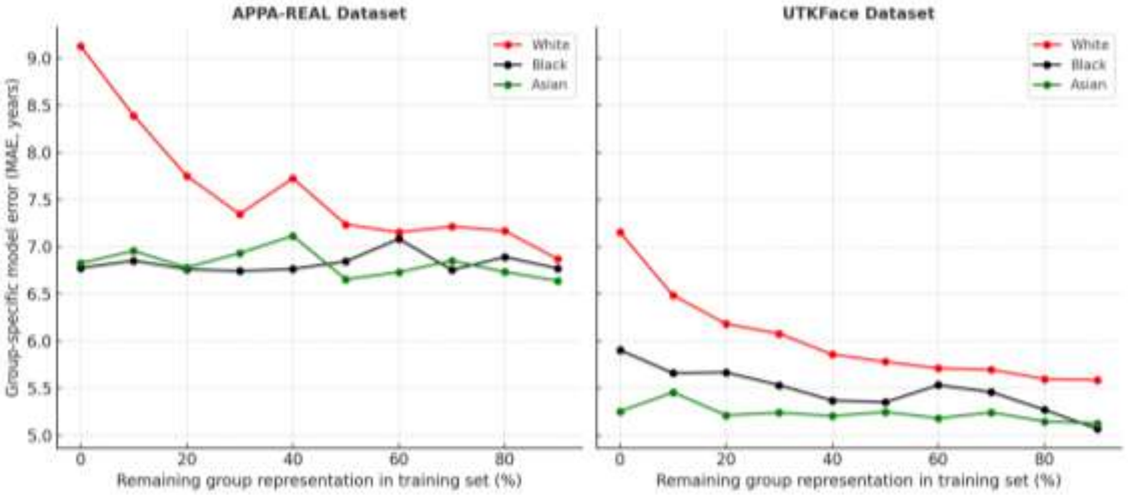

Figure 2. Relationship Between Group Representation and Model Error (MAE)
Source: Author

Through this graph, it can be seen that the relationship between the representation of the group in the training set and the model error is not linear, but that two distinct modes of behavior can be distinguished.

The first regime refers to low representation (below approximately 10–20%), where the error rises sharply or remains high. This indicates that the model fails to learn the specific patterns of the minority group, a phenomenon known as minority data underfitting. In extreme cases, such as the "White 100%" scenario in the APPA-REAL set, when there are no whites in training at all, the error reaches extreme values, which confirms that the complete absence of data leads to serious performance degradation. Although sometimes it seems that the error of the minority is similar to the majority group, as in the case of the black population with only 3% of data, this is not a sign of fairness, on the contrary, such uniformity arises because the model does not learn the peculiarities of that group, but uncritically transfers the patterns of the majority population, which gives seemingly similar, but fundamentally uninformed results.

The second regime includes moderate representation (between 20% and 60%), where the error shows greater variability between groups and datasets. Here, the amount of data is no longer a dominant factor, but qualitative differences in the structure of the data come to the forefront, such as age distribution, homogeneity of visual characteristics and consistency within each group. Thus, in the original UTKFace set, the Asian group, although it only accounts for 15.6% of the data, achieves the lowest MAE of 3.67 years, compared to 5.04 for the Black and 5.38 for the White group. This confirms that greater representation does not necessarily mean less error, the homogeneity and stability of the samples are often more important than the quantitative balance itself.

Overall, the graph shows that increasing the representation of a particular group leads to a smaller mean absolute error (MAE) on average, but in a highly non-linear way. After approximately 50% representation, the saturation effect becomes noticeable, the additional data no longer contributes to a proportional reduction of the error. Conversely, a sharp decrease in the proportion below the critical point (about 10%) leads to a degradation of performance and an increase in intergroup differences.

These findings confirm that the fairness of the model does not depend exclusively on the quantity of data per group, but above all on its information diversity, representativeness and intragroup homogeneity, which together determine the ability of the model to generalize fairly.

A key conclusion of this analysis is that achieving equity generally requires reducing imbalances in representation, but simply equalizing proportions does not necessarily equalize errors. In other words, the relationship between the distribution of data and model errors is not linear, there is an optimal equilibrium range in which the errors between groups are the smallest, while outside this range the differences increase again.

6. Discussion

The presented findings clearly demonstrate that data distribution has a substantial yet non-linear impact on model fairness. While model architecture determines general accuracy, it is the representativeness and structure of the dataset that predominantly dictate demographic error disparities. The following synthesis examines why particular fairness phenomena occur and what implications they hold for the design of fair age estimation models.

6.1. Why Did Oversampling Not Produce the Expected Improvement?

In both datasets, complete equalization of the distribution through oversampling did not reduce error variation, but on the contrary, it increased it. The reason lies in the nature of oversampling; it adds data quantity without increasing informational diversity. By replicating or augmenting existing samples, the model repeatedly encounters the same faces under slightly altered conditions such as rotation, illumination, or magnification, which leads to overfitting to patterns present in a limited number of original images.

In the APPA-REAL dataset, where the Black group contains 231 images, expanding this subset to 6,000 samples to match the White group means that each original image was repeated more than 25 times. This imbalance between quantity and variety causes the model to memorize specific individuals rather than develop a generalized understanding of demographic features. As a result, test performance deteriorates, the MAE for the Black group increases, and the overall deviation grows.

Oversampling also amplifies noise in the data. When minority groups contain mislabeled or visually suboptimal images, their repeated inclusion magnifies the effect of these errors in training, making the model more sensitive to irrelevant patterns and degrading performance across all groups, as noted by Ramyachitra and Manikandan (2014).

In summary, although oversampling aims to achieve a fairer class distribution, in practice it diminishes the model’s ability to learn robust representations. Quantitative equality does not mean informational equality.

6.2. Why Does Slight Undersampling of the Majority Improve Fairness?

Unlike oversampling, moderate reduction of the majority group proved considerably more effective in improving fairness. The mechanism behind this effect lies in the change of weight dynamics during optimization. When the majority class dominates, the model minimizes total loss primarily according to its samples, while errors on minority groups contribute marginally. By reducing the number of White group samples, the relative influence of minority groups on the loss increases, forcing the model to identify parameters that generalize more evenly across subpopulations.

This adjustment makes the training process “forcefully fairer,” as each group exerts a more proportional influence on the objective function. Moderate undersampling thus acts as a regulatory mechanism, enabling an equitable distribution of errors with negligible loss in overall accuracy.

This outcome aligns with findings from previous studies (Chawla et al., 2002; Branco et al., 2016), which emphasize that dataset balancing for fairness does not necessarily require increasing minority representation but can be achieved by carefully reducing the majority. In this way, the model develops more stable inter-group representations rather than overfitting to the dominant class.

However, the results also reveal an optimal threshold. In the APPA-REAL dataset, the fairness improvement was most pronounced when the White group was reduced by 20%. A further reduction such as to 50% led to deterioration of both fairness and accuracy, with standard deviation rising to 0.512 and MAE to 6.87. Excessive undersampling disrupts data diversity within the dominant class, creating a new imbalance that ultimately degrades overall model performance.

6.3. Interplay Between Ethnic and Age Distributions

An often-overlooked aspect of fairness analysis in age estimation is the uneven age distribution within ethnic groups, which can generate apparent model biases.

In the UTKFace dataset, the Asian group is concentrated in a narrower age range, predominantly between 20 and 40 years, while the White group spans the entire spectrum, including both children and older adults. Consequently, the model achieves a lower MAE for Asians, not due to inherent fairness, but because models generalize more effectively within age ranges that are most represented. Extremes such as childhood and old age remain more error-prone due to fewer samples and greater variability in visual aging cues.

This pattern is clearly visible in Figure 3, which shows that White samples are both more numerous and evenly distributed across ages, whereas Asian samples cluster around young to middle adulthood, and Black group samples exhibit an even narrower range.

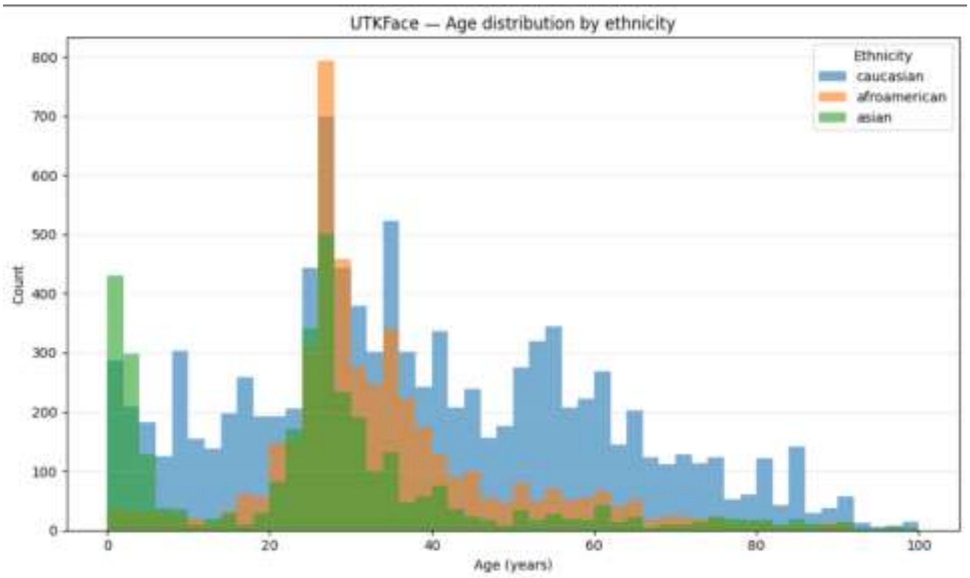

Figure 3. UTKFace - Age distribution by ethnicity
Source: Author

In contrast, the APPA-REAL dataset, visible in Figure 4, displays more uniform age-ethnicity distributions, explaining the smaller inter-group error differences observed in its experiments.

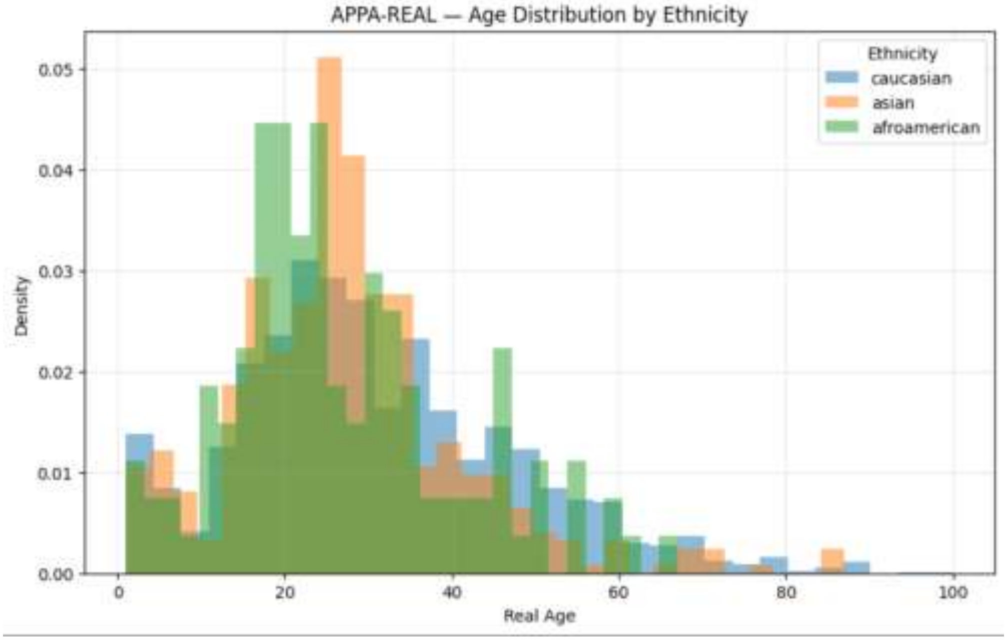

Figure 4. APPA-REAL - Age distribution by ethnicity
Source: Author

This phenomenon corresponds to what the literature describes as dataset shift, a change in the distribution of data between subsets. In our case, it represents an intra-dataset shift among ethnic-age subgroups. The combination of ethnicity and age varies substantially between groups, leading to systematic differences in model behavior even within a single dataset.

Therefore, achieving fairness requires not only balancing the number of samples by ethnicity but also their within-group age distribution. Oversampling the Asian group in UTKFace by merely replicating existing images would not remove bias, since it would only amplify the already dominant age category of younger adults. A more effective strategy would combine targeted oversampling and augmentation techniques that introduce variation and simulate underrepresented age subranges.

The key conclusion is that fairness in data depends on both quantity and qualitative diversity. Only datasets balanced across age, gender, recording conditions, and demographic representation can support models capable of learning equitable, generalizable features.

6.4. Distributional Fairness Baseline – Proof of Concept

The results of the empirical analysis confirm the usefulness of what we define as a distributional fairness baseline (DFB), a qualitative reference point of the expected performance ratio between groups given their representation and diversity in the data. When performance deviates significantly from this benchmark, it indicates additional sources of bias that are not explained by the numerical distribution alone.

When significant deviations from that expected parity are observed, it signals the presence of additional, qualitative biases that cannot be explained by the number of samples alone. Such deviations may indicate differences in age distribution, visual characteristics or quality of data by groups, but also possible effects during training.

In the case of the UTKFace set, the Asian group shows a significantly lower MAE than expected. Even though it has approximately a third of the number of samples compared to the White group, it achieves better results. Based on the distributional analysis, we see that this group includes a narrower age range. Thus, a lower MAE does not necessarily mean that the model is fairer to the Asian group, but that the group is distributionally more favorable to the task, confirming the usefulness of the baseline concept in uncovering hidden sources of imbalance.

We see the opposite example in the APPA-REAL set, where the Black group (with only about 1/30 the number of samples of the White group) achieves an MAE almost identical to the White group. That deviation from the expected trend suggests that the model did not take advantage of the quantitatively dominant group, which may mean that:

- a smaller group has a more representative coverage of the age spectrum, or
- the model generalizes roughly (predictions are clustered around average values), thus unintentionally reducing the difference between groups.

Both cases illustrate how the distributional fairness baseline functions as a diagnostic tool. It enables the identification of scenarios where the model overperforms or underperforms in relation to distributional assumptions. This does not measure the absolute fairness of the model, but maps signals of potential bias that require additional analysis, either through distribution correction, additional data balancing, or redefinition of the training strategy.

7. Conclusion

This study analytically re-examines previously published experimental findings on UTKFace and APPA-REAL to clarify how demographic data composition influences fairness in facial age estimation models. The results clearly show that dataset structure and representativeness, rather than sample size alone, determine disparities in Mean Absolute Error (MAE) across demographic groups. For example, in UTKFace the least represented Asian subgroup (15.6%) achieved the lowest MAE (3.67), while the White subgroup showed the highest error (5.38), whereas the highly imbalanced APPA-REAL dataset (White 88%, Black 3%) exhibited comparatively low disparities (SD ≈ 0.20). Redistribution scenarios further confirmed that fairness is a trade-off: the fairest configurations (UTKFace "Asian −90%" and APPA-REAL "White −20%") reduced inter-group error variation by 60–80% while slightly increasing total MAE (~10–12%), and full oversampling consistently increased disparities by overfitting duplicated minority samples. These findings demonstrate that moderate demographic rebalancing is more effective than naive equalization.

The proposed Distributional Fairness Baseline (DFB) provided a useful interpretive framework, showing that deviations from expected performance based on representation reveal where bias originates from qualitative dataset factors such as age-range compression or image homogeneity.

This work is limited to a theoretical and analytical interpretation of previously reported experiments, focused solely on ethnicity-based group-wise MAE and a specific three-group demographic taxonomy. Additional demographic attributes, data quality differences, and acquisition factors were not examined, and DFB was applied qualitatively rather than as a formal metric.

Future research should incorporate intersectional fairness (age × gender × ethnicity), develop quantitative DFB measures for dataset auditing, and validate these findings on more diverse datasets and modern model families. Overall, our results underline a practical conclusion: fairness must be engineered into the data first, as model-level interventions can only partially compensate for structural demographic bias.

References

Agustsson, E., Timofte, R., Escalera, S., Baro, X., Guyon, I., & Rothe, R. (2017). Apparent and real age estimation in still images with deep residual regressors on appa-real database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Washington DC, USA, pp. 87–94. https://doi.org/10.1109/FG.2017.20

Albiero, V., & Bowyer, K. W. (2020). Is face recognition sexist? no, gendered hairstyles and biology are. *arXiv, arXiv:2008.06989.*

Albiero, V., Zhang, K., & Bowyer, K. W. (2020). How does gender balance in training data affect face recognition accuracy? *IEEE International Joint Conference on Biometrics (IJCB),* Houston, USA, pp. 1-10. https://doi.org/10.1109/IJCB48548.2020.9304924

Angulu, R., Tapamo, J. R., & Adewumi, A. O. (2018). Age estimation via face images: a survey. *EURASIP Journal on Image and Video Processing*, 2018(1), 1-35. https://doi.org/10.1186/s13640-018-0278-6

Branco, P., Torgo, L., & Ribiero, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2), 1-50. https://doi.org/10.1145/2907070

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *1st Conference on fairness, accountability and transparency*, New York, USA, pp. 77-91.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16(1), 321-357. https://doi.org/10.1613/jair.953

Clapés, A., Bilici, O., Temirova, D., Avots, E., Anbarjafari, G., & Escalera, S. (2018). From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation. *IEEE conference on computer vision and pattern recognition workshops*, Salt Lake City, USA, pp. 2373-2382.

Dey, P., Mahmud, T., Chowdhury, M. S., Hosssain, M. S., & Andersson, K. (2024). Human Age and Gender Prediction from Facial Images Using Deep Learning Methods. *The 15th International Conference on Ambient Systems, Networks and Technologies (ANT)*, Hasselt, Belgium, pp. 314-321.

Hassanpour, A., Kowsari, Y., Shahreza, H. O., Yang, B., & Marcel, S. (2024). Chatgpt and Biometrics: an Assessment of Face Recognition, Gender Detection, and Age Estimation Capabilities. *IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, pp. 3224-3229. https://doi.org/10.1109/ICIP51287.2024.10647924

Hasib, K. M., Iqbal, M. S., Shah, F. M., Mahmud, J. A., Popel, M. H., Showrov, M. I. H., … & Rahman, O. (2020). A survey of methods for managing the classification and solution of data imbalance problem. *arXiv, arXiv:2012.11870*

Jacques, J.C.S., Ozcinar, C., Marjanovic, M., Baró, X., Anbarjafari, G., & Escalera, S. (2019). On the effect of age perception biases for real age regression. *IEEE International Conference on Automatic Face & Gesture Recognition*, Lille, France, pp. 1-8. https://doi.org/10.1109/FG.2019.8756595

Kärkkäinen, K., Joo, J. (2019). Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv, arXiv:1908.04913*

Khan, H., Perperoglou, A., & Majeed, H. (2020). A survey of methods for managing the classification and solution of data imbalance problem. *arXiv, arXiv:2012.11870*

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS international transactions on computer science and engineering*, 30(1), 25-36.

Narayan, K., Vibashan, V. S., Chellappa, R., & Patel, V. M. (2025). FaceXFormer: A Unified Transformer for Facial Analysis. *IEEE International Conference on Computer Vision (ICCV)*, Honolulu, Hawaii, pp. 11369-11382.

Michalski, D., Yiu, S. Y., & Malec, C. (2018). The impact of age and threshold variation on facial recognition algorithm performance using images of children. *IEEE International conference on biometrics (ICB)*, Gold Coast, Australia, pp. 217-224. https://doi.org/10.1109/ICB2018.2018.00041

Oladipo, O., Omidiora, E. O., & Osamor, V. C. (2024). Comparative analysis of features extraction techniques for black face age estimation. *AI & Soc*, 39(1), 1769-1783.

Paplhám, J., & Franc, V. (2024). A Call to Reflect on Evaluation Practices for Age Estimation: Comparative Analysis of the State-of-the-Art and a Unified Benchmark. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, pp. 1196-1205.

Panić, N., Marjanović, M., & Bezdan, T. (2024). Addressing Demographic Bias in Age Estimation Models through Optimized Dataset Composition. *Mathematics*, 12(15), 2358. https://doi.org/10.3390/math12152358

Puc, A., Štruc, V., & Grm, K. (2021). Analysis of race and gender bias in deep age estimation models. *28th European Signal Processing Conference (EUSIPCO)*, Amsterdam, Netherlands, pp. 830-834.

Ramyachitra, D., & Manikandan, P. (2014). Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)*, 5(4), 1-29.

Shou, Y., Cao, X., Liu, H., & Meng, D. (2025). Masked contrastive graph representation learning for age estimation. *Pattern Recognition*, 158, https://doi.org/10.1016/j.patcog.2024.110974

Srinivas, N., Ricanek, K., Michalski, D., Bolme, D. S., & King, M. (2019). Face recognition algorithm bias: Performance differences on images of children and adults. *IEEE/CVF conference on computer vision and pattern recognition workshops*, Long Beach, USA.

Terhörst, P., Kolf, J. N., Huber, M., Kirchbuchner, F., Damer, N., Moreno, A. M., … & Kuijper, A. (2021). A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, 3(1), 16-30.

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, USA, pp. 1521-1528.

Voigt, P., & Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676), 10-5555.

Xing, J., Li, K., Hu, W., Yuan, C., & Ling, H. (2017). Diagnosing deep learning models for high accuracy age estimation from a single image. *Pattern Recognition*, 66(1), 106-116. https://doi.org/10.1016/j.patcog.2017.01.005