Mitigating Racial Biasness in Facial Recognition Technology Using Generative Models

Hossenbux Muhammad Yaaseen[1]

Publication Date: 2026/02/17

How to Cite: Hossenbux Muhammad Yaaseen (2026) Mitigating Racial Biasness in Facial Recognition Technology Using Generative Models. *International Journal of Innovative Science and Research Technology*,11(2), 721-742.
https://doi.org/10.38124/ijisrt/26feb145

ABSTRACT

In recent years, several researchers have worked to lessen biasness in face recognition systems since it had produced numerous problems, from people getting falsely accused of minor offences up to murder, getting wrongfully arrested and in some extreme cases, getting killed, and these happened particularly for people of colour. This thesis aims to contribute to the data science field by investigating and implementing generative models as a potential assistance to expand the diversity of datasets used in facial recognition technologies and help mitigate racial biasness in facial recognition systems. The CelebA Dataset, containing more than 100,000 unique photos, was utilized to train our StyleGAN 2 model, to generate synthetic realistic images and the FairFace dataset which has a diverse dataset over 100k images of both males and females for our recognition model. East Asian, African, Caucasian, Indian, Middle Eastern, Latino, and Southeast Asian are the racial categories that have been identified. We used InceptionResnetV1 for feature extraction, MTCNN for face detection and then ran our recognition Model on our diverse dataset with and without synthetic images. After using our own generated synthethic data, we saw accuracy gains for a few of the races, including East Asian, African, Indian, Middle Eastern, and Latino, which demonstrated that the accuracy level increased by more than 10% in some cases.

TABLE OF CONTENTS

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb145

CHAPTER ONE
INTRODUCTION

In 1967, Woodrow W. Bledsoe, a pioneer in artificial intelligence developed a system that could input the coordinates of facial features into a computer, which would then compare the features to a database of known faces. [1] The primary intention behind the system would be to provide aid in law enforcement, particularly for identifying suspects or persons of interest and for security purposes. [1] This application further highlighted the potential of using computational methods for identification purposes, which was a significant step forward in the field of biometrics. [1] Bledsoe's work was groundbreaking for its time even though the early system required significant manual effort. He would take pictures of individuals and then manually record the coordinates of various facial features such as the eyes, nose, mouth, and the outline of the head using a RAND tablet, a device that allowed manual entry of the data points into a computer. One of the significant challenges was the manual nature of the process. It was time-consuming and lacked the efficiency and speed of modern automated systems. The system was also limited in its ability to handle changes in lighting, facial expressions, and angles. [1] However, despite its limitations, Bledsoe laid the foundation for the automated facial recognition systems that are widely used today. The evolution of facial recognition technology is a testament to the rapid advancements in computer science and artificial intelligence. In the 1970s and 1980s, advancements in digital image processing and the increase in computational power led to more sophisticated facial recognition systems. Modern facial recognition technology makes use of artificial intelligence, machine learning, and sophisticated algorithms. These systems can recognise facial traits automatically and instantly compare them with enormous databases. Contemporary systems exhibit greater versatility and accuracy because of their ability to accommodate differences in illumination, face expressions, and angles. [2]

The technology is currently employed in several industries. In the USA, the FBI's Next Generation Identification program uses facial recognition to match suspects' images with a national database of mugshots. [3] Large-scale CCTV networks using facial recognition software are used in cities like London to keep an eye out for illegal activities in public areas. Apple's Face ID technology, used in iPhones and iPads, allows users to unlock their devices and authenticate payments using facial recognition. Some smart home devices, like Google Nest Hub Max, use facial recognition to show personalized notifications and information to different household members. Delta Air Lines has implemented facial recognition technology at Atlanta's Hartsfield-Jackson Airport for a paperless and seamless boarding experience. HSBC Bank introduced the recognition system in its mobile banking app which allows customers to access their accounts securely. Even in the healthcare area where hospitals are using the system to match patients with their medical records, ensuring correct treatment and medication. In the education sector, Universities have implemented facial recognition systems for automatic student attendance tracking and enhancing campus security and monitoring visitor access. [4-6]

While facial recognition technology offers enhanced security and convenience, it also raises privacy concerns. Personalised advertising in retail and marketing can improve customer experience, but it also raises questions about consumer privacy and data use. Similarly, in education and the workplace, the use of facial recognition for attendance tracking must balance efficiency gains with students' or employees' privacy rights. The use of facial recognition by law enforcement agencies like the FBI has sparked debates over privacy rights of the public. Similarly, its use in public surveillance in cities like London has raised concerns about a surveillance state. The continuous use of the technology can lead to a loss of anonymity, creating a society where everyone's movements can be tracked and monitored. Furthermore, this could enable stalkers and harassers to locate and track individuals without their consent. The storage of biometric data also poses a significant security risk. If such sensitive data is hacked or improperly managed, it can lead to identity theft and other forms of cybercrime. These instances show how facial recognition technology is being widely used in a variety of industries, underscoring both its advantages and the moral conundrums it presents. The difficulty of reconciling the benefits of new technology with responsible use and privacy protection is not going away as it gets more and more ingrained in daily life. [7-9]

Additionally, this technology may cause problems to the extent of having people wrongfully accused of crimes, imprisoned and in some instances, even killed, the latter known as "racial biasness". [10]

Researchers and data scientists defined racial bias as follows: When an algorithm, model or system exhibits unfair, discriminatory, or damaging behaviour towards people based on their race or ethnicity. [10]

A facial recognition system is racially biased if it does not perform equally well for all racial groups. For instance, if a system is more accurate in identifying faces of one race compared to another. Bias often originates from the data used to train the system. If the training dataset is not diverse and lacks a representative sample of faces from various racial groups, the system will likely to be less accurate for underrepresented groups. The way algorithms are designed can inherently favour certain racial features over others. [11] This thesis focusses on this flaw of the technology and proposes a solution to mitigate racial biasness in the recognition software.

- *Problem Statement*

In September 2019, a photograph of Joshua Bada was rejected when he applied for a new British passport. The system mistook his lips for an open-mouth. He was told his application was not accepted because he needed to provide a neutral expression and a close mouth, something he had in fact done. However, the recognition system couldn't interpret his lips correctly. The 28-year old said he struggled with the technology as it kept on failing to recognise his features. He further added that it's also a problem that he faced on snapchat with the filters, where it hasn't quite recognised his mouth because of his complexion and his facial features. [12] Another similar incident happened with Cat Hallam, a black woman who lives in Staffordshire. She was frustrated after the system told her that her eyes were apparently closed, and it further couldn't find the outline of her head. She said: "The first time I tried uploading it and it didn't accept it," she said at the time. "So perhaps the background wasn't right. I opened my eyes wider, I closed my mouth more, I pushed my hair back and did various things, changed clothes as well – I tried an alternative camera." The educated technologist added that she found it inconvenient to pay more for a picture taken in a photo booth when others could use free smartphone images. She further said: "How many other individuals are probably either spending money unnecessarily or having to go through the process on numerous occasions of a system that really should be able to factor in a broad range of ethnicities? She said the problem was one of algorithmic bias and could not believe it amounted to racism. [12] Even Tech-giants like Amazon and its very own facial recognition software - Reckognition had a 100% accuracy in recognising lighter males but 68.6% accuracy in recognising darker females. [13]

The market growth for the technology is expected to reach more than £6 billion by 2024, in the USA alone. [14] And while the problems faced by both Joshua and Helen aren't life threatening, if the flaws aren't fixed, the flawed technology could lead to serious misjudgement and wrongful convictions by the police when it comes to identifying suspects or criminals as it carries the danger of misidentifying someone. Conversely, ethnic facial features could intentionally be used to identify some certain groups of people as they are suspected to be more prone to violence and crime [15] It also has the potential of being abused by law enforcement or other organisations for things like constant surveillance of the public. The Chinese government for instance is already using facial recognition to arrest jaywalkers and petty criminals which sparked debates of the line between protecting the public and encroaching on basic civil rights and privacy. [16]

➢ *Gaps & Aim*

The gap in current facial recognition technologies primarily revolves around the issue of racial bias, which manifests in several key areas. Existing datasets are predominantly composed of images from limited racial backgrounds, leading to insufficient representation of diverse racial features. This lack of diversity affects the algorithm's ability to accurately recognize faces from underrepresented groups. The biases in the data are intrinsically carried over into algorithms as they are frequently developed and refined using these limited datasets. Face recognition technologies as a result become less precise and may even discriminate against specific racial groupings. There is a substantial deficiency in the comprehensive examination and verification of these systems in various racial groups. Facial recognition technology's practicality in diverse racial contexts is frequently untested or understudied.

Given these gaps, the aim of this thesis is to explore and implement generative models as a solution to enhance the diversity of datasets used in facial recognition technology. This includes employing advanced generative models, like Generative Adversarial Networks (GANs), to create a rich and diverse set of facial images.

These models will generate synthetic faces with varied racial features, contributing to a more representative dataset. Augmenting existing facial recognition datasets with these generative model-produced images, thereby addressing the current lack of racial diversity.

This approach provides a more balanced dataset for training facial recognition algorithms. Furthermore, using these enriched datasets to train facial recognition systems, with a focus on improving their accuracy and fairness across all racial groups and conducting extensive testing of these improved systems to ensure their performance is validated across a wide spectrum of racial groups.

This thesis aims to establish a precedent in the face recognition space by using this study to make it a common practice to employ generative models for dataset diversification. The intention is to persuade the industry towards more ethical and equitable technology development. By tackling the important problem of racial bias, this thesis aims to close the present gap in facial recognition technology and eventually contribute to the creation of more equitable, accurate, and socially conscious technology.

CHAPTER TWO
LITERATURE REVIEW

The existence of bias in deep neural networks has been verified by numerous studies in recent years [17], [18], which could have unfavourable effects, particularly for face recognition [19]. A number of publications concentrate on minimising bias specifically, either by introducing data sampling procedures expressly for this goal or by actively altering the models.

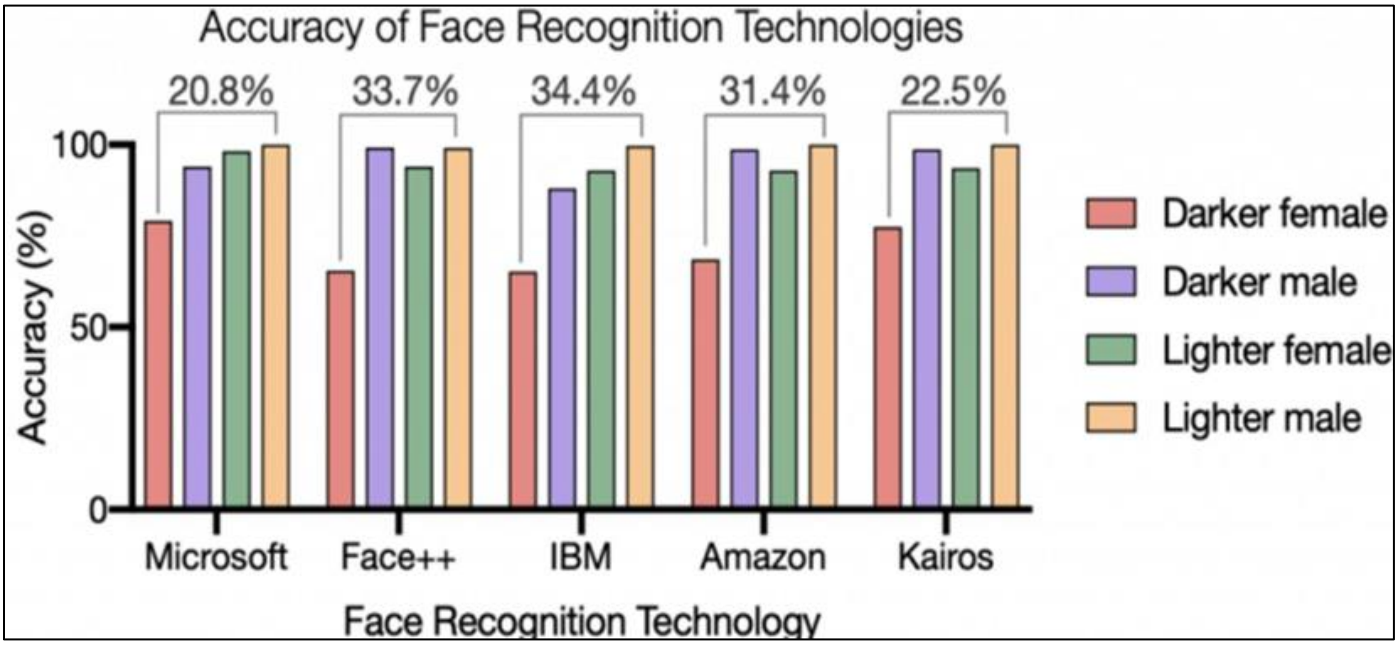


Fig 1 Existing Face Recognition Accuracy Levels Of Large Companies

Regarding model adjustments, [20] propound a model to balance representations of face data from other datasets, in addition to introducing a balance dataset. [21] use attention techniques and adaptive convolution kernels to reduce model bias. [22] take triplet loss into account to avoid discriminatory impacts. Additionally, [23] suggest learning an agnostic representation and protecting privacy to conceal sensitive data, such as gender and ethnicity. In terms of sampling techniques, [24] reduces bias by using dataset sampling based on reinforcement learning. [25] Introduces how sampling techniques can be used to reduce regional variations in picture ID document performance. [26] employs adversarial learning and demographic classifiers to strengthen face representations. [27] uses Cycle-GAN to racially balance training per individual in the sample. Additionally, [28] uses a balanced sampling technique during training to reduce bias, even though their study focuses exclusively on gender. A study was conducted whereby a hybrid algorithm combining the Gaussian Model and the Explicit Rule Algorithm was used to improve the detection of dark-skinned faces in face recognition systems. Morphological and anthropological techniques have been combined in this hybrid system to detect faces.

Additionally, the hybrid system has been verified using face corpora and eye blinking. The findings demonstrate that 87% accuracy was attained when the Gaussian model was used for skin detection. Conversely, 71% accuracy was achieved using the explicit-rule method. [29]

However, the accuracy increased to 89% when the hybrid Gaussian and Explicit rule was used. CCTV (Closed Circuit Television) cameras were used in an experiment to record digital images from video feeds. These digital photos were moved to prepare the portions. For skin recognition in an image, skin colour has always been the most important factor. However, due to racial variety and differences in skin tone, there may be colour variations. Another significant variable that had an impact on the outcomes was light. To distinguish between areas coloured like skin and those that weren't, the image was divided into distinct pixels. [29] A method that combines RGB and YCbCr values depending on thresholds has been used. The threshold range has been determined by taking into account the following factors:• Effect of illumination on the environment; • Features of the face, such as age and gender; • Colours, blurriness, and shadows in the background. In the end, face detection and segmentation were accomplished using skin detection/projection techniques. The skin detection process started by collecting skin samples. For training, over three hundred digital photos were collected. The training data set's mean, variance, and covariance were then estimated. The Gaussian model was then used to simulate human skin to identify the pixels that weren't skin. In addition, a greyscale picture representing the potential skin areas was produced. The following stage involved applying threshold methodology to further separate the skin and non-skin parts by transforming a greyscale image into a binary skin map. Explicit rules from 2-dimensional and 3-dimensional colour spaces were used to model human skin. [29] These guidelines specify a cutoff point at which the skin and non-skin regions of the image's pixels were separated. This method has already been determined to be the most successful method for classifying skin tones. To highlight the skin area for black skin colour tones, the CbCR digital images were applied to the explicit-rule skin segmentation technique. The AND operator was used to integrate the results of two algorithms, or the Explicit rule and the Gaussian model, for obtaining the skin region results, the RGB pictures were converted to greyscale. The use of Canny Edge Detector was made for binary conversion. The image's stance was adjusted using the location of the eyes. With the use of both horizontal and vertical projection, the eye pupil centre was found. The face alignment was done using the x0, y0 coordinates. [29] Initially, the angle rotation of the facial image—which is the angle of elevation—was used to make the line connecting the two eyes horizontal. This procedure, referred to as "de-skewing," increases the algorithm's efficacy. [30] The outcomes demonstrated that the hybrid Gaussian and Explicit Rule algorithm best increased the face detection rate for individuals with dark complexion. Based on the accuracy rate in skin detection and data preprocessing, a comparison has been done between the current and suggested techniques. Working with dark skin has resulted in an 89% detection rate, which is an improvement over the current 83.3% detection rate. Because fewer persons with dark complexion were willing to engage in the experiment, data set preparation was essential during that phase of the trial. [29] This thesis catered for this issue and proposes a solution in the cases when people with darker complexion refuses to engage in data set preparation experiments whereby synthetic faces are generated.

Another study was conducted to mitigate bias in face recognition technology using skewness-aware reinforcement learning. [31] According to the study, 2 RestNet-34 models with the guidance of Arcface and Softmax loss on the CASIA-Webface were trained. [31] Set-1 and RFW were used to calculate the intra-class and inter-class angles. The set-l included 500 randomly chosen identities for each race from the BUPT-Globalface dataset, and it brought the concept of adaptive margin to the problem of race balance. In order to minimise the skewness of angles between races and learn balanced performance for different races, ideal margins were adaptively selected for each coloured race, with the Caucasians' margin remaining unaltered. [31] Because existing datasets were not race-aware except RFW[32]. they used their BUFT-Globalface and BUPT-Balancedface datasets to train our models and used RFW to fairly measure performance of different races. The four testing subsets that made up RFW were Caucasian, Asian, Indian, and African. For face verification, each subgroup had roughly 10,000 photos of three thousand people. Furthermore, rather than relying solely on training sets, they also created a validation set to compute intra-class and inter-class distance in RL in order to precisely assess the generalisation capacity. There were 500 identities in each race within the validation set, and there were no subjects that overlapped with the RFW, BUPT-Globalface, or BUPT-Balancedface datasets [32]. Verification performance was measured by accuracy.

They utilized average accuracy of four races as metric to evaluate the total performance of the deep models and standard deviation (STD) and skewed error ratio (SER) were used as the fairness criterion. Several studies [32], [33], and [34] confirmed that, despite balanced training, non-Caucasians continue to perform worse than Caucasians and that faces with coloured skin are intrinsically hard for current algorithms to identify.

To investigate this phenomena further, they introduced noise and blur to the RFW [32] images and saw how this affected the performance of both Africans and Caucasians. Gaussian noise was used to further amplify the noise in the images. Two ResNet-34 models were then trained using the BUFT-Balancedface dataset, with guidance from Softmax and Arcface loss, and tested on the noisy and blurry RFW[32]. They observed that even with balanced training, there was still a performance discrepancy between African Americans and Caucasians. In order to verify the efficacy of their RL- RBN, the researchers trained their algorithms on training sets containing images selected at random from their BUPT-Globalface dataset and assessed the results on RFW. [31] The findings revealed several noteworthy findings: first, they demonstrated that racial bias is present in existing algorithms; for instance, when the racial distribution was 4:2:2:2, the accuracy of Norm-Softmax reached 89.69% on the Caucasian testing subset, but it drops dramatically to 84.17% on the African subset. Secondly, the results quantitatively validated their hypothesis that the accuracy of each race was positively correlated with its number in the training set, for instance, increasing the ratio of Caucasians (from 2/5 to 7/10). Training on BUFT-Globalface, they compared their RL-RBN with Softmax, Cosface and Arcface. [31] Furthermore, they compared their method with manual-margin based RBN(M-RBN). To put it simply, the M-RBN assigned distinct fixed margins to races based on the inverse relationship between the number of samples for each race. Asians' performance in M-RBN was always a hindrance to fairness, even though their approach was more equitable than M-RBN. This was due to the complexity of the racial bias problem, where the number affected accuracy out of balance but was not only a fact. Despite the fact that there were far more Asians in the BUFT-Global dataset than Indians and Africans, the group still required a greater margin since, even with balanced training, Asians are the hardest race to identify. This was the reason why the study proposed a reinforcement learning based race-balance network to alleviate racial bias and to learn more balanced features. It introduced the Markov decision process to adaptively find optimal margins for non-Caucasians. [31] This thesis therefore acknowledges the fact that even with a balanced dataset, some ethnic groups for instance Asians and Africans are harder to be recognised than for instance Caucasians.

Mitigating racial bias in face recognition via gradient attention was another study that was conducted. [35] The study used adversarial learning to improve the consistency of Gradient Attention Map (GAM) among different races. Their method was based on GAM alignment. Firstly, a GAM race classification network (NetworkGAM−CT) was trained to classify the races of GAM. The GAM of every image served as the input for NetworkGAM−CT, while the race associated with each GAM served as the output. Resnet-18 [36] and fully connected layer made up NetworkGAM-CT. The first convolution layer of resnet-18 [12] was modified to have one input channel (the GAM dimension is W × H×1). The conventional cross entropy loss function (Lcls) was used as the loss function of NetworkGAM−CT. Then, they used NetworkGAM−CT as discriminator and face recognition network (NetworkID) as generator for adversarial training. NetworkGAM−CT provided an adversarial loss Ladv for face recognition network (NetworkID). Ladv would make NetworkID generate GAM in which NetworkGAM−CT cannot distinguish race. It was well known that a uniform distribution had the highest entropy and presented the most randomness. If an optimal classifier operating on GAM always produced a posterior probability of 1N for all categories in the racial attribute, it meant that the GAM generated by the NetworkID has a consistent sensitive region of different races. [36] They employed GAM guided sensitive facial region erasure (GAM-SFRE) to enlarge the facial region of those with darker complexion to further enhance fairness. They started by looking through every pixel in the GAM and arranging them in descending order based on how much attention each pixel received. They then chose the mask's centre point from among the pixels with the highest attention value. They employed a rectangular block as their mask, which was smaller in width than a w mask and smaller in height than a H mask. Lastly, they input the image into the network for training after superimposing the mask over the original image at the appropriate location. The researches believed that this inconsistency in confidence was also one of the factors leading to racial bias. To lessen this discrepancy, they consequently incorporated a confidence balance loss based on ArcFace's objective function. The maximum probability value Pmax, also known as confidence, was obtained by feeding the normalised features and weights into the ArcFace (additive angular margin penalty) loss function in order to obtain the probability value Pi for each class.

The number of samples in the current batch with Pmax smaller than the threshold Tconfidence had to be determined next, with nbatch standing for the total number of samples in the batch. Ultimately, they obtained the ultimate objective function in GAM-CT by combining Lconf, LID, and adversarial loss Ladv.

RFW datasets were utilised for testing, and BUPT-Balancedface and BUPT-Globalface datasets for training. Four racial groups' faces made up BUPT-Balancedface: Asian, Indian, African, and Caucasian. 1.3 million photos of 28K celebrities, with roughly 7K identities per race, were included in the dataset. 2 million photos of 38K celebrities were included in BUPT-Globalface, and their racial makeup closely matched that of the global population. Four racial groups' faces made up RFW: Asians, Indians, Africans, and Caucasians. RFW subsets each included roughly 10K photos representing 3K identities. Using the average accuracy of four races, they computed the skewed error ratio (SER) and standard deviation (STD). The criteria for fairness were STD and SER. They preprocessed the photos by cropping and resizing them to 112 × 112 pixels and using five facial landmarks for similarity transformation. They employed the RestNet-34 CNN architecture. Using three GPUs (NVIDIA GEFORCE 1080Ti), the SGD technique was used to train the model. [31] Using the ArcFace loss function, they trained a ResNet-34 model on BUPT-Globalface and then communicated the results via RFW Protocol. In contrast to the SOTA findings, GABN decreased SER to 1.60 and STD to 0.75. They also experimented on BUPT-Balancedface using the same methodology. In contrast to the SOTA findings, GABN decreased STD to 0.56. In contrast to DebFace, another adversarial learning-based approach, theirs produced lower STD and SER as well as increased model accuracy overall. Their technique performed competitively on datasets that were both balanced and unbalanced. Caucasians had a big sensitive zone while persons with darker skin had a small sensitive region in the average GAM of the four races determined by the baseline. While according to this study, mitigating racial bias in face recognition can be done via gradient attention, this thesis attempts to mitigate racial bias by using generative models to generate synthetic faces to further have a more balanced dataset to be trained on which could result into a better accuracy level in a less complicated manner. [31]

Fan-Shaped GAN for racial transformation is another study that was conducted. [37] The researchers adapted the non-saturating loss [38] with R1 regularisation in order to increase the generated samples' realism and make them more similar to actual photos. They generated high-resolution facial images using the non-saturating loss technique combined with R1 regularisation, yielding realistic results. They changed the initial classification loss suggested by StarGAN [39], which trained the discriminator using only real images as input. They employed the racial domain's one-hot label as an attribute and fed both actual and artificially created phoney photos into the discriminator. In addition to being trained to categorise the genuine photos into their respective domains, the discriminator was also trained to recognise that the created picture's attribute labels are all zero. The generator learnt how to confuse the discriminator and made it classify the fake image into the target class. They also forced the generator to recreate the original picture x using the translated image G(x, t) and the original domain label s by applying a cycle consistency loss [39] on it. As the reconstruction loss, they went with the L1 norm. The researchers were unsure if the created image and the original image should be consistent in the feature space, like previous face editing tasks like cosmetics transformation and face ageing, because their aim was to create different ethnic domain images. To reduce the L1 distance in various feature space layers, perceptual loss was applied to both the original and the rebuilt images. Besides, they also employed a VGG-16 to extract feature, which is pre-trained on ImageNet. They chose the pool1, pool2, and pool 3 output as feature maps, same like in [40]. However, the outcome of the former's ethnic change was superior to the latter. They combined the encoder, which is the front end of the multi-domain translation generator. However, they retained the various deconvolutional components at the generator's end, referred to as the decoder. They also corrected various ethnic domain data in the decoders, with each decoder matching to produce a distinct domain's face image. In the encoder, they used batch normalisation, while in the decoder, they used instance normalisation. They used PatchGANS [41] to assess the legitimacy of the discriminator's output, which determined if the overlapping images that were patched were authentic or fraudulent. They employed a 34 × 34 patch for training images with a resolution of 128 × 128. After using the RFW dataset, they separated all of the photos into four groups: Caucasian, Asian, Indian, and African. They created four domains, took 5,000 photos from each ethnic group as training data, and used similarity transformation to align the original 400 × 400 size to 128 × 128. 500 photos from each ethnic group were then chosen at random for the test set. They chose the two training domains in order, which required updating the two decoders at the conclusion of the generator each time, to improve the balance of the transition between the different races.

During training, learning rate maintained constant in the first 24 epochs and declined to 0 linearly in the next 48 epochs.

Africans had a poorer recognition rate than other races, which was noted in [42] and demonstrated in the real-world application of deep facial recognition models, thus they decided to employ them as their data augmentation target.

To train their FGAN model, they reprocessed all of the African and Caucasian images and aligned them to 112 x 112. A race has roughly 300k photos. In a similar manner, 1,000 photos were chosen at random to serve as the test set, and the other images served as the training set. They employed the Caucasian subset described in [42] as a training set for the deep facial recognition model, which included roughly 500K labelled photos of 10,000 Caucasians. Arcface loss was used to guide the training of the ResNet–34 in all experiments. They established 200, 0.9, and 0.0005 for the batch size, momentum, and weight decay, respectively. When errors reached a plateau, the learning rate was reduced twice by a factor of ten. It was initially set at 0.1. With the aforementioned settings, the baseline model was only trained on 10,000 Caucasians. Simultaneously, they extracted an additional 10,000 African photos from the Caucasian training set by applying the learned FGAN model to all the images in the training set. Augmented images were added to the original training set as new ids, and the training details were same as baseline. [37] Using the RFW evaluation set and other widely used face recognition evaluation datasets, they compared the performance of the baseline and enhanced models. The RFW assessment set included 6,000 pairs of photos for each race. The images were chosen based on cosine similarity to prevent saturated performance and to challenge the recognizer with variants of the same persons and people with similar appearances. They decided to assess the model's performance using LFW, CFP-FP, and AgeDB-30 as additional common assessment sets. According to the findings, the supplemented model's recognition rate for African people increased by roughly 1%, while its performance on other racial groups and standard assessment sets did not noticeably decline. [37] To further conceal racial information from the original photograph, they attempted two different approaches to create ethnically independent facial images. One involved taking a pixel-by-pixel average of the four racial images that FGAN directly produced. Another strategy was adversarial learning; they constructed a racial independent generator with a basic convolutional layer after training a race classifier based on ResNet18 on the RFW training set. The four racial images produced by FGAN were fed into the generator, which used adversarial learning with a trained ethnic classifier to attempt to produce realistic images while giving the classifier the same score for each of the four races. In terms of visual effects, both approaches could conceal some features of the original ethnic domain because they were comparable to the original photographs, even if the images obtained by the two methods couldn't fool the pre-trained racial classifier. Their research yielded a novel approach to minimising racial bias in deep facial recognition and balancing the recognition rate of different races. Furthermore, they demonstrated how the average of different races might be used to create an ethnically independent facial image. Their technique also managed to preserve the model recognition effect while partially masking the racial information included in the original photographs. [37]

Moreover, another study which involved measuring hidden bias within face recognition via racial phenotypes was conducted. [43] The goal of the project was to develop a novel approach for analysing racial bias in face recognition by using facial phenotypic features. In order to investigate racial bias in face recognition, the suggested approach added racial phenotypic traits in place of perhaps protected or ill-defined subject attributes. They avoided unintentionally offending anyone by designating race-related phenotypic traits with descriptive names, as per the research of [44], [45]. Two distinct face datasets that were openly accessible were used for their research (VGGFace 2 and RFW). They took two restrictions into consideration when adopting graphs and measures for facial recognition. First, tight cropped, low-quality photos with occlusion, shadows, and illumination fluctuations were needed for both the training and test stages in order to evaluate face recognition tasks efficiently. [43] Compared to real-world human faces, this made phenotype attribute recognition on the unique features of face dataset photos more challenging. Second, the wider classification led to a greater quantity of possible clusters, rendering bias assessment ineffective for facial recognition systems. They therefore chose to employ the following six key attributes—skin type, eyelid type, nose shape, lip shape, hair type, and hair color—to determine the phenotypic categories for their study. Under the six core attributes, they subsequently had twenty-one distinct attribute categories. They selected to employ Fitzpatrick Skin Types for skin tones as it provided greater granularity, {Type 1, Type 2, Type 3, Type 4, Type 5, Type 6}, than binary skin-tone groupings, {lighter skin-tone, darker skin-tone}. [43] Several cosmetic industry rules have categorised the look of the human eye according to its location, shape, and settings[46].

They did not, however, have a strong connection to race or a scientific background. As a more distinguishing characteristic for racial bias, the researchers instead examined epicanthal folds and examined eyelid difference [47]. They admitted that one attribute category might be seen in several racial groups. But finding the easiest-to-observe racial phenotypic traits on photos was their primary priority to assess bias.

They examined the nasal breadth to create two categories for the nose's appearance: wide and narrow [48]. By measuring the frequency of twists, waves, and curve diameter, [49] classified hair texture into eight groups. Here they utilised eight categories and grouped them into three main hair texture types: straight, wavy, curly, in addition to bald.[43] Since hair colour correlated with skin tone, they kept their hair colour even though it was the most artificially manipulable characteristic. [50] The classifications they employed for hair colour were: red, grey, black, blonde, brown. They selected the most well-known face recognition datasets ahead of the annotation procedure in order to verify our suggested approach. They selected the RFW dataset for the face verification job because it offered a comparatively wider range of racial subjects, with three to five photos per subject. They employed the VGGFace2 closed-test set, which included at least 300 photos per individual, for face identification. They created an annotation interface for each dataset in order to make the annotation process reliable and easy to use. In order to prevent improper annotation caused by challenging samples such greyscale photos, facial makeup, and dim scene lighting, they supplied many sample images of a subject. Within the annotation interface, each subject was shown a series of face photos along with attribute category selectors. [43] Each subject was then annotated via the interface by a seasoned annotator with knowledge of morphological variations between races. They acquired 11654 subjects annotations from the RFW and VGGFace2 benchmark datasets. They found that Skin Type 3, Straight Hair, Narrow Nose, Other (non-monolid) Eyes, and Small Lips were the most common phenotype attribute categories for both datasets, and that these categories were connected with the predominance of Caucasian appearances. They used the VGGFace2 benchmark datasets, which included 8631 subjects with a racially unequal subject distribution, to train ArcFace using a ResNet 100. Their specific selection of VGGFace2 resulted from their investigation into the effects on their suggested evaluation technique of imbalanced training data, including data bias. [43] They made use of a ResNet34 backbone architecture with the Softmax loss, which has been trained on the 28,000 face subjects in the BUPT-Balanced benchmark dataset. With 7000 face subjects each, the four groups in the BUPT-Balanced were distributed racially evenly: African, Asian, Indian, and Caucasian. This was primarily done to evaluate, using their suggested phenotype-based methodology, the effects of a racially balanced training dataset on findings over the bias. In contrast to the imbalanced training data, they compared the extent to which a racially balanced training dataset reduced that performance gap. The impact of a single attribute (attribute-based) and appearance-based facial groups (subgroup-based) on the evaluation performance of face verification was then investigated using two pairing procedures. [43] They started by creating pairings from pictures that belonged to the same attribute category. As a result, they evaluated the performance of each individual attribute for face verification using both training sets. They randomly chose 20k positive and 20k negative pairs for attribute-based face verification out of all potential pairs for each attribute. To determine which pairs were the most difficult, they computed the cosine similarity of feature encoding for each chosen pair of positive and negative features. Subsequently, they picked the most similar 3000 couples from the negative samples and the least similar 3000 pairs from the positive samples for each attribute category. To demonstrate how much the standard deviation (σ) varied between balanced and imbalanced training data, they employed both training settings. They discovered that the accuracy were worse for monolid eyes, black hair, large lips, and wide nose in setup 1 (imbalanced training data) and setup 2 (racially balanced training data) than for the other eye, blonde hair, tiny lips, and narrow nose, respectively. [43] They also detected a minor link between darker skin tones and higher false matching rates when they matched from the same attribute categories. Moreover, despite the imbalanced training configuration resulted a higher performance difference (σ), the amount of difference between two setups was tiny, showing that a racially balanced dataset distribution was not enough to overcome performance bias. [43] They then divided the dataset into a number of subgroups based on distinct combinations of phenotypic attributes. One such subgroup, for instance, was made up of individuals with skin type 3, thin lips, straight hair, large noses, and monolid eyes. [43] The primary goal of these pairings was to illustrate the consequences of individual attribute changes on a group. In order to study subgroup-based performances, they also created every conceivable subgroup with a variety of phenotypic attribute category combinations.

They deduced from their observations that groups with a wide nose, full lips, and monolid eye type consistently performed less accurately than the other groups with a narrow nose, tiny lips, and other eye. Additionally, they found that there were significantly less subgroup variations among darker skin tones than among lighter tones, which led to a variety of evaluation and analytic issues. [43] During the test phase, it was not sufficiently interpreted; there were minority with monolid eyes and dark skin throughout the world's population, as well as other less prevalent variations.

[43] There were insufficient representatives of these minority groups in benchmark datasets. A more comprehensive evaluation dataset that encompassed a wider range of phenotypic combinations, ensuring that its distribution accurately reflected the global population, would have been ideal. Finally, they used training setup 2 to estimate these differences between various grouping algorithms. Since these grouping methodologies were frequently used in the literature, they adopted the racial groupings {African, Asian, Indian, Caucasian} and the binary skin tone groupings {lighter skin-tone, darker skin-tone}. They contrasted them using their grouping technique based on phenotype. They demonstrated how three distinct techniques' sub-groups differed in accuracy and standard deviation. Greater variety exposed implicit prejudice that could have gone unnoticed under restrictive, inaccurate racial or binary skin tone classification schemes. The phenotype-based grouping approach resulted in a more precise measurement of performance bias since it allowed for a more detailed observation of performance variability (greater standard deviation). [43]

➢ *The Rest of this Thesis is Organised as Follows:*

- Section 3 – Data Collection & Methodology
- Section 4 – Results & Analysis
- Section 5 – Conclusion, followed by Statement of limitations and suggestions for future works.

CHAPTER THREE
METHODOLOGY

This study aims to have generative models as standard practice for dataset diversification. The goal is to influence the industry towards more ethical and equitable technology development. This thesis seeks to bridge the current gap in facial recognition technology and ultimately aid in the development of more fair, accurate, and socially conscious technology by addressing the significant issue of racial bias.

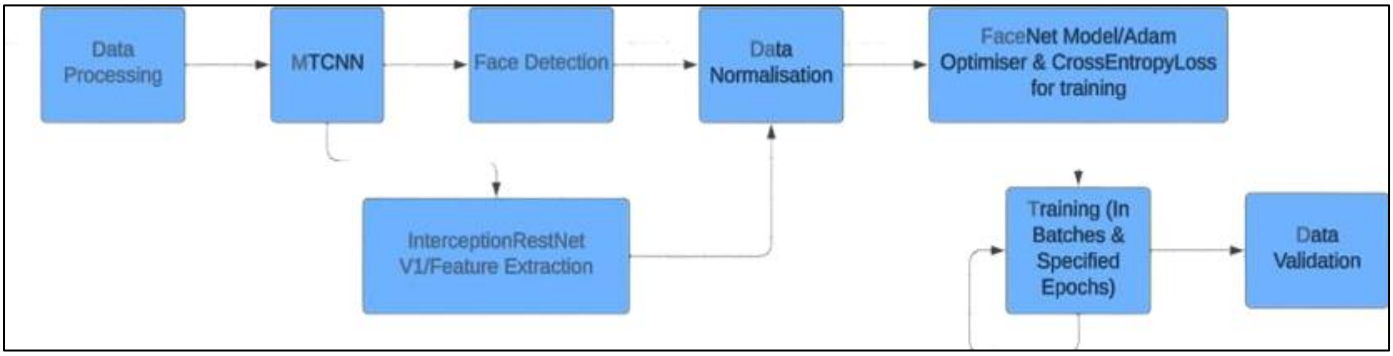

Data Processing
MTCNN
Face Detection
Data Normalisation
FaceNet Model/Adam Optimiser & CrossEntropyLoss for training
InterceptionRestNet V1/Feature Extraction
Training (In Batches & Specified Epochs)
Data Validation

Fig 2 Methodology and Workflow Development for Our Recognition Model

This project fine-tunes the FaceNet model and using the InceptionRestnetV1 on a custom dataset whereby one has real and generated images together and the other having only real images. The CelebA Dataset was used to generate realistic synthetic images. Once that is done, the race labels in the datasets are encoded using 'LabelEncoder' to convert categorical labels into numerical format. The images are loaded and pre-processed using MTCNN module for face detection and InceptionrestNetV1 for feature extraction. The 'Load_and_preprocess_images' function takes care of loading and preprocessing images in batches. The data are then organised and then transformed whereby resizing and normalisations are applied. Pixel values are normalized to the range [0,1] for both training and validation image arrays.

The FaceNet model is fine-tuned on the dataset and an Adam optimizer and CrossEntropyLoss/SoftMax are used for optimizing and as loss function. The model is trained for 10 epochs. The training loop includes loading batches of data, performing forward and backward passes, and updating the model parameters. After each epoch, the model is evaluated on the validation set. For each race category, the accuracy, standard deviation(STD) and Skewed error ratio(SER) are calculated based on the entire validation set.

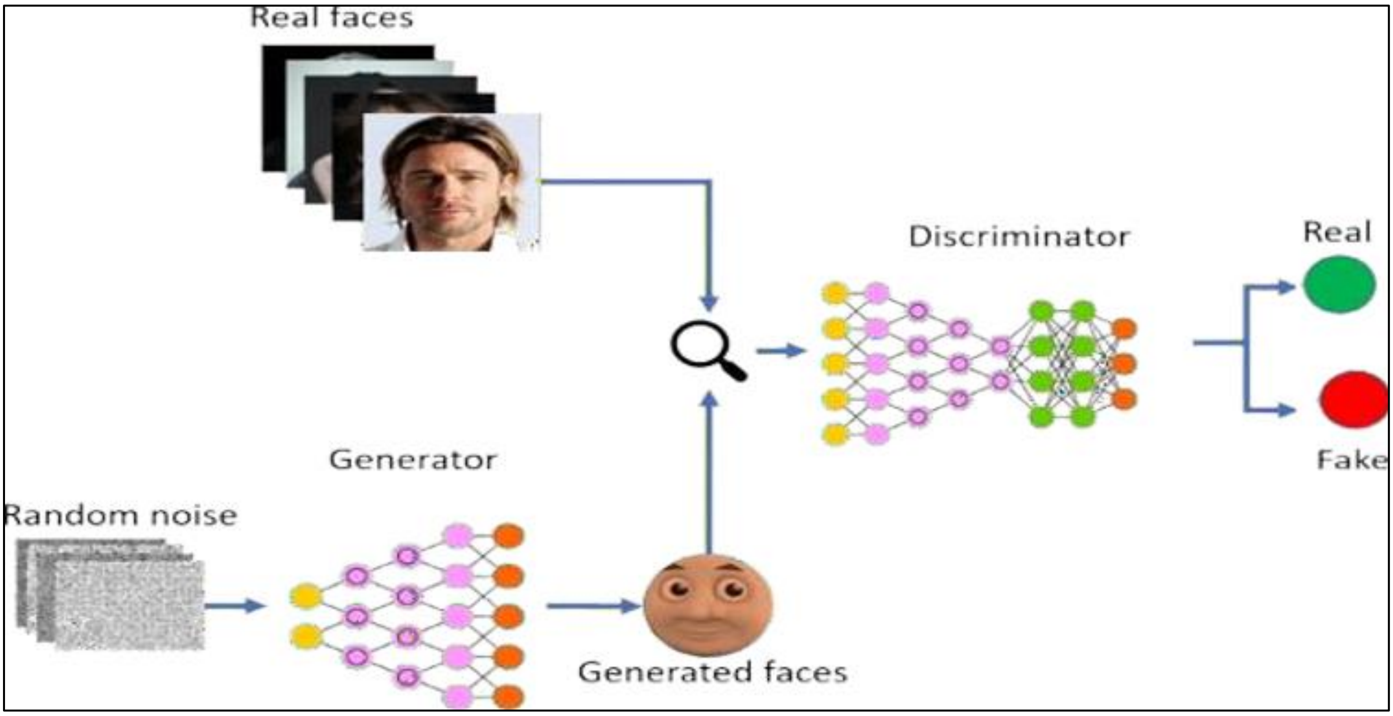


Fig 3 GAN Model

By putting two neural networks against one another, GANs can determine the probability distribution of a dataset. While the other neural network, known as the discriminator, creates new instances of data, the generator creates new examples of data, and the discriminator determines whether each instance of data it analyses is authentic. In the meantime, the discriminator receives fresh, artificial, or phoney images from the generator. It does this in the hopes that, although being fraudulent, they too will be accepted as genuine. By employing transposed convolution, the inverse of convolution, a 100-dimensional noise (uniform distribution between -1.0 and 1.0) is created, which is used to create the fake image. The generator's objective is to provide acceptable visuals, allowing users to lie covertly. The discriminator's task is to recognise images that originate from the generator as fraudulent. The steps a GAN takes are as follows: 1) After receiving random numbers, the generator outputs an image. 2) The discriminator receives this created image in addition to a stream of photos obtained from the real, ground-truth dataset. 3) The discriminator receives inputs in the form of actual and phoney images, and outputs probabilities—a value between 0 and 1 that indicates phoney and 1 that indicates a prediction of authenticity. The discriminator and the known ground truth of the images are in a feedback loop. The discriminator and generator are in a feedback loop.

All the analysis and computation necessary for this project such generating the synthetic images, the analytics of distribution, data visualisations, development and evaluation of the model performances were conducted on a simple machine with Processor: AMD Ryzen 7 5800H 3.20 GHz and 16 GB Ram. The chosen programming language was Python.

- *Dataset Collection*

Two datasets were collected; one being the FairFace dataset with has a diverse dataset over 100k images of both males and females and the other being the CelebA dataset which we used our GAN model to generate our synthetic images. To test our recognition system, we also took random pictures from the Chicago Dataset. [51]

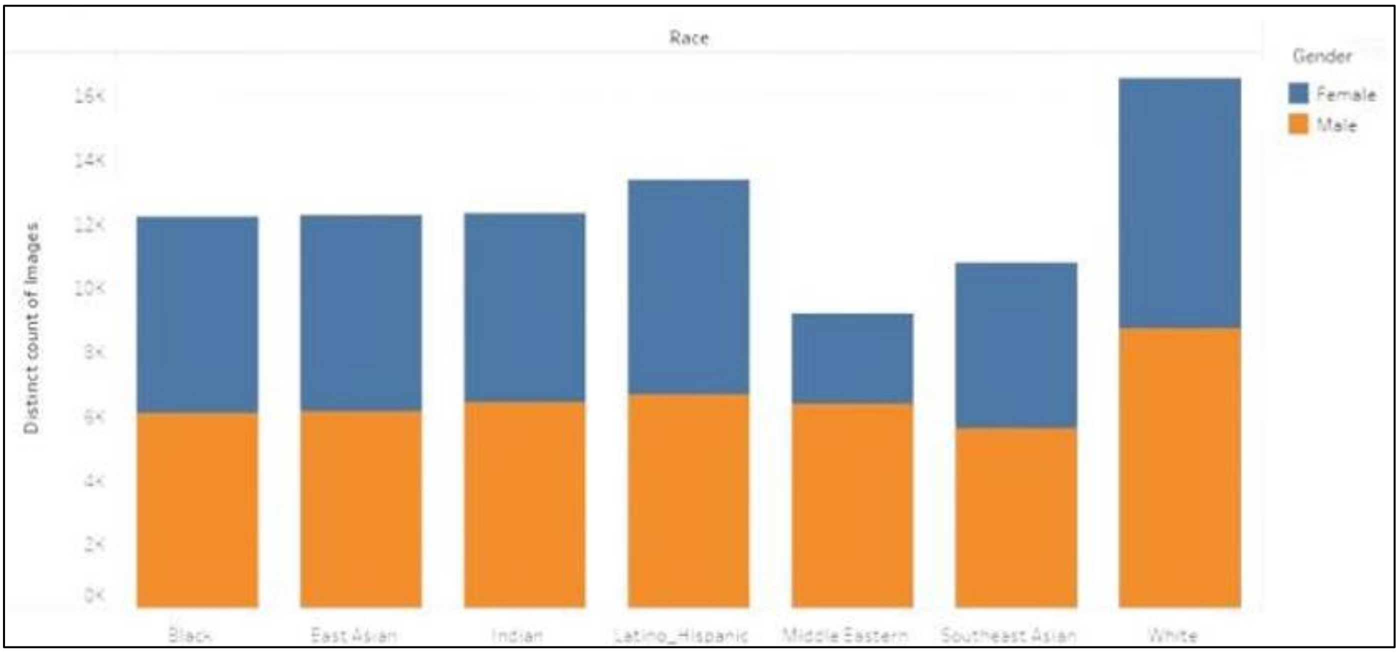


Fig 4 FairFace Dataset [52]

The Fairface dataset is a high-quality dataset of human faces designed to address issues of bias and fairness in facial recognition systems.[52] The FairFace dataset's main objective is to lessen prejudice and advance fairness in facial recognition software. Its main goal is to give a fair representation of faces from various racial, gender, and age categories. FairFace includes images of faces from seven major racial or ethnic groups: African, East Asian, Indian, Latino_Hispanic, Middle Eastern, Southeast Asian, and Caucasian. This diversity is essential for training models that perform well across various demographic groups. Apart from diversity of race, FairFace also tackles representation of gender and age. It adds to a completer and more representative sample by containing faces from diverse age groups and genders. Social media sites, the internet, and publicly accessible databases are the sources of the photos used in FairFace. This varied collection of sources contributes to the dataset's ability to depict a broad variety of real-world situations. Researchers may assess and train models for demographic traits thanks to the dataset's annotations for gender, age, and race. These annotations are useful for comparing facial recognition algorithms that take fairness into account. The FairFace dataset contains a mixture of high-resolution and low-resolution photos, with varying resolutions. This variance reflects the range of image quality that one could potentially come across in practical situations. FairFace recognises the difficulties of building an impartial and fair dataset. Unbalances in training data can introduce bias into facial recognition algorithms; FairFace attempts to address this issue by providing a more balanced face representation. The dataset can be used to train and assess fairness-focused facial recognition methods. It is used by researchers to evaluate how well algorithms work for various demographic groups and to find possible bias sources.

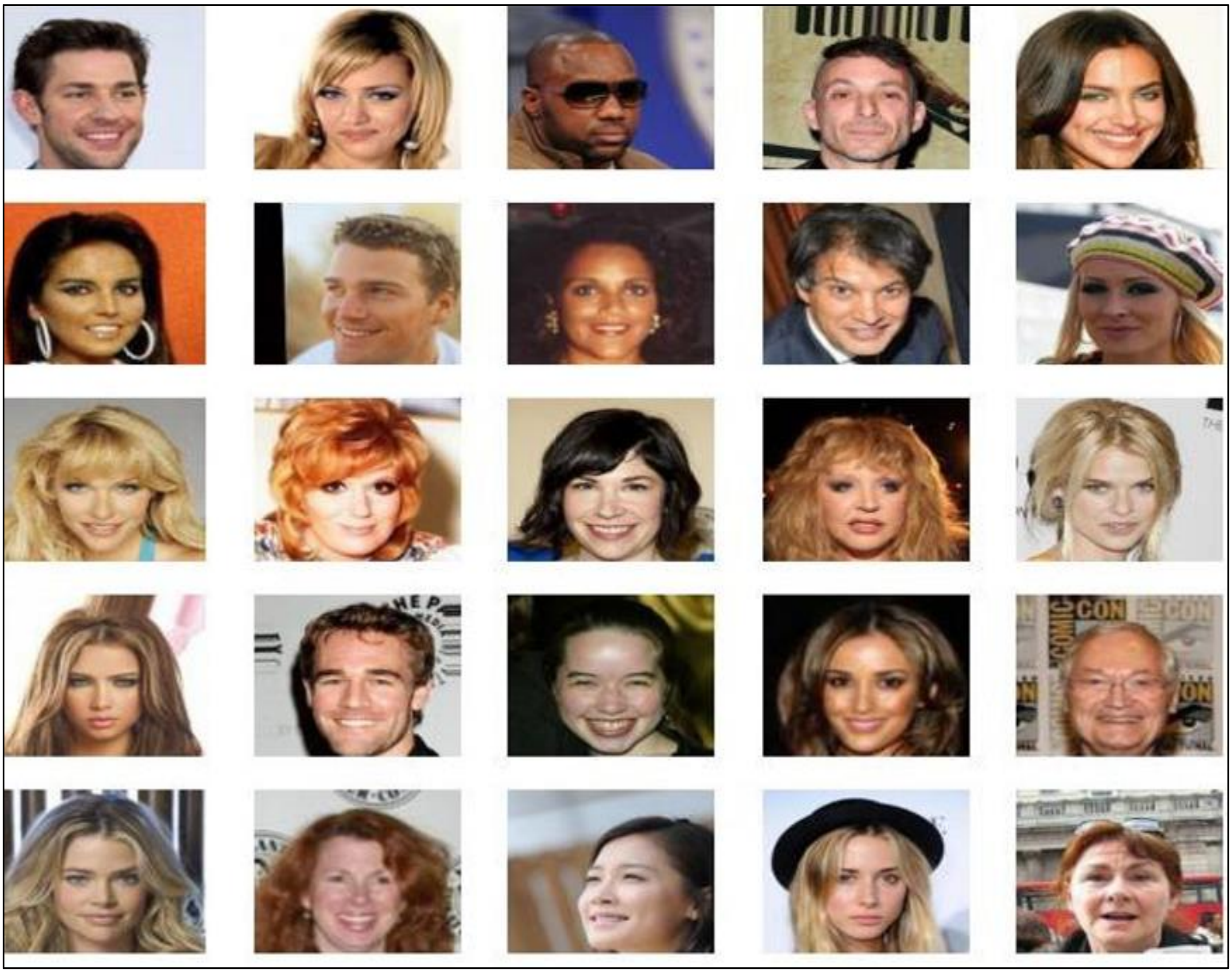
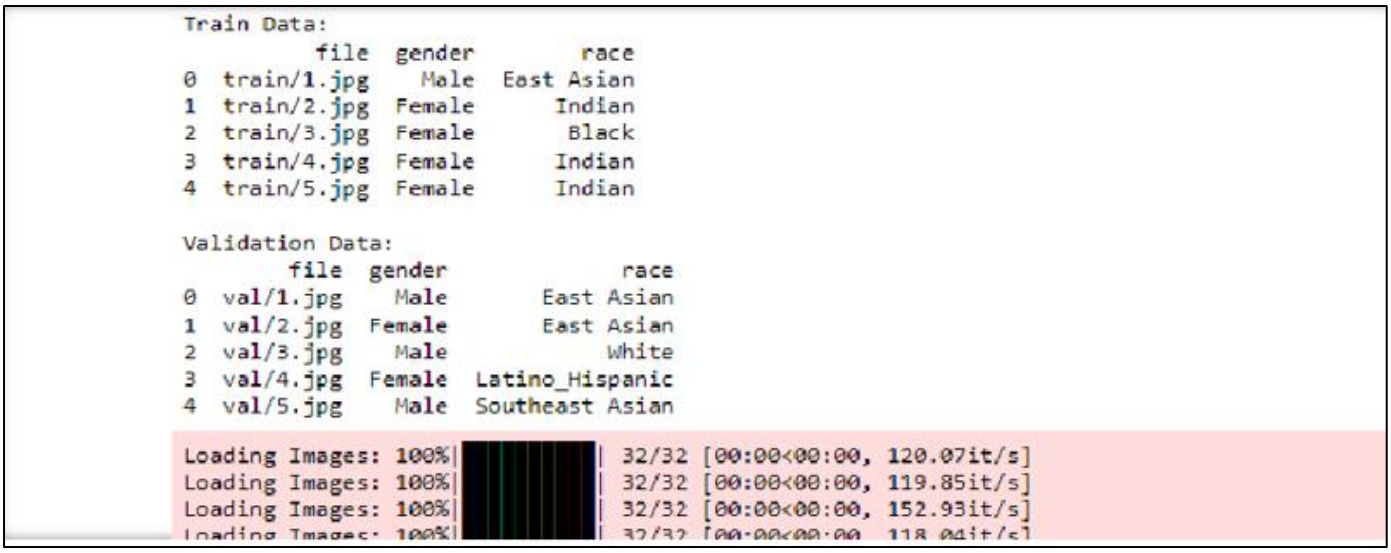


Fig 5 CelebA Dataset

The CelebA dataset contains over 200,000 celebrity images, making it one of the largest publicly available datasets for face-related tasks. The dataset features images of approximately 10,000 unique celebrity identities. Each identity is associated with multiple images, capturing different facial expressions, poses, and lighting conditions. The dataset includes a diverse set of celebrities, encompassing various ethnicities, ages, and genders. This diversity is essential for training models that generalize well across different demographic groups. CelebA images exhibit a wide range of variability in terms of facial expressions, head poses, occlusions, and backgrounds. This variability is crucial for training models that can handle real-world scenarios. The dataset has a resolution of 178 x 218 pixels. While not extremely high-resolution by modern standards, it is sufficient for many facial recognition and attribute prediction tasks. Due to its large size, detailed annotations, and diversity, the CelebA dataset has become a popular choice for researchers and developers working on facial analysis tasks. As such, this was the chosen dataset that was used for our GAN model to train and generate synthetic realistic human images.

- *Pre-Processing Data Images*

Two csv files are loaded using pandas which contains information about the training and validation data, including the file paths, gender, and race labels. The structure of the loaded data is then displayed as shown in Figure 5. 'LabelEncoder' from scikit-learn is used to encode the categorical race labels into numerical values. This is done for both the training and the validation sets. The original race labels are replaced with the encoded values. The 'Load_images' function is defined to read and convert images into NumPy arrays. This function is used to load images in batches for both training and validation sets. The images are loaded in batches using batch size 32 and the pixel values are normalized to the range [0,1] by dividing by 255.0. The resulting image arrays 'X_train' and 'X_val' are created by extending the batches.



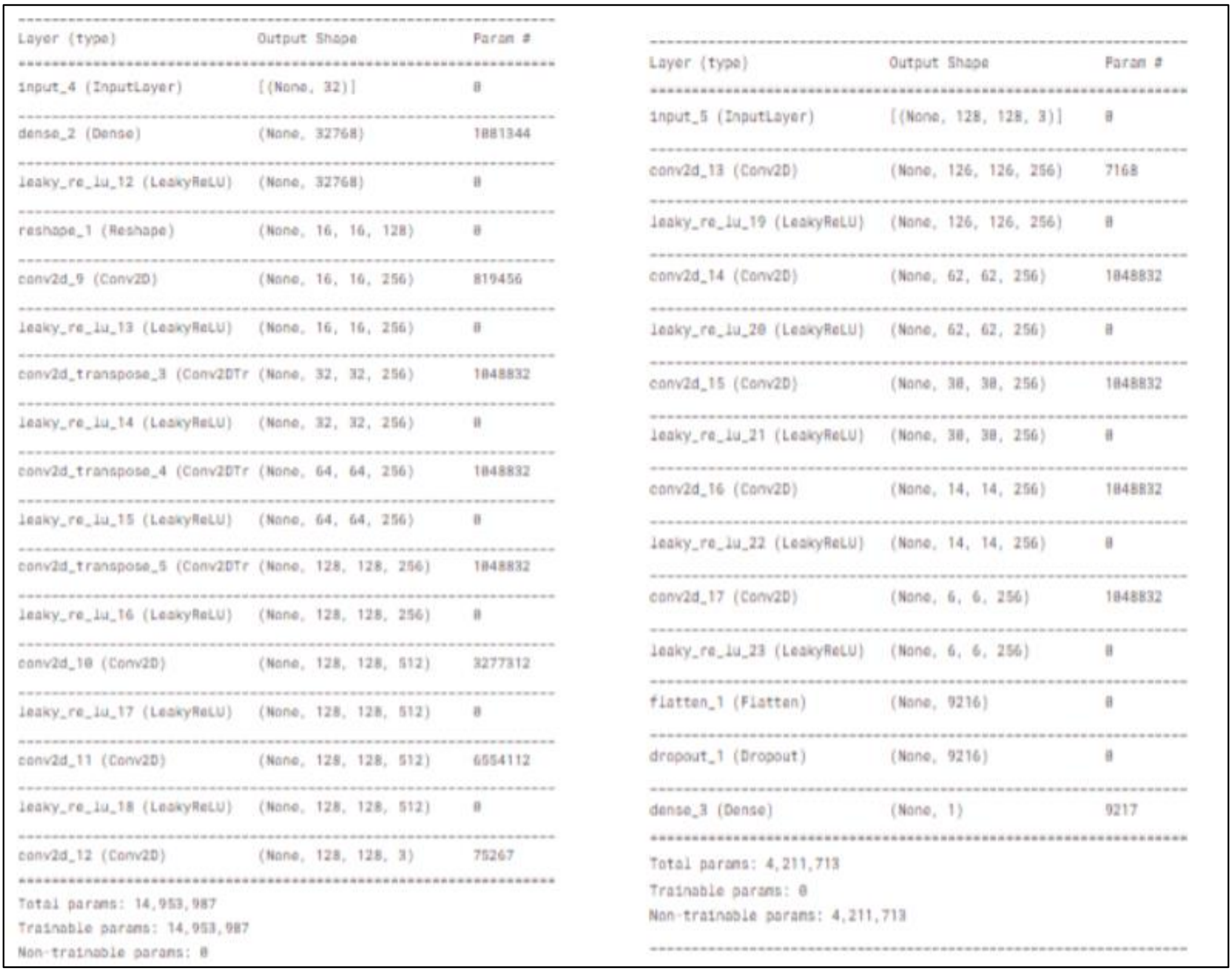
```
Train Data:
          file  gender        race
0  train/1.jpg    Male  East Asian
1  train/2.jpg  Female      Indian
2  train/3.jpg  Female       Black
3  train/4.jpg  Female      Indian
4  train/5.jpg  Female      Indian

Validation Data:
        file  gender             race
0  val/1.jpg    Male       East Asian
1  val/2.jpg  Female       East Asian
2  val/3.jpg    Male            White
3  val/4.jpg  Female  Latino_Hispanic
4  val/5.jpg    Male  Southeast Asian

Loading Images: 100%|██████████| 32/32 [00:00<00:00, 120.07it/s]
Loading Images: 100%|██████████| 32/32 [00:00<00:00, 119.85it/s]
Loading Images: 100%|██████████| 32/32 [00:00<00:00, 152.93it/s]
Loading Images: 100%|██████████| 32/32 [00:00<00:00, 118.04it/s]
```

Fig 6 Image Loading & Data Preprocessing

- *A Model Development – GAN Model*

The CelebA data is loaded in pre-processed. It was resized and cropped down to 128x128 pixels. Next step was to create a generator. The generator takes the opposite tack, claiming that the artist is the one attempting to trick the discriminator. There are eight convolutional layers in this network. Here, we begin by feeding our input—called gen_input—into the first convolutional layer. After performing a convolution, each convolutional layer also executes a leaky ReLu and batch normalisation. The tanh activation function is then returned. Next, we create a discriminator. Similar to the generator, the discriminator network is made up of convolutional layers. We will first execute convolution for each layer of the network, followed by batch normalisation to increase the network's speed and accuracy, and lastly, a leaky ReLu on each layer. Subsequently, a GAN model that integrates the discriminator and generator models into a single, larger model can be defined. The discriminator model's output and error will be utilised to train the generator's model weights utilising this larger model. To make sure that only the weights of the generator model are changed, the discriminator model's weights are marked as not trainable in this larger GAN model because the discriminator model is trained independently. This modification to the discriminator weights' trainability solely impacts the combined GAN model during training; it has no effect on the discriminator while training alone. This larger GAN model accepts a point in the latent space as input.

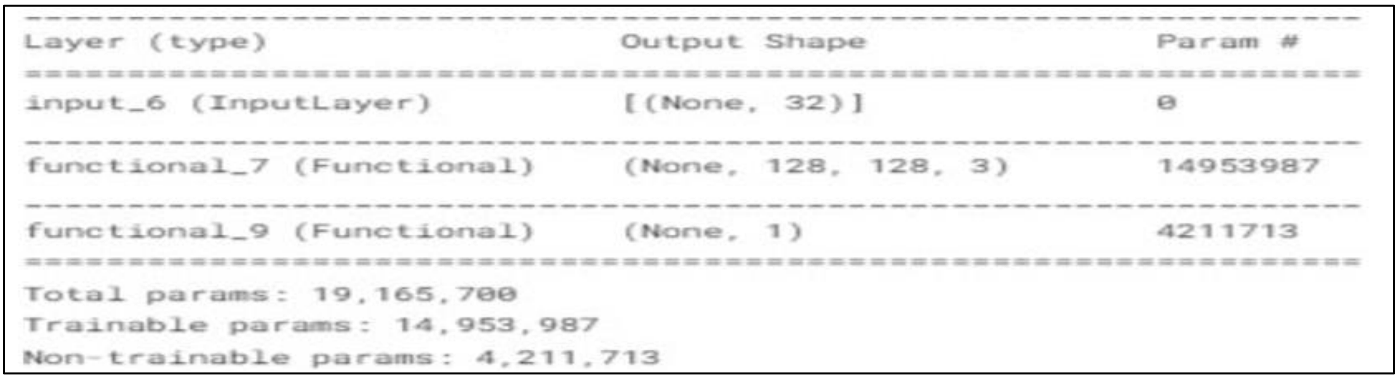
It then creates an image using the generator model, feeds it into the discriminator model, and outputs or classifies it as real or false. Since the Discriminator yields sigmoid output, we apply binary cross-entropy to the loss function. In this instance, RMSProp as an optimizer produces more lifelike phoney images than Adam. There is 0.0001 learning rate. In the later stages of the training, learning is stabilised by weight decay and clip value. The goal of GANs is to mimic a probability distribution. Consequently, we ought to use loss functions that capture the disparity between the true data distribution and the data distribution produced by the GAN.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_4 (InputLayer)         [(None, 32)]              0
_________________________________________________________________
dense_2 (Dense)              (None, 32768)             1081344
_________________________________________________________________
leaky_re_lu_12 (LeakyReLU)   (None, 32768)             0
_________________________________________________________________
reshape_1 (Reshape)          (None, 16, 16, 128)       0
_________________________________________________________________
conv2d_9 (Conv2D)            (None, 16, 16, 256)       819456
_________________________________________________________________
leaky_re_lu_13 (LeakyReLU)   (None, 16, 16, 256)       0
_________________________________________________________________
conv2d_transpose_3 (Conv2DTr (None, 32, 32, 256)       1048832
_________________________________________________________________
leaky_re_lu_14 (LeakyReLU)   (None, 32, 32, 256)       0
_________________________________________________________________
conv2d_transpose_4 (Conv2DTr (None, 64, 64, 256)       1048832
_________________________________________________________________
leaky_re_lu_15 (LeakyReLU)   (None, 64, 64, 256)       0
_________________________________________________________________
conv2d_transpose_5 (Conv2DTr (None, 128, 128, 256)     1048832
_________________________________________________________________
leaky_re_lu_16 (LeakyReLU)   (None, 128, 128, 256)     0
_________________________________________________________________
conv2d_10 (Conv2D)           (None, 128, 128, 512)     3277312
_________________________________________________________________
leaky_re_lu_17 (LeakyReLU)   (None, 128, 128, 512)     0
_________________________________________________________________
conv2d_11 (Conv2D)           (None, 128, 128, 512)     6554112
_________________________________________________________________
leaky_re_lu_18 (LeakyReLU)   (None, 128, 128, 512)     0
_________________________________________________________________
conv2d_12 (Conv2D)           (None, 128, 128, 3)       75267
=================================================================
Total params: 14,953,987
Trainable params: 14,953,987
Non-trainable params: 0
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_5 (InputLayer)         [(None, 128, 128, 3)]     0
_________________________________________________________________
conv2d_13 (Conv2D)           (None, 126, 126, 256)     7168
_________________________________________________________________
leaky_re_lu_19 (LeakyReLU)   (None, 126, 126, 256)     0
_________________________________________________________________
conv2d_14 (Conv2D)           (None, 62, 62, 256)       1048832
_________________________________________________________________
leaky_re_lu_20 (LeakyReLU)   (None, 62, 62, 256)       0
_________________________________________________________________
conv2d_15 (Conv2D)           (None, 30, 30, 256)       1048832
_________________________________________________________________
leaky_re_lu_21 (LeakyReLU)   (None, 30, 30, 256)       0
_________________________________________________________________
conv2d_16 (Conv2D)           (None, 14, 14, 256)       1048832
_________________________________________________________________
leaky_re_lu_22 (LeakyReLU)   (None, 14, 14, 256)       0
_________________________________________________________________
conv2d_17 (Conv2D)           (None, 6, 6, 256)         1048832
_________________________________________________________________
leaky_re_lu_23 (LeakyReLU)   (None, 6, 6, 256)         0
_________________________________________________________________
flatten_1 (Flatten)          (None, 9216)              0
_________________________________________________________________
dropout_1 (Dropout)          (None, 9216)              0
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 9217
=================================================================
Total params: 4,211,713
Trainable params: 0
Non-trainable params: 4,211,713
_________________________________________________________________
```

Fig 7 The GAN Architecture



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_6 (InputLayer)         [(None, 32)]              0
_________________________________________________________________
functional_7 (Functional)    (None, 128, 128, 3)       14953987
_________________________________________________________________
functional_9 (Functional)    (None, 1)                 4211713
=================================================================
Total params: 19,165,700
Trainable params: 14,953,987
Non-trainable params: 4,211,713
```

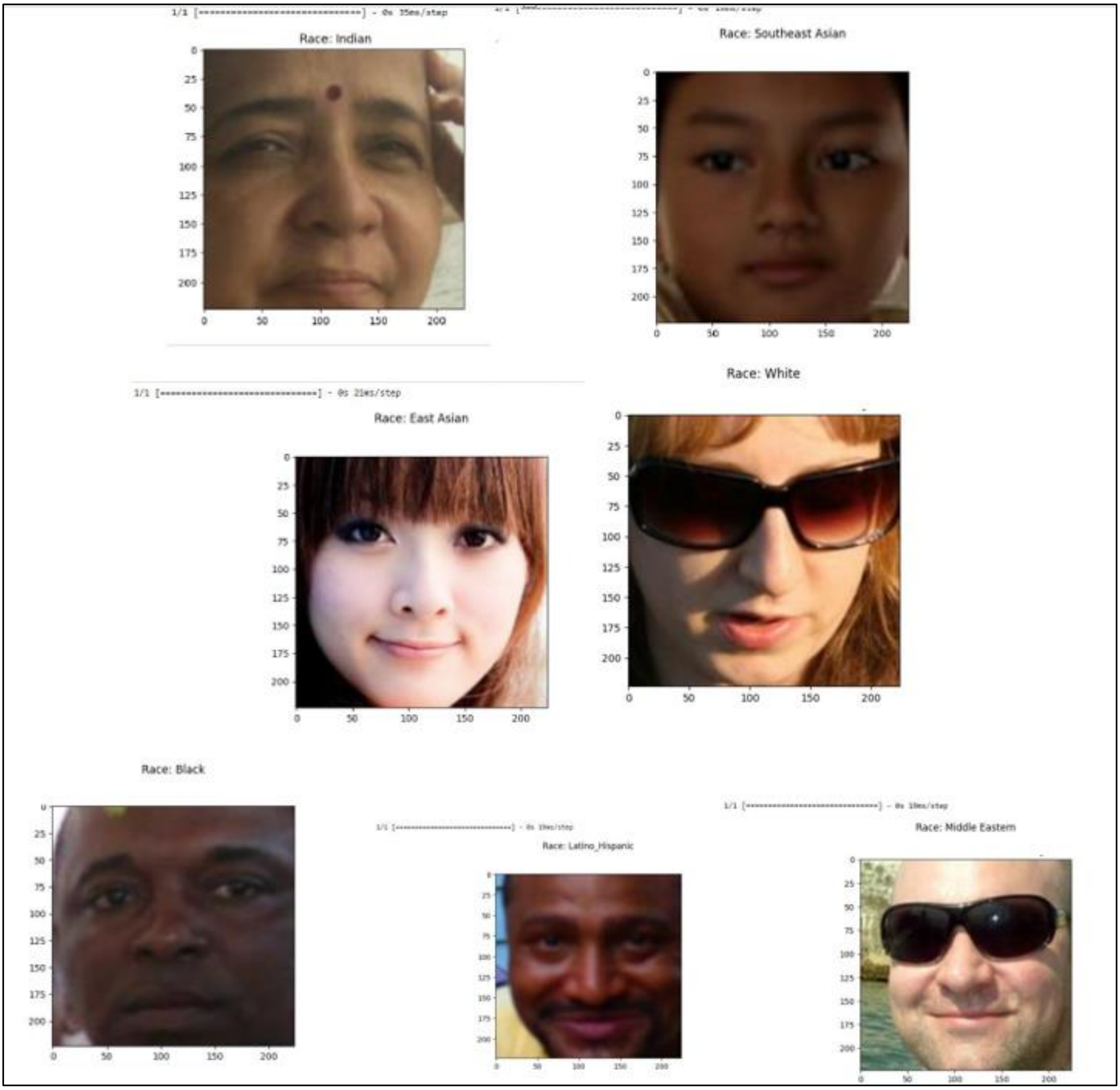
Fig 8 The GAN Summary

Next, we must train our GAN model which is tricky as GANS contain two separately trained network. The generator and discriminator are the two types of training that GANs must balance. GAN convergence is difficult to detect. The performance of the discriminator decreases as the generator grows better with training because it becomes more difficult for the discriminator to distinguish between real and false. The discriminator has a 50% accuracy rate if the generator operates flawlessly. To make a forecast, the discriminator essentially flips a coin.

The discriminator feedback becomes less significant over time, which is a concern for the GAN's overall convergence. The generator starts to train on garbage feedback and its quality may collapse if the GAN keeps training after the discriminator starts providing entirely random feedback.

- *B Model Development – Recognition Model*

Our model takes face images as input. These images are pre-processed to ensure consistent lightning, alignment and other factors. This has been done in our pre-processed stage by using MTCNN. It detects faces in the image and provides bounding boxes. We then construct our model using Keras with three convolutional layers, max-pooling layers and a flatten layer, and two fully connected layers. The output layer also includes softmax activation. The model is compiled using Adam optimizer and sparse crossentropy loss. The model is trained on the training data with 10 epochs. The evaluation metrics used were the accuracy level, STD (Standard Deviation) and SER (Skewness Errored Ratio). They were used to assess the performance of the model on different racial categories. Low, or small, standard deviation indicates data are clustered tightly around the mean, and high or large, standard deviation indicates data are more spread out. Skewed distribution is important as linear models assume that the distribution of the independent variable and the target variable are very similar. Knowing about the skewness of data could help develop more accurate models. Hence, these are the reasons why they were the chosen metrics for our model, alongside accuracy level. Once the model was trained and validated, we tested it on our Chicago Dataset as it has a diverse set of images of several races. We passed through selected images from the dataset and measured the recognition accuracy it was giving for each race. The results will be mentioned and discussed in the 4th chapter of this thesis.



```
1/1 [==============================] - 0s 35ms/step
```

```
1/1 [==============================] - 0s 21ms/step
```

```
1/1 [==============================] - 0s 19ms/step
```

```
1/1 [==============================] - 0s 19ms/step
```

Fig 9 Our Model Recognition of Different Races Accordingly After Adding Generated Dataset

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb145

➢ *Training & Evaluation of Model*

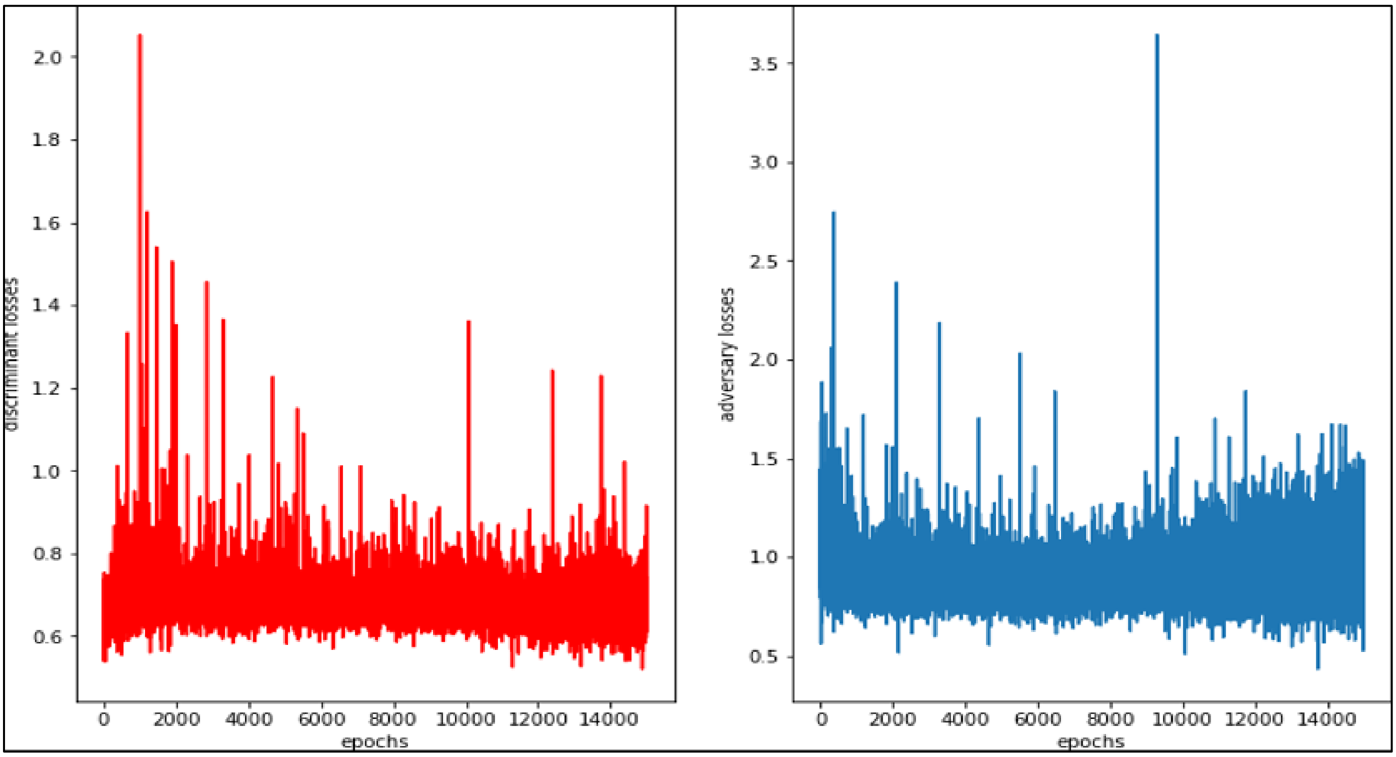
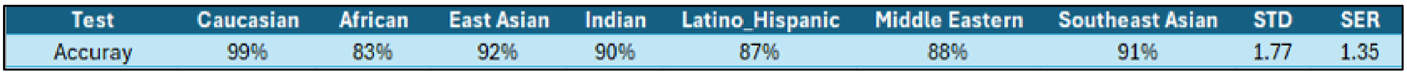


Fig 10 Training Summary of GAN Model

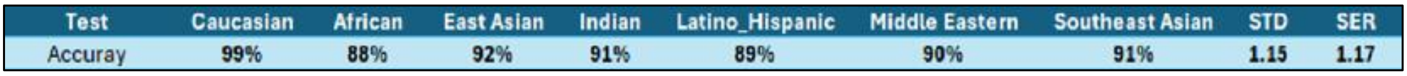
The training for our GAN model is scheduled for 15,000 iterations. The discriminator loss measures how well the discriminator is able to distinguish between real and generated samples. A lower value is better for the generator as it suggests that the generator is successfully fooling the discriminator. The generator loss measures how well the generator is fooling the discriminator. A lower value is better for the generator as it suggests that the generator is generating samples that are more convincing to the discriminator. Generally, we'd want the generator and the discriminator to improve together. The generator wants to minimize its loss by generating realistic data and the discriminator wants to minimize its loss by correctly classifying real and generated data. According to our results, both losses are decreasing over time which is a positive sign for the training process. Our GAN model helps us generate our realistic synthetic human faces which we will then and combine and use to compare with a dataset without synthetic images and see if we can identify any differences. About 25 thousand diverse images are generated and classified into the 7 mentioned races above. This was the thesis' most challenging task as it was tricky because GANS contain 2 separately trained network: Discriminator and Generator that they must balance, and it demanded extremely high computational performances.



Test	Caucasian	African	East Asian	Indian	Latino_Hispanic	Middle Eastern	Southeast Asian	STD	SER
Accuray	99%	83%	92%	90%	87%	88%	91%	1.77	1.35

Fig 11 STD, Accuracy, SER levels Before Synthetic Data

For our recognition model, The STD of 1.77 indicates the variability or spread of the predictions around the mean (average). A higher STD suggests that the predictions for different instances (images) vary more from the average prediction. A higher STD could be an indication that the model is not consistently confident or accurate across different instances. It struggles with certain cases or exhibit inconsistency in predicting races for instance Africans, Latino_Hispanic. The SER of 1.35 suggests a skewed error distribution. SER is a measure of how imbalanced the error is among different classes. A SER greater than 1 indicates that some classes may be more prone to misclassifications than others. It also implies that certain races might be more challenging for the model, leading to a higher error rate for those specific races.



Test	Caucasian	African	East Asian	Indian	Latino_Hispanic	Middle Eastern	Southeast Asian	STD	SER
Accuray	99%	88%	92%	91%	89%	90%	91%	1.15	1.17

Fig 12 STD, Accuracy, SER levels After Synthetic Data

After adding our synthetic data, we can see that now our STD is at 1.15 and SER is at 1.17. For the decrease in STD, it indicates a reduction in the spread of model predictions, suggesting an increased stability for consistency across the different races. The decrease in SER indicates a reduction in the skewed error rates, suggesting a more balanced performance across the different races.

CHAPTER FOUR

RESULTS AND ANALYSIS

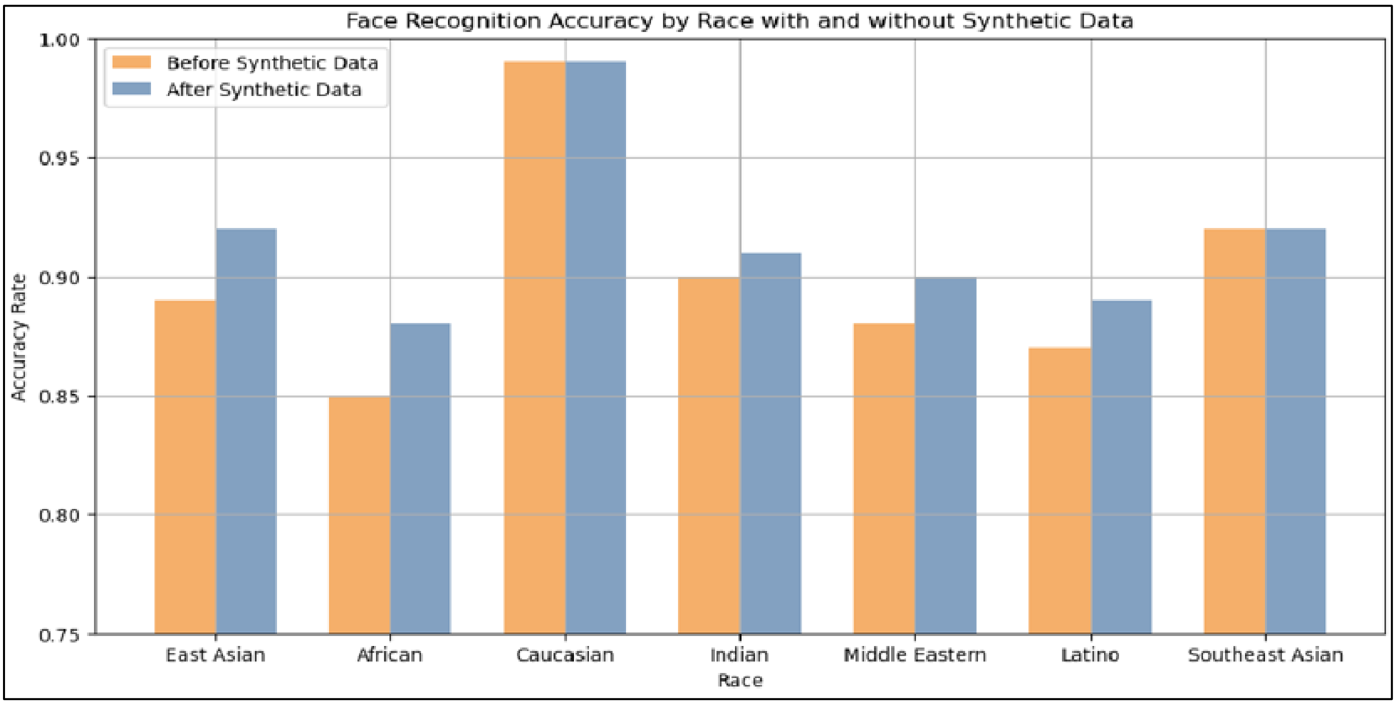


Fig 13 Results And Comparison With And Without Our Synthetic Data

Caucasians have the highest recognition accuracy, followed by East Asians. Southeast Asians, Indians. Middle Eastern, and Latino groups show moderate accuracy levels. African individuals have the lowest face recognition accuracy. The significant variation in accuracy rates indicates potential biases in the facial recognition software's performance among different races. We could see that after we added our synthetic data to enrich training datasets, it can increase the diversity of facial features that the facial recognition system was exposed to. This, in turn helped mitigate biases and improved accuracy across different racial groups. The synthetic data acted as a supplement, which could also fill in gaps where real-world data may be scarce or non-representative. Some further studies showed that the noise found in real facial images and the noise found in synthetically generated images are different for people of colour. The added synthetic images appear to contribute positively to the model's performance, resulting in increased accuracy. It's noteworthy that the accuracy for the African race shows a significant improvement, going from 85% to 88%. Races with initially high accuracy levels maintained their accuracy, while races with lower initial accuracy showed improved. This showed that the model has stability.

In summary, according to our result and analysis, we found the following:

➢ *Positive Impact on Accuracy:*

- Synthetic data has positively influenced accuracy, especially for races with lower initial accuracy levels.
- The model is now more adept at recognizing faces across diverse racial groups.

➢ *Enhanced Stability:*

- Reduction in STD indicates improved stability and consistency in predictions.
- Synthetic data has contributed to a more uniform and reliable model.

➢ *Balanced Performance:*

- SER reduction signifies a more balanced error rate distribution across races.
- Synthetic data has mitigated biases and improved fairness in predictions.

➢ *Model Robustness:*

- The model has become more robust, as seen in increased accuracy and reduced variability.

CHAPTER FIVE
CONCLUSION

The aim of this thesis was to explore and implement generative models as a solution to enhance the diversity of datasets used in facial recognition technology. According to our results, we can say that the use of generative models can be helpful and as such this thesis incites the use of the latter as a mean to mitigate racial biasness as much as possible. The questions that were posed by a lot of people was whether AI is bias; we have numbers which suggests that it is if we take into consideration our model before we added our synthetic data but as of 2023, we know that AI aren't sentient being and as such we know that they do not have a mind of their own and if we go by the definition of biasness which is: "Prejudice in favour of or against one thing, person, or group compared with another, usually in an unfair or negative way", we can say that AI is not biassed at least on its own. The problem could come from a biased training data. If the training data used to develop the face recognition system is not representative of the global population and is skewed towards specific racial or ethnic groups, the model will likely exhibit bias. If certain racial or ethnic groups are underrepresented in the training data, the model may have difficulty accurately recognizing faces from those groups, leading to biased performance. Imbalances in the number of images per racial or ethnic group can lead to biased outcomes. If the dataset is skewed, the model may be more proficient in recognizing faces from overrepresented groups. Biases in the process of labelling or annotating training data can contribute to racial bias. Subjective judgments made during data labelling may unintentionally introduce bias into the training set. Facial recognition algorithms may struggle with diverse facial features, and certain features common in specific racial or ethnic groups may be poorly represented in the training data. Unequal representation of lighting conditions and environments in the training data can result in biased performance. If certain groups are disproportionately represented in specific conditions, the model may struggle in other scenarios.

This thesis aids to mitigate the biases that could arise if the above-mentioned issues arise. However, there are several other causes of bias that this thesis cannot fully tackle. This thesis, however, serves as a stepping stone for other researchers to work further on generative models.

➢ *Statement of Limitations and Future Works:*

This thesis faced a lot of challenges and limitations which includes: Limited CPU/GPU Power to being able to handle big data and the time allocated for this thesis is merely a few months whereas researchers spend years on this topic and as of January 2024, we still haven't found a concrete solution that completely eradicates racial biasness in AI. It maybe because currently, we are faced with these following issues: Algorithmic Bias: The design and implementation of the face recognition algorithm itself can introduce bias. If the algorithm is not thoroughly tested for fairness across different racial and ethnic groups, it may exhibit disparities in performance even with a balanced dataset. Historical biases and societal inequalities can be reflected in training data. If the data collection process is influenced by systemic biases, the resulting models can inherit and perpetuate those biases. Homogeneous development teams may unintentionally embed biases into algorithms. Diverse teams with varied perspectives can contribute to more comprehensive and fair face recognition systems.

For future works, since addressing racial bias in face recognition involves a holistic approach, including diverse and representative training data, rigorous testing for fairness, ongoing monitoring, and ethical considerations throughout the development lifecycle, more resources in terms of testing needs to be put, for instance real world test cases via the use of live webcams, amongst others.

REFERENCES

[1]. Patel, Ripal & Rathod, Nidhi & Shah, Ami. (2012). Comparative Analysis of Face Recognition Approaches: A Survey. 57.

[2]. Eric Sullivan, Legislative Research Analyst. (2021). MONTANA LEGISLATIVE SERVICES DIVISION Office of Research & Policy Analysis. Facial Recognition Technology.

[3]. C. Bouras and E. Michos, "An online real-time face recognition system for police purposes," 2022 International Conference on Information Networking (ICOIN), Jeju-si, 2022, pp. 62-67, doi: 10.1109/ICOIN53446.2022.9687212.

[4]. P. J. Thilaga, B. A. Khan, A. A. Jones and N. K. Kumar, "Modern Face Recognition with Deep Learning," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 1947-1951, doi: 10.1109/ICICCT.2018.8473066.

[5]. M. F. A. Rahman et al., "Facial Recognition Development to Detect Corporate Employees Stress Level," 2019 IEEE International Conference on Engineering, Technology and Education (TALE), Yogyakarta, Indonesia, 2019, pp. 1-6, doi: 10.1109/TALE48000.2019.9225909.

[6]. Tara MitchellRyann, M. HawJeffrey E, PfeiferChristian, A. MeissnerChristian, A. Meissner “Racial Bias in Mock Juror Decision-Making: A Meta-Analytic Review of Defendant Treatment”, 2015, ResearchGate, doi: 10.1007/s10979-005-8122-9

[7]. B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge and A. K. Jain, "Face Recognition Performance: Role of Demographic Information," in IEEE Transactions on Information Forensics and Security, vol. 7, no. 6, pp. 1789-1801, Dec. 2012, doi: 10.1109/TIFS.2012.2214212.

[8]. L. Masupha, T. Zuva, S. Ngwira and O. Esan, "Face recognition techniques, their advantages, disadvantages and performance evaluation," 2015 International Conference on Computing, Communication and Security (ICCCS), Pointe aux Piments, Mauritius, 2015, pp. 1-5, doi: 10.1109/CCCS.2015.7374154.

[9]. K. Marwa and O. Kais, "Current Challenges of Facial Recognition using Deep Learning," 2022 19th International Multi-Conference on Systems, Signals & Devices (SSD), Sétif, Algeria, 2022, pp. 1980-1986, doi: 10.1109/SSD54932.2022.9955857.

[10]. E. Ntoutsi, "Keynote Speech 2: Bias and Discrimination in AI Systems: From Single-Identity Dimensions to Multi-Discrimination," 2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS), Valencia, Spain, 2023, pp. 2-2, doi: 10.1109/ICCNS58795.2023.10193086.

[11]. H. F. Menezes, A. S. C. Ferreira, E. T. Pereira and H. M. Gomes, "Bias and Fairness in Face Detection," 2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Gramado, Rio Grande do Sul, Brazil, 2021, pp. 247-254, doi: 10.1109/SIBGRAPI54419.2021.00041.

[12]. Zamira Rahim. 19.September.2019, Independent.co.uk/news/uk/home-news/black-man-lips-passport-photo-home-office-joshua-bada-a9111711l, “'I was a bit annoyed': Black man's lips flagged by passport checker as open mouth. ‘My mouth is closed, I just have big lips,’ Joshua Bada tells machine”.

[13]. J. Goodrich, “Facial Recognition Faces More Proposed Bans Across U.S,” 2019, The.Institude.

[14]. M. Jha, A. Tiwari, M. Himansh and V. M. Manikandan, "Face Recognition: Recent Advancements and Research Challenges," 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2022, pp. 1-6, doi: 10.1109/ICCCNT54827.2022.9984308.

[15]. Alex Najibi. “Racial Discrimination In Face Recognition Technology”. 2020. Science Policy, Special Edition: Science

[16]. Policy And Social Justice. Harvard University, The Graduate School Of Arts And Sciences.

[17]. Ntungila, Jef. (2021). The Application of Facial Recognition Technology by Law Enforcement. 10.13140/RG.2.2.13570.40646.

[18]. Joy Buolamwini and Timnit Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification", Conference on fairness accountability and transparency, pp. 77-91, 2018.
[19]. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman and Aram Galstyan, "A survey on bias and fairness in machine learning", ACM Computing Surveys (CSUR), vol. 54, no. 6, pp. 1-35, 2021.
[20]. Shruti Nagpal, Maneet Singh, Richa Singh and Mayank Vatsa, "Deep learning for face recognition: Pride or prejudiced?", arXiv preprint arXiv:1904.01219, 2019.
[21]. Harini Suresh and John V Guttag, "A framework for understanding unintended consequences of machine learning", arXiv preprint arXiv:1901.10002, 2019.
[22]. Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao and Yaohai Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network", The IEEE International Conference on Computer Vision (ICCV), October 2019.
[23]. Sixue Gong, Xiaoming Liu and Anil K Jain, "Mitigating face recognition bias via group adaptive classifier", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3414-3424, 2021
[24]. Yuning Qiu, Teruhisa Misu, Carlos Busso, “Unsupervised Scalable Multimodal Driving Anomaly Detection”, IEEE
[25]. TRANSACTIONS ON INTELLIGENT VEHICLES, VOL. 8, NO. 4, APRIL 2023
[26]. WEI LI , JINBAO SUN2 , JING ZHANG , BOCHENG ZHANG, “Face Recognition Model Optimization Research Based on Embedded Platform”, DOI: 10.1109/ACCESS.2023.3277495 ,June 2023
[27]. J. Engel, V. Koltun and D. Cremers, "Direct Sparse Odometry," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 3, pp. 611-625, 1 March 2018, doi: 10.1109/TPAMI.2017.2658577.
[28]. H. Mav, A. Mokashi, S. Nanduri and V. Pinjarkar, "Face Recognition and Adversarial Masking Techniques," 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022, pp. 1-7, doi: 10.1109/INCET54531.2022.9825044.
[29]. Seyma Yucer, Samet Akçay, Noura Al-Moubayed and Toby P Breckon, "Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 18-19, 2020.
[30]. Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, et al., "Towards fairness in visual recognition: Effective strategies for bias mitigation", Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8919-8928, 2020.
[31]. M. Gwilliam, S. Hegde, L. Tinubu and A. Hanson, "Rethinking Common Assumptions to Mitigate Racial Bias in Face Recognition Datasets," 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 2021, pp. 4106-4115, doi: 10.1109/ICCVW54120.2021.00458.
[32]. N. Stojanović, Z. Qiang, C. Prodaniuc and F. Karinou, "Eye deskewing algorithms for PAM modulation formats in IM-DD transmission systems," 2017 Optical Fiber Communications Conference and Exhibition (OFC), Los Angeles, CA, USA, 2017, pp. 1-3.
[33]. M. Wang and W. Deng, "Mitigating Bias in Face Recognition Using Skewness-Aware Reinforcement Learning," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 9319-9328, doi: 10.1109/CVPR42600.2020.00934.
[34]. Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao and Yaohai Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network", Proceedings of the IEEE International Conference on Computer Vision, pp. 692-702, 2019.
[35]. James Zou and Londa Schiebinger, Ai can be sexist and racistlits time to make it fair, 2018.
[36]. Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge and Anil K Jain, "Face recognition performance: Role of demographic information", IEEE Transactions on Information Forensics and Security, vol. 7, no. 6, pp. 1789-1801, 2012.
[37]. L. Huang et al., "Gradient Attention Balance Network: Mitigating Face Recognition Racial Bias via Gradient Attention," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 2023, pp. 38-47, doi: 10.1109/CVPRW59228.2023.00009.
[38]. Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, "Deep residual learning for image recognition", Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.
[39]. J. Ge, W. Deng, M. Wang and J. Hu, "FGAN: Fan-Shaped GAN for Racial Transformation," 2020 IEEE International Joint Conference on Biometrics (IJCB), Houston, TX, USA, 2020, pp. 1-7, doi: 10.1109/IJCB48548.2020.9304901.
I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative adversarial networks", *NIPS*, 2014.
[40]. Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim and J. Choo, "Stargan: Unified generative adversarial networks for multidomain image-to-image translation", *CoRR*, 2017
[41]. Y. Jo and J. Park, "SC-FEGAN: face editing generative adversarial network with user's sketch and color", *CoRR*, 2019.
[42]. P. Isola, J. Zhu, T. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks", *CoRR*, vol. abs, 2016.
[43]. M. Wang, W. Deng, J. Hu, J. Peng, X. Tao and Y. Huang, "Racial faces in-the-wild: Reducing racial bias by deep unsupervised domain adaptation", *CoRR*, 2018.
[44]. S. Yucer, F. Tektas, N. A. Moubayed and T. P. Breckon, "Measuring Hidden Bias within Face Recognition via Racial Phenotypes," 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022,
[45]. pp. 3202-3211, doi: 10.1109/WACV51458.2022.00326.

[46]. Cynthia Feliciano, "Shades of race: How phenotype and observer characteristics shape racial classification", American Behavioral Scientist, 2016.

[47]. Abdulla Fakhro, Hyung Woo Yim, Yong Kyu Kim and Anh H Nguyen, "The evolution of looks and expectations of asian eyelid and eye appearance" in Seminars in plastic surgery., Thieme Medical Publishers, 2015.

[48]. Theiab Alzahrani, Waleed Al-Nuaimy and Baidaa Al-Bander, "Integrated multi-model face shape and eye attributes identification for hair style and eyelashes recommendation", *Computation*, 2021

[49]. Yoonho Lee, Euitae Lee and Won Jin Park, "Anchor epicanthoplasty combined with out-fold type double eyelidplasty for asians: do we have to make an additional scar to correct the asian epicanthal fold?", *Plastic and reconstructive surgery*, 2000.

[50]. Ziqing Zhuang, Douglas Landsittel, Stacey Benson, Raymond Roberge and Ronald Shaffer, "Facial anthropometric differences among gender ethnicity and age groups", Annals of occupational hygiene, 2010.

[51]. De La Mettrie, Didier Saint-Léger, Genevievève Loussouarn, Annelise Garcel, Crystal Porter and André Langaney, "Shape variability and classification of human hair: a worldwide approach", Human biology, 2007.

[52]. Jonathan L Rees, "Genetics of hair and skin color", Annual review of genetics, 2003.

[53]. https://www.chicagofaces.org/

[54]. https://www.kaggle.com/datasets/kleinertee/fairface

[55]. https://www.kaggle.com/datasets/jessicali9530/celeba-dataset