INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

AUDITING BIOMETRIC EQUITY: A COMPARATIVE STUDY OF DEMOGRAPHIC AND AGE-BASED DISPARITIES ACROSS FACIAL RECOGNITION ARCHITECTURES

Supreet Kaur Sahi[1], Vandana Kalra[2], Medhansh Banga[3]

[1]Assistant Professor, [2]Professor, [3]Student
[1,2,3]Department of Computer Science, Sri Guru Gobind Singh College of Commerce, University of Delhi, New Delhi, India
Corresponding author:
Medhansh Banga (medhansh.224033@sggscc.ac.in)

Abstract

Even though the facial recognition systems are extremely accurate, there is a well-known issue of bias when it comes to facial recognition. In this paper, in order to overcome the problem of bias, we test 5 very different, but related models, with the help of the DeepFace Library (Python). Two different tasks have been carried out, classification and verification. The metrics used to obtain the results were different for both the tasks. This study is intended solely for the purpose of comparison for bias and fairness only and will not consider any other topics. The use of this can be of great benefit for facial recognition in real-world situations where there is no need for a new model to be developed.

1 Introduction

The recent advancement in the deep convolutional neural networks (CNNs) show remarkable performance in terms of accuracy and efficiency, but evidence shows that demographic bias still exists. As face recognition systems are adopted on a larger scale for security purposes, it is important to address the bias in such a way that it is equitable to all parties.

Recently, face recognition has shifted from heavy models to lightweight models as it is utilized very much in cell phone devices these days, and started off as a trend in the iPhone X which introduced the Face ID adhering to a proper scan of the face. This must be done so that the model requires little use of resources and will fit into the limitations of a device. It is not acceptable to have a trade off with efficiency – it should not be at the expense of fairness. Now, there are plenty of Face Recognition models being developed with multiple loss functions and architectures and comparison of these models to find the best one for quick development for various work situations is required. It is not viable to have a set of individuals just to create a new model according to a certain situation in the work environment. There might be bias due to differences in the models, which may affect different demographic groups. Hence we implemented the same pre-processing for all the datasets of the images. This study involves two tasks to evaluate bias and fairness, one task involves classification and the other involves verification (1:1 matching), verification is how face recognition is mostly used these days. In order to conduct research, we developed a grey box audit, which is based on face detection and alignment from RetinaFace[6]. Then we apply the similar pre-processing steps to each image for quality control (to avoid blurry image) and it is very tedious to have to do it for every image out of the 100k of images (both datasets combined). The datasets utilised are very demographically balanced for this research, and are FairFace [12] and Balanced Faces in The Wild [16] rather than a dataset with a million images that would take hours or days to assess. Also with the sample set of clean images, it analysed what was causing the difference to the images and by use of Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) it was clear how the model was behaving on these sample images, which may have contributed to the difference in results. It is important to note that we made use of the DeepFace library [19] which included pretrained models and weights, this more or less likely simulates a real world deployment situation where one would not need to train and test a completely new model.

2 Related Work

The race and gender bias in face recognition has been examined for the last 10 years since the Gender Shades study[4] was performed. The skin toning difference was seen to have resulted in significantly lower error rates in light skin persons as compared to darkier skin persons. If it continued like that, it would result in the majority of future artificial intelligence based systems to have a biased attitude towards them, which would be bad news, and it is better if it was fixed early on.

The National Institute of Standards and Technology (NIST) has conducted evaluations on this topic as well; the ones included in this research are the Face Recognition Vendor Tests (FRVT) in 2019 and 2022[10, 9]. The goal of this study was to investigate whether bias exists with regard to race, sex, and age in verification and identification tasks.This study confirmed that bias exists between races, ages and sex in verification and identification tasks. While the accuracy of most models has significantly improved, there's still bias within different demographics.

What was a academic issue in the first place is now a regulatory concern. Under the EU Artificial Intelligence Act 2024 [8] biometric identification systems are categorized as high-risk AI and explicitly require developers to perform subgroup-level accuracy checks

as well as overall accuracy. This takes a standard of high average accuracy of a deployable face recognition system to a new level: documented fairness across demographic subgroups and comparative audits of the sort that can be done here are no longer only academic, but a legal requirement for practitioners working within regulated markets. Combining the technical evidence provided by NIST [10, 9] with the evidence from the intersectional analysis, as provided by Gender Shades [4], and the requirements of the EU AI Act [8] makes it clear that a need exists for controlled, architecture-level fairness evaluations conducted using metrics of subgroups.The need for controlled, architecture-level fairness evaluations, using metrics of subgroups, is reinforced by combining the evidence provided by NIST [10, 9] with the evidence provided by the intersectional analysis, Gender Shades [4], and the requirements of the EU AI Act [8].

The recent studies still fail to deem bias and demographic fairness as a resolved problem, either because of the algorithm used or the data set [13]. Many studies, however, have been conducted for the study of the problem of designing algorithms for architectures and for compressing models, from the perspective of bias and fairness. Evidence shows that model compression in mobile devices can lead to different results for underrepresented demographic groups [5].

However, though the development of face recognition has been remarkable, pre-processing inconsistencies can negatively affect the overall architecture, and cause them to suffer from bias problems. Pipeline structured preprocessing leads to a meaningful boost in the quality of the model as well, which translates into a better accuracy [17]. To achieve controlled results, it is crucial to have uniform pre-processing operations [19].

Another research area has focused on the influence that the choice of loss function during training has on the geometric structure of identity embeddings and the resulting demographic bias and fairness. Most of the early models of face recognition used the standard Softmax Loss that was developed for object detection and classification, and which does not exist any constraint on angular separation in feature space. This enabled models to have good overall accuracy, using the tool such as skin color or hair color as a shortcut. With the addition of metric learning methods, such as that proposed by FaceNet in [18] which introduced the Triplet Loss, the goal shifted to learning a small Euclidean space where the distance relates to the similarity of identities. The Euclidean distance, however, is sensitive to the magnitude of the vector, which can be influenced by the brightness of the image, thus having a significant impact on the darker skin tone of people who are underexposed in hardware level [9]. The later work on Angular Margin Losses led to ArcFace [7] which normalized the embeddings in all directions into the unit hypersphere, and thus the computation of similarity between two identities is only about their directions and does not benefit from any low-cost demographic proxies. The study's motivation comes from this architectural evolution, which encompasses the entire range of legacy Softmax-based architectures to modern architectures that feature angular margins.

3 Methodology

We chose a few metrics that had been used in previous studies to make the comparison valid with regard to bias and fairness. We created a gray box comparison framework, since it uses explainability techniques and we also have access to the embeddings of the images. We should emphasize that we did not train the models from scratch, rather, we used pre-trained ones to just compare with other models with a similar structure in the same topic, so as not to invent this wheel again. Since we do not have control over how the model was trained, we used the same face detector (InsightFace) and preprocessing conditions, to avoid bias in any of the classification or verification tasks.

3.1 The Unified Benchmarking Framework

For the FairFace dataset, we applied the same face detector to all images to ensure we are able to use a common structure to analyse the five models. The other option is the pre aligned crops from Balanced Faces in the Wild, these were already available by the authors themselves and we ensured that we used these rather than processing through a detector again which could be useful if required. So in this way, it was possible to take a consistent approach to identifying a demographic bias.

3.1.1 Preprocessing and Quality Control

Dataset-Specific Pipelines:

For the FairFace dataset (Task A), raw images were processed using the InsightFace [6] library with the buffalol detection backend for robust face localization and geometric alignment to the center. While buffalol performs alignment, it does not perform any quality filtering. To minimize the impact of low-quality data on our bias and fairness metrics, we applied a strict quality filter to the FairFace inputs. Images that did not meet the following thresholds were removed prior to feature extraction:

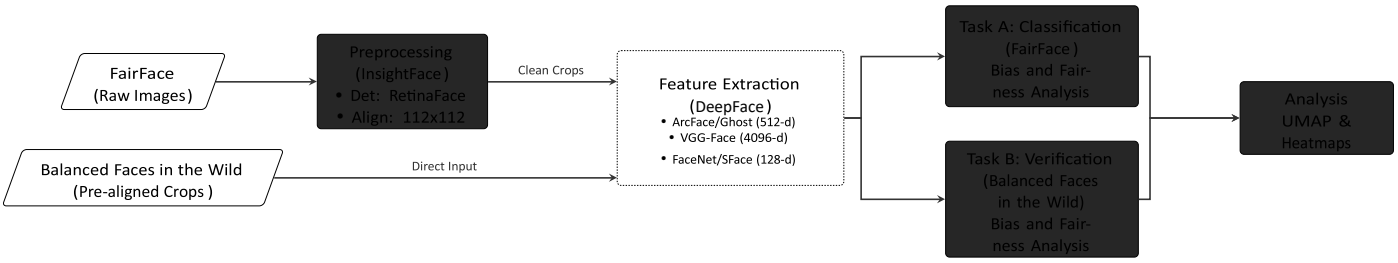


Figure 1: The combined audit pipeline. FairFace is preprocessed in the same manner as a standard detector (InsightFace), and in the case of Balanced Faces in the Wild, the faces are cropped according to the author's alignment, so that compounding detection errors are avoided. Embeddings are sampled in their original dimensionality and then tested on the dimension of fairness for demographics.

Confidence Score ($Conf > 0.5$):
Make sure that the detected object is a valid face, filtering out false positives such as background textures or clothing patterns.

Blur Variance ($\sigma^2 > 15$):
Rejects blurry images that are lacking in detail for feature extraction using the Variance of the Laplacian [20].

Minimum Resolution ($Size_{min} > 30$px):
Prevents resize when upscaling extremely small faces to the size that the model requires as input (usually 112 × 112).

Max Yaw ($|\theta_{yaw}| < 30^{\circ}$):
Excludes side angles of extreme range and profiles to achieve strict control of pose, so that the model only sees a near frontal view and will not be biased because the looker is looking right or left.

Max Pitch ($|\theta_{pitch}| < 20^{\circ}$):
Eliminates faces that are looking either up or down, so that there is no bias caused by angle or geometric disturbance.

These variables are important in order to ensure that the bias and fairness metrics calculated are not altered by a change in the angle, blur, or resolution of the input image.
On the other hand, for the Balanced Faces in the Wild (BFW) dataset (Task B), we did not use the InsightFace detection stage, and utilized the pre-aligned images that had been supplied by the dataset authors. Task B is a verification task, and so the images should be of natural poses (not photoshoot poses) and natural lighting and angle of view, thus, allowing for proper evaluation of identity matching. Images provided are used to keep the mentioned features.
This was also an important methodological control as the pre-aligned images were instantly fed into the feature extractors, thus avoiding an additional error from the face detection stage. This configuration limits the impact of the preprocessing errors and thus the differences found in Task B is mainly due to the differences in the models' learned representations.

3.1.2 Model Selection and Feature Extraction

For the sake of uniformity and repeatability, we used the DeepFace framework [?] for face recognition, which is a simple yet powerful hybrid face recognition suite that centralizes various architectural models under a single interface. This enabled us to carefully maintain the input pipeline and give all the model architectures preprocessed input that was exactly the same. We chose 5 different models to reflect the architectural variety of present models with the condition that it must be a "architectural and functional diversity.'. In the case of the high capacity of the servers, we used ArcFace [7] and VGG-Face [15] that are well-known and are considered for high-security applications due to their deep network architecture. To make a contrast with these heavy models, models like GhostFaceNet [2] and SFace [3] were used as lightweight models that are appearing with the constraints of storage. These models are actually specifically developed for mobile applications, where computational capabilities are extremely restricted. Besides, FaceNet [18] was added as being based on "Triplet Loss", which differs from the other classification-based loss functions in the other architectures such as ArcFace using "Additive Angular Margin Loss" [7].
Firstly, we extracted embedding vectors of each model. The output data showed that the test models gave out vectors of different dimensions: ArcFace and GhostFaceNet gave standard vectors of size 512; the other two test models FaceNet and SFace output vectors of size 128. The "VGG-Face architecture, on the other hand, produced much more extensive 4096-dimensional representations, for which it is famous. We intentionally did not use Principal Component Analysis (PCA) or any other reduction technique to keep these original dimensions, and thus our fairness audit did not rely on any reduction in the information capacity of the pre-trained models. Applying reduction techniques might make it unfair for one pre-trained model over the other.

3.2 Evaluation Protocols

Two different tasks were used to simulate real life deployment scenarios for the audit.
Task A: Demographic Classification (Bias in Representation) The FairFace dataset was used for the classification task. Classification was done using linear probing [1] with a logistic regression classifier trained with the frozen feature embeddings. Stratified 5-fold cross validation technique was used to make sure that each fold contained the same and representative amount of

each demographic class. We employed liblinear solver using a fixed regularization parameter, $C = 1$ for the reproducibility. The measurement of Macro-F1 Disparity in intersectional subgroups is stable and accurate, thanks to this standardised approach.

Task B: Identity Verification (Cosine Similarity)

For the Balanced Faces in the Wild verification task (1:1 matching), we directly computed the similarity scores of the embeddings. We took a pair of images, (x_1,x_2), and computed their corresponding feature vectors, v_1,v_2; and compute the *Cosine Similarity*:

$$\text{Score}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1^{\top} \boldsymbol{v}_2}{\|\boldsymbol{v}_1\|_2 \, \|\boldsymbol{v}_2\|_2} \quad (1)$$

Then we computed the optimum threshold to meet False Positive Rate (FPR) at 10% False Negatives (FN) and compute True Positive Rate (TPR) (the probability that it is accepted) at that FPR, which would be different for the different models, which is 1 out of 100. We then examine the *Max Equal Opportunity Difference* (Max EOD) which measures the gap between false rejection rates of different demographic groups, and highlights the group who is worst affected by 'false rejections'.

3.3 Metrics Used

We chose evaluation metrics based on known legal and ethical criteria when auditing algorithms.

3.3.1 Accuracy Disparity

TTo measure the "worst-case" scenario, we computed the Accuracy Disparity (*AD*) values [4]. This measure is also a measure of the bias gap, indicating differences between the top performing group (G_{max}) and the bottom performing group (G_{min}, which would be the best and worst demographic groups respectively):

$$\text{AD} = \text{Acc}(G_{\text{max}}) - \text{Acc}(G_{\text{min}}) \quad (2)$$

3.3.2 Macro-F1 Disparity

Accuracy Disparity indicates the absolute difference between the performances, but the raw accuracy might hide the underlying distributions of errors in multi-class attribute prediction. For a rigorous assessment of Demographic Classification (Task A), MacroF1 Disparity was used. The F1-Score is a harmonic mean of Precision and Recall that prevents that either False Positives or False Negatives are unfairly taken up by certain demographics. The difference between the highest and lowest F1-Scores for the groups analysed is computed:

$$\text{F1}_{\text{disp}} = \text{F1}(G_{\text{max}}) - \text{F1}(G_{\text{min}}) \quad (3)$$

A large $F1_{Disparity}$ value indicates that the model is not reliable and inconsistent. It works well for some groups but not for others, as it is not able to recognize the unique face patterns of minority groups and classifies their data as a 'blur.

3.3.3 Verification Fairness Metrics (EOD and Equity)

For the verification task (Task B), the Equal Opportunity Difference [11] was chosen as a priority. This is especially important for security systems as it quantifies the difference in False Negative Rates (FNR). A high EOD suggests that there are various groups who are more likely to be "locked out" or excluded:

$$\text{EOD} = \max_{g \in \mathcal{G}} \left| \text{TPR}_g - \text{TPR}_{\text{overall}} \right| \quad (4)$$

Successful successful legit access is measured by TPR (True Positive Rate).

For the sake of context, we also report an absolute error gap as well as a normalized Fairness Metric. TPR is measured relative to the lowest-performing demographic subgroup (G_{min}) using typical algorithmic auditing practices.

$$\text{Fairness Metric} = \frac{\text{TPR}(G_{min})}{\text{TPR}(G_{max})} \quad (5)$$

A Fairness Metric closer to 1.0 (or > 0.80), suggests that the verification performance is more similar for the different demographic groups when the model is applied through the same decision threshold. On the other hand, a lower value means that the groups perform differently more during the identity verification.

3.3.4 Justification of Metric Selection

We chose fair measures that would reflect differences in the quality of the models. To measure the consistency of performance of each model across the different demographic groups, we chose two measures for demographic classification: *Accuracy Disparity* and *Macro-F1 Disparity*. These metrics reflect the differences in reliability for classification and works to take into account class imbalance.

Table 1: Task A: Intersectional Bias Analysis (FairFace Classification). Mean ± std of results over Stratified 5-Fold Cross-Validation. Lower Bias Gap and $F1_{Disparity}$ mean a better embedding representation for all demographic groups.

Model	Accuracy	Acc Std	Accuracy Disparity	Gap Std	F1 Disparity	Best Group	Worst Group
ArcFace	84.41%	0.0022	0.148	0.0215	0.044	Middle Eastern Male	Black Female
GhostFaceNet	86.40%	0.0026	0.158	0.0096	0.046	Middle Eastern Male	Black Female
VGG-Face	87.27%	0.0021	0.168	0.0197	0.049	Middle Eastern Male	Black Female
SFace	81.64%	0.0055	0.167	0.0169	0.052	Latino Hispanic Female	White Female
FaceNet	82.92%	0.0036	0.217	0.0245	0.066	Middle Eastern Male	Black Female

To verify a user, we need a set of systems that are able to accept legitimate users, so we chose *Equal Opportunity Difference (EOD)* and the TPR-based *Fairness Metric* for Task B - Verification. EOD calculates an average TPR of the whole dataset and then measures it against every model's TPR. Such metrics quantify variation in True Positive Rates among demographic groups and are more suitable to assess fairness for biometric security systems.

3.4 Explainability via SHAP and Occlusion Sensitivity

Applying two model-agnostic explainability techniques, SHAP and Occlusion Sensitivity, to the models, we investigated how their decisions are made without looking into their original training code. Such was done for both demographic classification (Task A) and for identity verification (Task B).

SHapley Additive exPlanations (SHAP):

SHAP assigns an importance score to each "input region" to quantify its contribution to the overall output to the region by the model. SHAP is like a spectator, looking at how the model changes the output it outputs as the input changes, rather than analyzing what it considers in its calculations.

SHAP shows the parts of images that are important for the prediction of image attributes, like age, gender and ethnicity, for the demographic classification task (Task A). It will help to see if the model uses the real face, or peripheral features.

SHAP computes the regions on which the decision to make a match or non-match for the identity verification task (Task B) depends. A positive SHAP value means pixels that make the model more likely to make a match, and negative SHAP values mean that pixels make the model more likely to reject a match. Contrastive Occlusion Sensitivity:

As a complement to SHAP, we mask or hide small portions of the input image to assess the impact on the model output, which is also called Occlusion Sensitivity. A large shift in the model's self-confidence after masking the region is regarded as an important region for recognition. Both Task A (FairFace) and Task B (BFW) were shown to be occlusion sensitive to facilitate comparing spatial attribution behavior between architectures.

4 Results and Analysis

This section makes a comparison between the five different AI models after the testing conducted under the same scenario.In this section, comparison of the five different AI models under the same testing conditions is done. We consider two aspects: Models' guesses on people's characteristics such as age/gender (Task A) and models' guesses on face identification (Task B). We assess whether the models are equally reliable for all members of the population or whether they are more reliable for some groups than others using the fairness rules outlined in Section 3. For solid and reproducible results, we used 5-fold stratified cross-validations for the classification tasks and a more strict 10-fold cross-validations for the verification tasks.

4.1 Demographic Classification and Intersectional Bias (Task A)

Intersectional Representation Bias: As with other large studies such as *Gender Shades* [4], our analysis of the FairFace dataset (Table 1) shows Black Females are the most vulnerable group. In all of the models, the internal map (embeddings) was the least effective for Black Females, with Middle Eastern Males consistently having the highest scores.

Most importantly, the old FaceNet architecture presented the greatest representation bias with the highest Accuracy Disparity (0.217) and the highest Macro-F1 Disparity (0.066). This matches well with our results from Section 4.6 depicting the fragmented features that make FaceNet unreliable for minority groups. In contrast, ArcFace was the most competitive, and fair ($F1_{Disparity}$ = 0.044) method, and the much lighter and faster GhostFaceNet ($F1_{Disparity}$ = 0.046) showed that mitigating fairness does not need to be at the expense of speed and simplicity.

4.2 Age-Related Performance Degradation

In addition to the phenotypic characteristics, there was a general performance decline in the 0–2 Age Group (Infants and Toddlers). All of the models evaluated exhibited the highest Accuracy Gap (¿ 28%) and had a significant F1 Disparity (¿ 0.30) in the ageing category of a face less than three years old, detailed in Table 2.

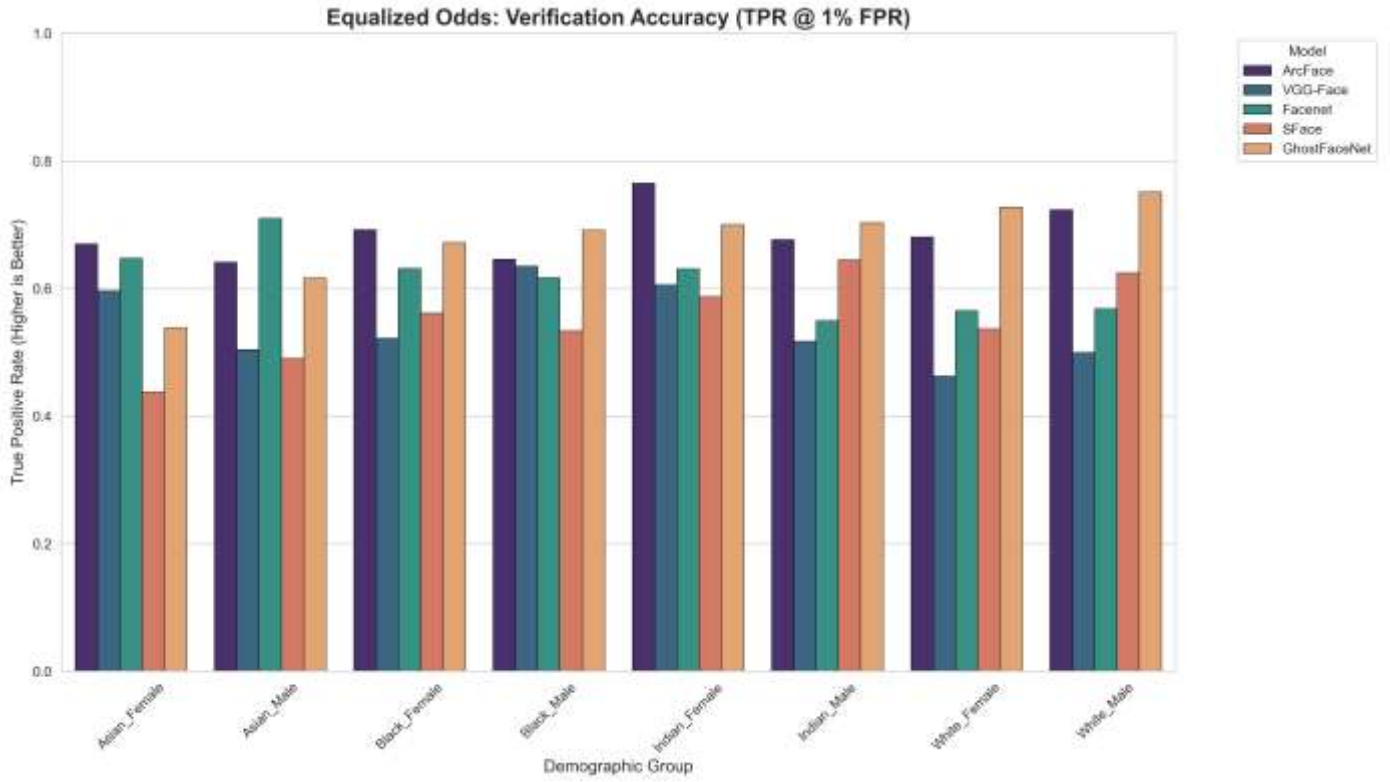


Figure 2: Equalized Odds Analysis of BFW (5 Models). This chart provides a comparison of Verification Rate (TPR) for each of the demographic subgroups. The own fairness profile of the GhostFaceNet (Orange) is competitive, it significantly outperforms SFace (Red) and approaches the stability of ArcFace (Purple).

This poor performance is probably because infant's bones are not fully developed and they don't have distinct facial features. In particular, VGG-Face achieved the lowest age-related bias gap (0.281) and the lowest Macro-F1 Disparity (0.302), which may indicate that this is due to its representational capacity (4096-d), but differences in the age distribution of the old training data cannot be excluded. By contrast, the over-compressed SFace had the worst F1 Disparity (0.340), so it had serious problems when applied to distributions outside the dominant adult distribution. Importantly, this novel architectureGhostFaceNet was able to outperform other older architectures, such as FaceNet (0.334), even under strict constraints on the number of parameters. Thereby it is again confirmed that it has an excellent mechanism of feature extraction, which provides for demographic equality across various axes of variation.

Table 2: Task A: Age Bias Analysis (FairFace Classification). This is assessed by Stratified 5-fold Cross-Validation.

Model	Bias Gap	F1 Disparity	Worst Age	Best Age
VGG-Face	0.281	0.302	0–2	30–39
ArcFace	0.292	0.314	0–2	40–49
GhostFaceNet	0.297	0.324	0–2	30–39
FaceNet	0.306	0.334	0–2	30–39
SFace	0.309	0.340	0–2	40–49

4.3 Identity Verification Disparities (Task B)

In the verification domain (BFW), bias is evidenced by a difference in True Positive Rate (TPR) between subgroups at a given False Positive Rate. The TPR breakdown of the five architectures analyzed and audited is depicted in figure 2 for all the demographic groups.

- model with the highest fairness score (0.837) is ArcFace (Purple) which performs the most consistently of all the models.

It has the highest scores for the Indian Females and the lowest scores for Asian Males, but the range of scores between the top and bottom group is very small (0.125). This indicates that its complicated mathematics are able to treat various groups equally.

- There are significant group differences in the reliability of VGG-Face (blue) and SFace (red). VGG-Face is pretty good at recognizing Black Males, but much less successful at White Females. Likewise, the smaller SFace model works best for Indian Males, and performs poorly for Asian Females, with the least fairness score (0.692). From these results, it is observed that for certain identities, the outcomes of these old designs can also be unstable and making models too small can also bring in unstable results for certain identities.

- A surprise is GhostFaceNet (Orange). Although it is designed for use on mobile devices, it is fairly accurate albeit in the same trend as the larger ArcFace model (0.734). This is an example that a model doesn't need to be large to be fair; its design is more important.

Overall, it was found that the increased amount of compression of a model and the way a model is built are directly related to its fairness as demonstrated by the verification tests. Other models are the same for all others, others become much less reliable for certain groups under the same security settings.

4.4 Pose Control and Methodological Robustness

In order to maximize attention to the model's own bias, we were interested in filtering out the FairFace dataset (Task A) in order to get rid of the "noise. We discarded faces turned at a high angle (more than 30° turned to the side or 20° turned up/down) from the pipeline, InsightFace. This reduced our images to 90,000 to a small group of high quality images about 50,000 of which were clear and front facing images. It was essential to clean the data in this way because of two reasons:
Speed: It made testing each of the five models very speedy. Justice: It excluded poor performance due to bad pose. When the image is blurry or is not looking directly at the camera, it can be difficult to determine whether the AI is biased or whether the image was just too challenging to be read by the AI. If we have only clear and frontal images, we can feel more confident that the inaccuracies—such as things that we've seen in FaceNet and SFace—are due to the embeddings (internal map) of the model, and not from camera angle alone.
Preserve Variance for Verification (Task B): For the verification task, we used a different approach to the Balanced Faces in the Wild (BFW) dataset.
But first, the developers of the BFW dataset have already provided the images as "face-prints" – which were already pre-aligned and cropped. We did not need to use our own detection tool on top of their work and might have added some new errors.
Second, the point of a "verification" test is to see how these models work in the real world. The classification task (Task A) should be as "clean" as possible so as to be able to visualise the model mapping inside, but the verification task (Task B) should be more realistic. In the real world, the security system must cope with people facing various directions or angles. These natural variations in BFW data ensure that this data more accurately reflects the performance of the model in a real security scenario.

Table 3: Task B: Identity Verification Fairness Metrics (BFW). Trained and tested at 1% FPR, 10-fold cross validated.

Model	EOD	Bias Gap	Fair Metric	Best Group	Worst Group
ArcFace	0.0793	0.1254	0.8372	Indian Female	Asian Male
FaceNet	0.0831	0.1548	0.7793	Asian Male	Indian Male
VGG-Face	0.0959	0.1706	0.7335	Black Male	White Female
GhostFaceNet	0.1278	0.1967	0.7342	White Male	Asian Female
SFace	0.1136	0.1968	0.6917	Indian Male	Asian Female

4.5 UMAP Analysis of the Embeddings

To understand why some models are more biased than others, we have employed Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [14]. The tool maps out these multi dimensional embeddings—the digital signatures that the AI has created for faces—in a simple 2D map to visualize the organization of different groups, and presents them to us.
ArcFace (The Balanced Map): The dots in the ArcFace map (Fig. 5a) are intermingled between the racial groups. A small number of "islands": Single races. This implies that the construction of ArcFace contributes to it having an individualistic approach instead of just a race-based approach. The lower gap in reliability $F1_{Disparity}$ can be attributed to the groups being fairly mixed, thus giving the model even treatments of all the groups.
FaceNet (The Fragmented Map): The FaceNet map (Fig. 5b), however, demonstrates the "clumping," which can be seen. It is possible to observe individual spaces where certain racial groups are being pushed out of the other groups. This "fragmentation"

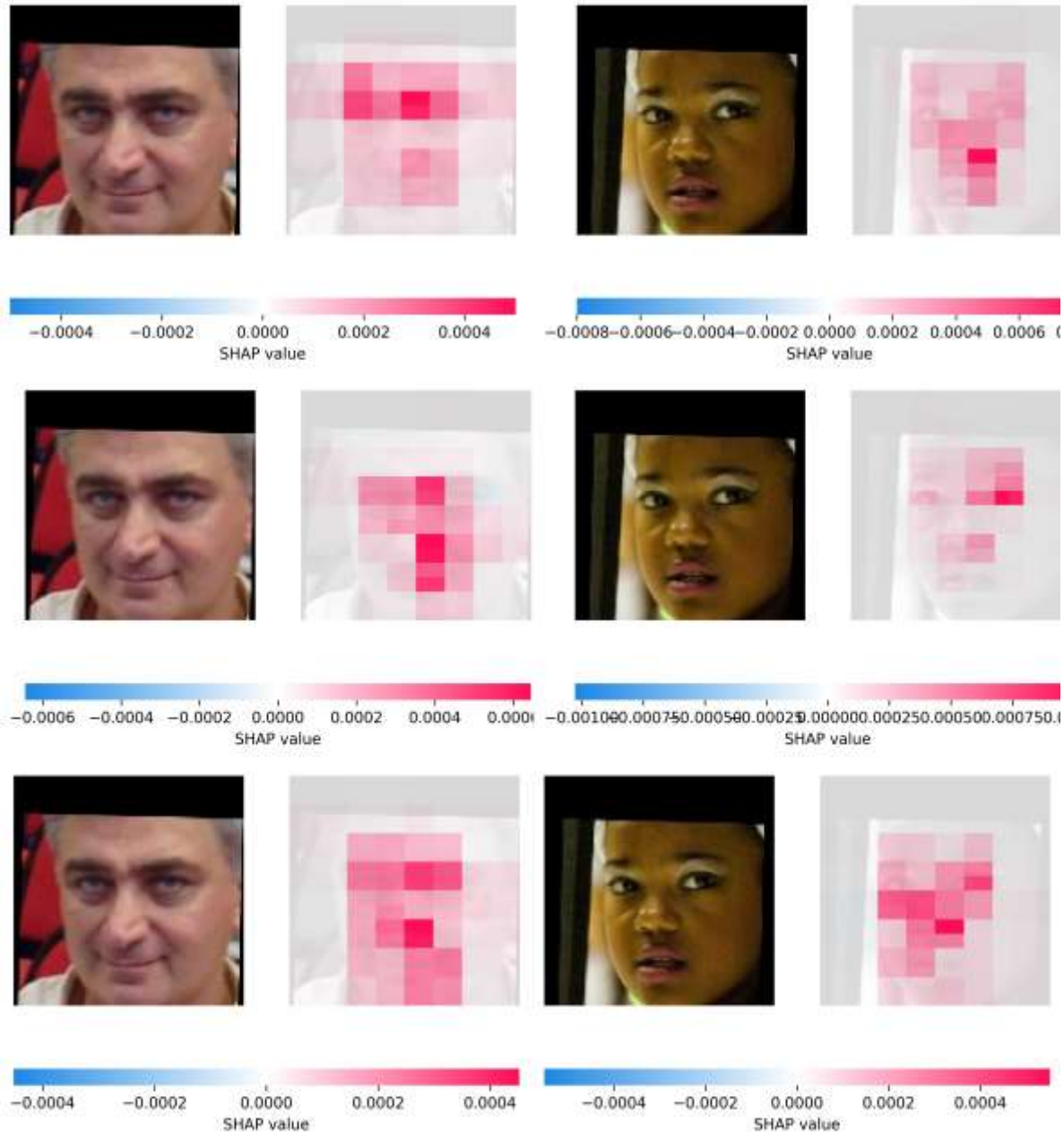


Figure 3: Task A (Classification): SHAP Feature Importance on FairFace. This visualization represents which area(s) of the image is being focused by the models. Both ArcFace (Top) and GhostFaceNet (Bottom) remain attuned to important features of the face such as the nose and the eyes. FaceNet (Middle) on the other hand suffers from ”attention drift,” using background and/or hair artefacts for the classification of minority groups, which results in less reliable classification results.

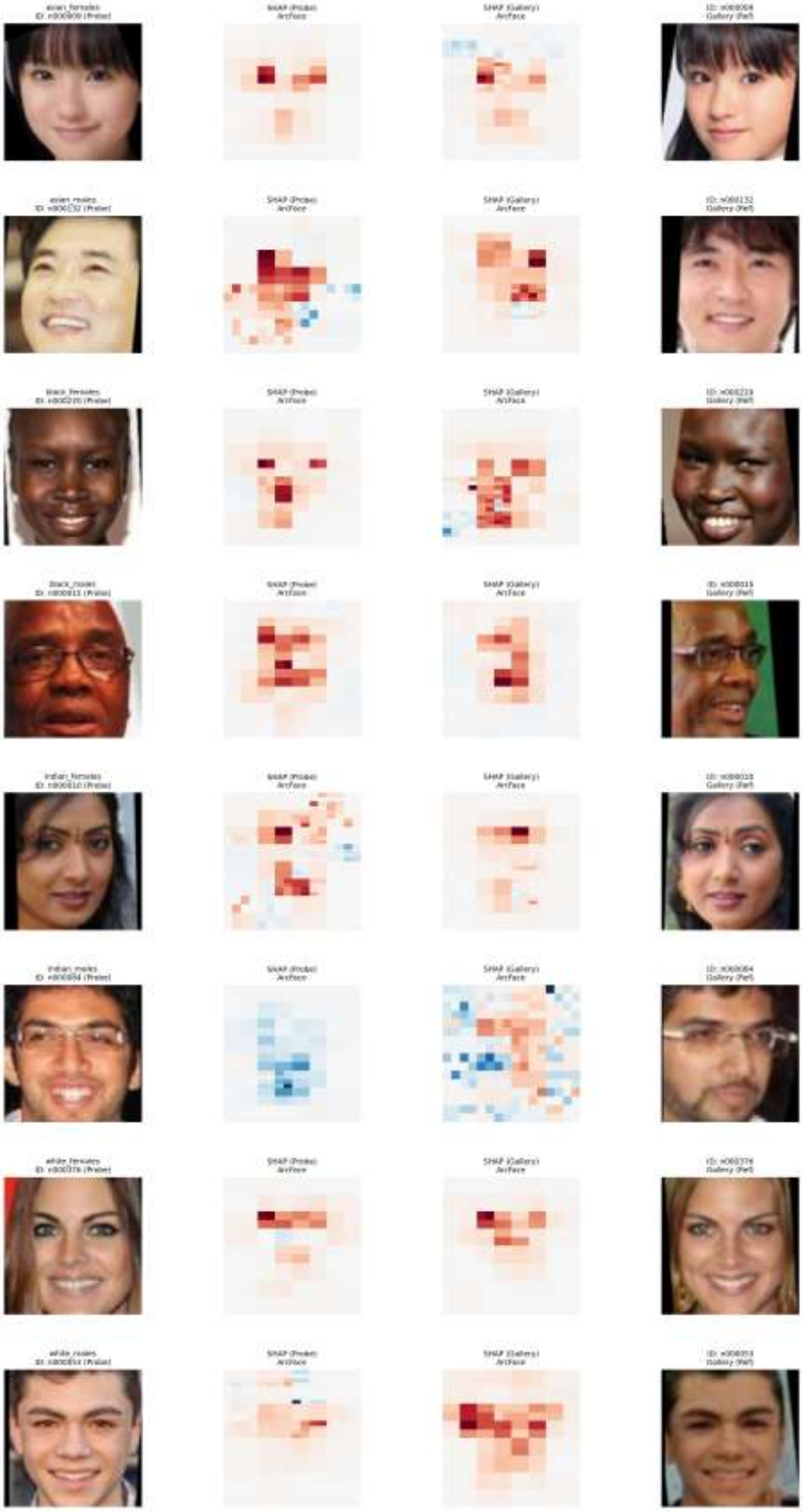
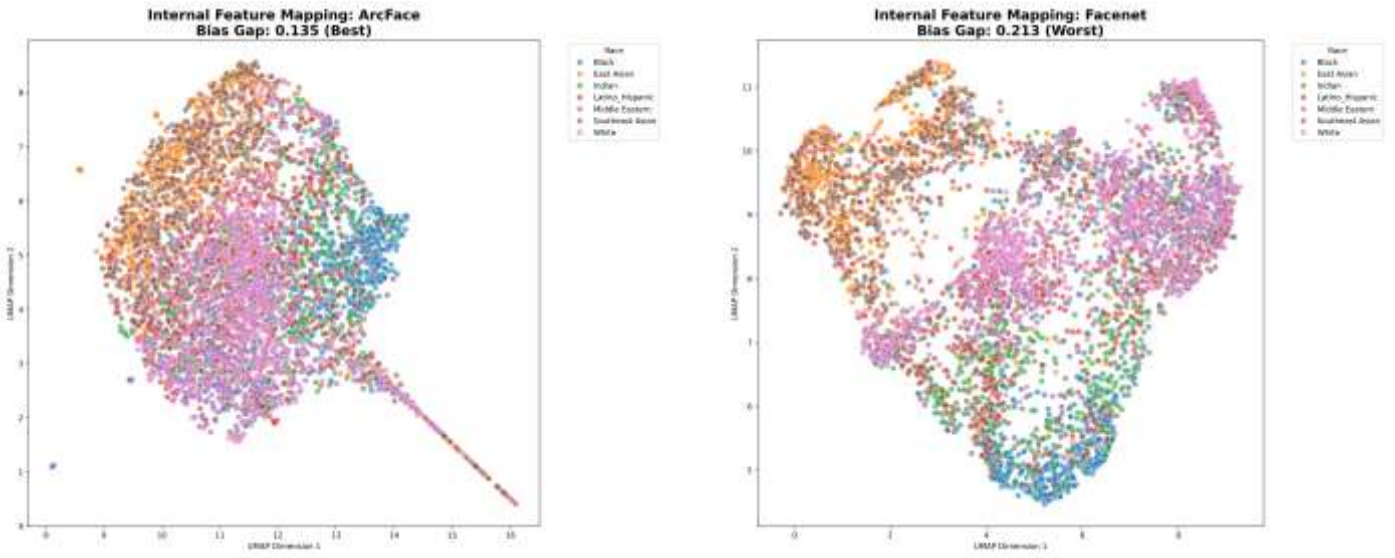


Figure 4: Task B (Verification): Intersectional Pairwise SHAP Audit on ArcFace. The matrix compares image pairs split equally amongst eight demographic pairs (Asian, Black, Indian, and White by gender), looking at the spatial attribution patterns both for the probe input images (left column) and the gallery reference images (right column). The high intensity of the red zones illustrates the most important landmarks affecting a positive face similarity evaluation in terms of biometric data.

refers to the fact that the internal map of the model is divided into various zones in a way that is different for various individuals. The more this mess the map is in, the bigger the chance of being wrong in the view of the AI, such as getting both persons of a



(a) ArcFace (Integrated)(b) FaceNet (Segregated)

Figure 5: The embeddings of the (FairFace Dataset) under UMAP are visualized and show (a) a balanced distribution and (b) the scatter of the points of different races and thereby provide evidence of the bias in the demographic.

minor group confused, or not recognizing them at all. This explains why FaceNet had the highest reliability gap in our tests.
The overall point of these maps is that the less fair the model, the more ”blended” the internal map becomes. Those models that do not separate groups are likely to be more stable and reliable for all.

4.6 Explainability Analysis

Here we visualize the models using SHAP (Lundberg et al., 2017) and Occlusion Sensitivity (Zeiler et al., 2014) to learn what they ’look at’ when they make a decision. The aim is to determine whether there are any visual patterns in these heat maps that are consistent with the fairness numbers we calculated above.
Task A: What the Models See during Classification (FairFace)
To interpret these results, we looked at the heatmap of their respective models to understand what each of them ”looks at” (Figure 4). A good model should be based on the characteristics of the face, such as the nose, eyes, mouth, which are related to the race.
Consistent Focus: Both ArcFace and GhostFaceNet display a central focus. They are more fair and accurate across groups as they look at the most important structures of the face for all.
But FaceNet has its moments. When it comes to minority groups, its eye is often drawn to the hair or background, rather than the face (as indicated in the visual grids). That implies that the model is learning incorrectly from the ‘noisy’ clues, rather than from the next obvious facial features, which is why it does not do as well in these categories.
Task B: What the Models See during Identity Matching (BFW)
The overall results of the verification tests suggest that the structure and degree of compression of the model have a direct correlation with the fairness. There are some models that remain consistent across all users, and others that are much less reliable to certain groups, given the same level of security.
The resulting feature attribution heatmaps show a distinct difference in the focus of the models at the local level when matching two images of the same person. The local spatial attribution maps show scattered or indistinct clusters of deviations, which for the sub-populations with less verification fidelity for a given network, largely move away from the anatomical structures. This translates directly into greater false rejection rates, excluding people who are supposed to be within the biometric security system.
Stable Core Focus (ArcFace & GhostFaceNet): Modern architectures such as angular margin (e.g., ArcFace) and resilient and lightweight architectures (e.g., GhoutFaceNet) show excellent and sharp, symmetric feature anchoring as shown in Figure 4. The SHAP attribution layers (concentrated deep red segments) stay firmly attached to the ocular canopy, perinasal T-zone and mouth plane. These models focus calculations on these bone structures that remain relatively constant across the population and filter out non-biometric factors such as hairstyle changes, cosmetics or peripheral skin temperature disparities to enforce strong security margins between different demographics.
Fragmented Attention and Negative SHAP Deviations (FaceNet): Older Euclidean models, in turn, such as FaceNet, exhibit clear spatial instability and attention drift. The spatial focus for the lowest performing segments (for example, Indian Male and Indian Female) deviates from central facial features for these compute similarity scores.
More importantly, as described in the bottom half of Figure 4, deep blue pixel patterns are seen throughout the face of the target. For each blue block, explicit parts of the image that actively deceive the network into deciding against a match, the feature is called negative SHAP feature. For other subjects, such as the Indian Male, when you add in reflections on their eyes, facial shapes, very slight postural adjustments, etc., the focus becomes diluted, and they have competing positive (red) and negative (blue) pulls on

their features. This pattern of a non-conforming image is clearly a distraction from the deep biometric geometry and is the primary reason for the high False Rejection Rate of FaceNet.

Summary of Visual Patterns in SHAP:

- Steady Focus (ArcFace): The model always emphasizes the same basic facial structures in each racial and gender group. This local geometric discipline is similar to its high geometric fairness scores.
- Scattered Attention (FaceNet): This older architecture exhibits significant saliency shift, drifting toward non-biometric textures and peripheral noise when confused. This vulnerability directly impacts minority groups.
- Efficient Stability (GhostFaceNet): Despite extreme resource and parameter compression, this architecture maintains highly focused multi-scale feature tracking. This proves that lightweight mobile deployment can be achieved without compromising demographic equity.

4.6.1 Occlusion Sensitivity

Occlusion Sensitivity was used to verify our SHAP results. The approach described here is to systematically ”block” small squares of the face to observe the face’s confidence as it changes. When the accuracy of the model decreases at a particular area, we know that the area is important.
Key Findings (Figure 6):
FaceNet and ArcFace mainly target the central part of the face when relying on the Face Center in groups whose model performances are good (such as “Race Best”).
When the models have trouble (such as ”Race Worst” or ”Infants”), ”Edge” is the problem - FaceNet is very sensitive to the hair and background. However, the model’s ”focus” is moving away from the eyes and towards the edges of the image as seen in Figure 6(b) and 6(d). Using unreliable clues to make a decision is indicated by this.
Age and Structure: For adult faces, models are based on distinct facial features such as the jawline and cheekbones. These features are less well-developed for infants (0–2 years). The problem is that something was scattershot in face detection; ArcFace remains fairly focused on the face (why it is so accurate for young children) while FaceNet got cumbersome and less accurate.
This is a summary of the Visual Audit. The results of our visual tests (SHAP and Occlusion) are comparable to our numerical results overall. The most consistent and fair models will focus on the center of the face. There are models that are ”distracted” by background or hair, such as the older FaceNet models, which have the greatest differences in reliability among groups.
While these images don’t necessarily capture how the AI “thinks”, it is evident from the pattern that a steady focus creates a fairer model making.These images are an attempt to give one representation of how the AI “thinks” but it is clear from the pattern a steady focus creates a fairer model making.

5 Limitations

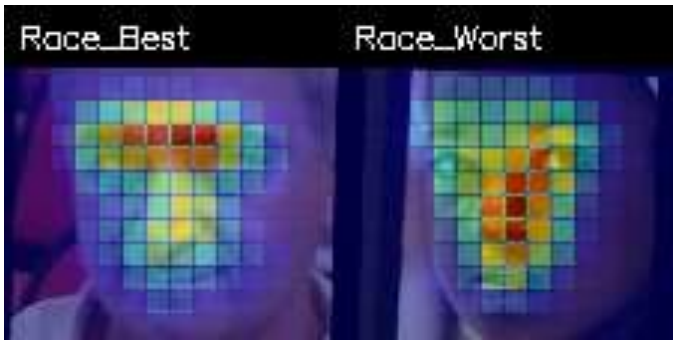
This study illustrates various face recognition models in detail but there are a few drawbacks in our approaches.

5.1 Model Training and Data

These models were used as ”gray-box” models. This would have made it possible to view the final embeddings–the digital face-prints–and the weights of the models but we were unable to control the way the models were actually trained. Therefore, it is not possible to precisely distinguish between the ”data” that the model was given for training and the ”design” itself.
This is because the more recent designs of ArcFace and GhostFaceNet [7, 2] may be responsible for the improved performance, or perhaps because they were trained on newer, cleaner datasets. The older models which were trained years ago, such as VGG-Face and FaceNet [15, 18] were trained on datasets known to be more imbalanced in some way. If all models were retrained on the same data, it would be more accurate, but this was not part of the scope of this project. The models can be evaluated in order to get a more realistic view of their performance in real world apps.

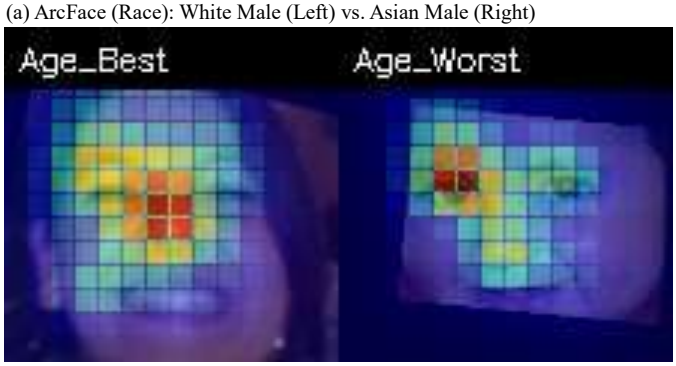
5.2 Broad Demographic Groups

The labels for the FairFace and BFW datasets were used for our analysis; these labels use broad categories such as ”Asian,” ”Black,” and ”White. We note that these categories are a social classification and not hard and fast biological groups.
These large categories may contain more specific and significant differences that might be differences in skin color or ethnic subgroups within a category. Therefore, the gaps in performance for more specific identities found in our study may be underestimated. Also, the identities that we could use were limited to what was in our data sets.

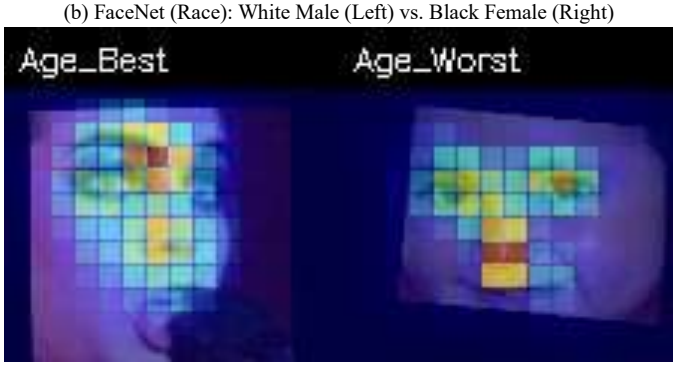


(a) ArcFace (Race): White Male (Left) vs. Asian Male (Right)

(b) FaceNet (Race): White Male (Left) vs. Black Female (Right)



(c) ArcFace (Age): Adult (Left) vs. Infant (Right)



(d) FaceNet (Age): Adult (Left) vs. Infant (Right)

Figure 6: Occlusion Sensitivity Across Models. (Left Column: ArcFace) the image is consistently focused in the center of the face. It is quite dependent on the area of the eye in its best performing groups; for poorer or weaker groups it still maintains its attention on more fundamental features such as nose and mouth. (Right Column: FaceNet) shows a more ”scattered” pattern for its lowest-performing groups, such as minority identities and infants. If so, the model will be distracted by inconsequential details such as hair, background, and edge of the image. These are visual patterns which are congruent with the greater fairness gap and issues with reliability from our previous numerical test.

5.3 Explainability Tools

The tools previously mentioned to view model behavior after reaching the decision (SHAP and Occlusion Sensitivity) offer a view into the behavior of the model. These heatmaps indicate on which area of a face the model focuses, but not on the actual ”hidden” math or neuron activity inside the model.

SHAP correlations demonstrate how focus relates to fairness, but do not necessarily paint a full picture of the model’s inner workings. A more detailed “white-box” study, focused on each individual layer, could provide further detail in the future.

5.4 Data Choice and Ethics

The FairFace and Balanced Faces in the Wild datasets were selected here because they were carefully created to be balanced. Some older datasets were also criticised for the insufficient number of images of lighter-skinned males, which makes it difficult to assess fairness in tests. Balanced data would allow us to find out how the model functions, instead of revealing a problem with the data.

We did not go with much larger quantities of images, e.g. more than a million of AI-generated images. Each of these tested models was already a large undertaking, taking considerable time and computing power to generate embeddings for the 70,000+ images included in this study, and by concentrating on these “benchmark” models, the authors were able to conduct in-depth comparisons of the individual models.

6 Conclusion

This study seeks to make comparisons between five different face recognition models [7, 15, 18, 2, 3] to understand how each model’s design impacts the level of fair performance. We tried a few big and cumbersome models, and a few small and “light” models for mobile devices. By using the same testing rules for every model including cleaning the FairFace [12] data for better accuracy and using the realistic Balanced Faces in the Wild [16] data for verification—we were able to see how each model handles different demographic groups.

Our study is inspired by earlier studies in which size and architecture have been shown to be as relevant as data to achieve fairness, as in recent work by [5]. Key Findings

We show that the model construction process can significantly influence the fairness of the model towards individuals.

Most Reliable: Modern models such as ArcFace, which offer greater capacity were most stable. They had the smallest racial and gender gap in performance.

The other models like the older one FaceNet and the very small one SFace, had much larger differences. They were much less accurate with the minority population and infants, resulting in more “false rejection” rates for the minority users and infants.

Efficient and Fair: one significant result is that GhostFaceNet, being both small and fast, is nearly as fair as the larger models. This is proof that it is feasible to make a model smaller when using one with a mobile phone without making it less fair. Visual Evidence Our tests used this procedure with these numbers. We found that the internal maps of fairer models were nicely mixed whereas internal maps of biased models were "clumpy" or "messy", as the maps used tools like UMAP. In addition, the SHAP and Occlusion heatmaps remained in the center of the face (eyes and nose) in fair models. Instead, however, a biased model is distracted by the background or hair when it is confronted with the minority groups.

The overall result that can be deduced from this study is that there is not a single measure of the 'total accuracy' of a model that is enough. The design of the model can also result in significantly varying levels of fairness, even if the models are tested under the same conditions. As face recognition systems are increasingly used in everyday life, it is crucial for these issues to be standardised and measured by, and built into, face recognition systems during their development and selection.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[2] Mohamad Alansari, Oussama Abdul Hay, Sajid Javed, Abdulhadi Shoufan, Yahya Zweiri, and Naoufel Werghi. Ghostfacenets: Lightweight face recognition model from cheap operations. *IEEE Access*, 11:35429–35446, 2023.

[3] Fadi Boutros, Marco Huber, Patrick Siebke, Tim Rieber, and Naser Damer. Sface: Privacy-friendly and accurate face recognition using synthetic data. In *IEEE International Joint Conference on Biometrics, IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022*, pages 1–11. IEEE, 2022.

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018.

[5] Eduarda Caldeira, Pedro C. Neto, Marco Huber, Naser Damer, and Ana F. Sequeira. Model compression techniques in biometrics applications: A survey. *Information Fusion*, 114:102657, 2025.

[6] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019.

[8] European Parliament and Council of the European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act). *Official Journal of the European Union*, 2024.

[9] Patrick Grother. Face recognition vendor test (frvt) part 8: Summarizing demographic differentials. *National Institute of Standards and Technology (NIST)*, 8429:8, 2022.

[10] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019.

[11] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[12] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

[13] Ketan Kotwal and Sebastien Marcel. Demographic fairness issues survey in face recognition. *Computer Vision and Image Understanding*, 2025.

[14] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[15] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. volume 1, pages 41.1–41.12, 01 2015.

[16] Joseph P. Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: Too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–1, 2020.

[17] Brinda Sakhiya. Architectural comparative analysis of data preprocessing techniques for large language models: From linguistic fundamentals to scalable cloud-native pipelines. *International Journal of Innovative Research in Technology*, 12(5):2066–2074, October 2025.

[18] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[19] Sefik Serengil and Alper Ozpinar. A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules. *Journal of Information Technologies*, 17(2):95–107, 2024.

[20] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer Science & Business Media, London, UK, 2010.