Fairness Evaluation in Facial Obfuscation Methods

Shubham Shah
University of Texas
Arlington, Texas
shubhambhavinku.shah@mavs.uta.edu

Shraddha Varekar
University of Texas
Arlington, Texas
sxv1887@mavs.uta.edu

Suhas Holla Karkada Chandrashekar
University of Texas
Arlington, Texas
sxk8003@mavs.uta.edu

Rahul Nagireddi
University of Texas
Arlington, Texas
rxn8893@mavs.uta.edu

Keywords: Differential Privacy, Face Obfuscation, Comparative Evaluation, Fairness detection

ACM Reference Format:
Shubham Shah, Shraddha Varekar, Suhas Holla Karkada Chandrashekar, and Rahul Nagireddi. 2018. Fairness Evaluation in Facial Obfuscation Methods. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 34 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The rapid growth of social media and image-sharing platforms has led to an unprecedented increase in the volume of image data generated and shared daily. This massive proliferation of visual content has enabled the development of new technologies and applications centred around image data. However, as a consequence, concerns have arisen regarding the privacy protection of individuals depicted in these images. Face de-identification has become a critical process that aims to eliminate identifying information from photographs to safeguard privacy.

Face recognition technology is employed in various applications, such as law enforcement, security systems, and social networking platforms. Despite their utility, these technologies may also infringe on individuals' privacy by tracking them without their consent. Face de-identification methods offer a means to preserve privacy by removing personally identifiable information from facial images.

In this paper, we explore the fairness and effectiveness of various face de-identification methods, focusing on live video de-identification and analysing their performance across diverse demographic, race, and gender groups. We evaluate four state-of-the-art de-identification methods—DP-PIX, DP-SAMP, DP-SNOW, and CNN based Live deidentification—by applying them to three distinct datasets: BFW[9], DemogPairs[5], and RFW[11]. Our research questions aim to assess the obfuscation success of each method, measure their False Negative Rate (FNR) across different demographics, and identify any significant performance differences among the methods. By examining the performance of these face de-identification methods, we aim to contribute to the development of more fair and effective techniques that can better protect individuals' privacy in an increasingly interconnected world.Also, we tried to answer the following researhch questions:

- How do different obfuscation methods perform in terms of obfuscation success for different datasets/live video de-identification?
- How does the performance of these face de-identification methods vary across different datasets?
- What is the obfuscation success for each demographic, race, and gender when applying these methods?
- Are there any significant differences in performance among the methods when comparing obfuscation success?
- How can the findings of this research contribute to the development of more fair and effective face de-identification methods?

2 Background and Related Work

In recent years, there has been an increasing concern for privacy preservation of image data, particularly sensitive information such as faces and irises. Traditional obfuscation methods such as blurring or pixelization have been widely used, but they are susceptible to inference attacks and do not provide quantifiable privacy guarantees. To address this issue, several differentially private obfuscation methods have been proposed, which provide rigorous privacy guarantees. One such method is DP-SVD, which outperforms other methods on several privacy and utility measures. In Reilly et al. [8], an extensive evaluation of differentially private obfuscation methods, including DP-SVD, was conducted in the context of obfuscating face and iris images. The authors

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

synthesised existing differentially private obfuscation methods, analysed their privacy guarantees, and conducted an empirical evaluation with real-world datasets. Their results highlight the importance of considering both privacy and utility when applying differentially private image obfuscation methods. Another approach to image privacy preservation is to enhance traditional obfuscation methods with differential privacy. DP-Blur is a method that extends the standard differential privacy notion to image data, which protects individuals, objects, and/or their features. In [3], the authors develop differentially private methods for two popular image obfuscation techniques: pixelization and Gaussian blur. They demonstrate the utility and privacy of their methods through empirical evaluation with real-world datasets and show that their methods effectively reduce the success rate of re-identification attacks.

Similarly, in [2], the authors propose sharing pixelated images with rigorous privacy guarantees by extending the standard differential privacy notion to image data. Their method is shown to effectively reduce the success rate of re-identification attacks despite its simplicity.Empirical evaluation with real-world datasets demonstrates the utility and efficiency of their approach, highlighting the importance of considering privacy preservation techniques in the context of sharing image data. Face de-identification in order to preserve privacy of an individual was an ongoing research over years and several methodologies are suggested to solve this challenge. GAN based models are one of the popular models used to address this problem. Major contributions are done in de-identification in still images. In the paper, authors proposed a novel approach to solve this live video while maintaining pose, expression.

Our study differs from existing work in several ways. First, by employing diverse datasets, we provide a more comprehensive understanding of each method's capabilities across various demographic, race, and gender combinations. Second, we emphasise fairness and investigate the FNR for each demographic, race, and gender, ensuring that the methods are unbiased and effective for different population groups. Third, we apply these techniques to live video de-identification, adding a layer of complexity and real-world applicability to our research. Lastly, we utilise the FacePlusPlus compare tool as an external metric for performance evaluation, ensuring a standardised and objective assessment of the obfuscation success of each method.

3 Ethical Considerations

As we delve into the realm of face de-identification, it is crucial to consider the ethical implications of our research and the technologies we evaluate. While the primary goal of face de-identification methods is to protect individuals' privacy, it is essential to acknowledge the potential for misuse or unintended consequences. One ethical consideration is the potential for bias in the algorithms used in face de-identification methods. Ensuring fairness across demographic, race, and gender groups is a significant aspect of our research, as biased algorithms can exacerbate existing social inequalities and discrimination. By evaluating the performance of de-identification methods across diverse datasets, we aim to identify and address potential biases and contribute to the development of more equitable solutions.

Another ethical consideration is the possibility of malicious actors using face de-identification methods to conceal their identity while engaging in illegal or unethical activities. While our research aims to develop effective techniques for privacy protection, we must also consider the potential for these methods to be exploited for nefarious purposes. This underscores the importance of developing robust solutions that can distinguish between legitimate use cases and harmful intent. Furthermore, the application of face de-identification methods in public spaces may raise concerns about the right to privacy versus the right to freedom of expression and information. Striking a balance between these rights requires a careful examination of the contexts in which face de-identification is employed and the potential impact on the public interest.

Finally, as researchers, we must adhere to ethical guidelines in our data collection and usage. This includes obtaining informed consent from individuals whose images are used in the datasets, ensuring that the data is used solely for the purposes of the research, and taking necessary measures to protect the privacy and confidentiality of the data subjects. In conclusion, the development and evaluation of face de-identification methods must be approached with a strong sense of ethical responsibility. By considering the potential ethical implications of our research, we can contribute to the creation of more fair, effective, and ethically sound privacy protection solutions.

4 Data Collection

To ensure a comprehensive evaluation of face de-identification methods across diverse demographic, race, and gender groups, we employed a data-driven approach that encompasses three distinct datasets: BFW, DemogPairs, and RFW. These datasets were chosen for their specific focus on demographic balance, ethnic balance, and racial diversity, providing a solid foundation for our research.

The first dataset, Balanced Faces in the Wild (BFW)[9], focuses on gender and ethnic balance. It contains 20,000 images of multiple individuals, collected from public domain sources. By incorporating a diverse sample of faces, the BFW dataset allows us to analyse the performance of face de-identification methods across various gender and ethnic groups.**Table 31** in **Appendix A** gives a brief metadata about the dataset.

The second dataset, DemogPairs[5], emphasises demographic balance. It consists of 10,800 images of multiple individuals and has been curated with an equal representation of age, gender, and ethnicity. This dataset enables the evaluation of face de-identification methods across a wide range of demographics, providing insights into the potential biases and effectiveness of each method. **Table 30** in **Appendix A** gives a brief metadata about the dataset.

Lastly, we have the Racial Faces in the Wild (RFW) dataset[11], which focuses on racial diversity. This dataset contains 40,607 face images from four racial groups—African, Asian, Caucasian, and Indian—allowing for the assessment of face de-identification methods across different racial groups. By utilising the RFW dataset, we can better understand how these methods perform in various racial contexts and identify areas for improvement. **Table 29** in **Appendix A** gives a brief metadata about the dataset.

Our data collection approach ensures a thorough analysis of the face de-identification methods under consideration. By utilising diverse datasets that represent various demographic, ethnic, and racial groups, we can effectively evaluate the performance of each method and assess their fairness and effectiveness. This comprehensive approach enables us to draw meaningful conclusions and contribute to the development of more fair and effective face de-identification methods.

5 Methodology

Differential privacy is a rigorous privacy-preserving technique that aims to protect sensitive information in datasets while still allowing useful data analysis. The primary goal of differential privacy is to ensure that the output of a data analysis process does not reveal any information about individual records in the dataset. This is achieved by adding a controlled amount of random noise to the data, which helps maintain the privacy of individual records while preserving the overall statistical properties of the dataset.

Differential privacy is particularly relevant in the context of face de-identification methods, as it can help mitigate the risk of unauthorised access to personal information in facial images. By applying differential privacy techniques, it is possible to ensure that the de-identified faces do not reveal any sensitive information about the individuals they depict, while still allowing for the analysis and use of the anonymized data.

In our research, we evaluate four state-of-the-art face de-identification methods that incorporate differential privacy techniques: *DP-PIX, DP-SAMP, DP-SNOW* and *CNN-based.*

5.1 DP-PIX

The *DP-PIX* method is a pixel-based de-identification technique that applies differential privacy by perturbing individual pixel values in the image. The method adds carefully calibrated noise to the pixels, ensuring that the output image is both private and visually similar to the original image. This allows for privacy preservation while maintaining the overall structure and appearance of the face.

Algorithm 1: DP-Pix

Input : Input image $\mathcal{I}$, Privacy budget ϵ,
Block size b, Number of pixels m
Output: Obfuscated image satisfying ϵ-DP

blocks $\leftarrow$ partition $\mathcal{I}$ into blocks of size $b \times b$
foreach block *in* blocks **do**
  average $\leftarrow$ average pixel intensity of block
  noise $\leftarrow$ noise drawn from $Laplace(0, \frac{255m}{b^2\epsilon})$
  assign intensity of pixels in block to average + noise
end
$\hat{\mathcal{I}} \leftarrow$ output image constructed from blocks

The *DP-PIX* method [2] is the first approach to provide differential privacy guarantees when publishing individual images. Considering the vast number of pixels in a typical image, perturbing each pixel in the source image directly would result in low utility. To strike a balance between privacy and utility, *DP-PIX* employs pixelization and the Laplace mechanism to comply with differential privacy.

5.1.1 m-neighbourhood. The concept of an m-neighbour [2] is introduced to define neighbouring images in the context of differential privacy. Two images, I_1 and I_2, are deemed neighbouring if they differ by no more than m pixels. By adjusting the value of m, the data owner can manage the privacy protection provided by DP-PIX: larger m values imply indistinguishability in a broader neighbourhood, thus offering stronger privacy protection.

5.1.2 Private Pixelization. To minimise the amount of noise required for differential privacy, DP-PIX utilises pixelization. Pixelization, also known as mosaicing, breaks down an image into blocks by overlaying a grid on the source image. Each grid cell (i.e., super-pixel) consists of $b \times b$ pixels. The value of each super-pixel is calculated by averaging all pixels contained within the grid cell. To achieve ϵ-DP, a perturbation noise is sampled from a *Laplacian distribution* with a mean of 0 and a scale of $\frac{255m}{b^2\epsilon}$ and added to each grid cell. The steps of the DP-PIX method are outlined in **Algorithm 1**.

5.2 DP-SAMP

The *DP-SAMP* method is a sampling-based approach that leverages differential privacy. It involves randomly selecting a subset of facial features from the original image and replacing them with corresponding features from other images in the dataset. This process ensures that the de-identified face is a composite of multiple individuals, making it difficult to link the output image to any specific person.

Algorithm 2: DP-Samp

Input : Input image $\mathcal{I}$, Privacy budget ϵ, Number of clusters k, Number of pixels m
Output: Obfuscated image satisfying ϵ-DP

perform pixel clustering to generate k clusters
calculate most frequent intensity in each cluster $(\Psi_1, ..., \Psi_k)$
// Budget allocation
foreach $\Psi_i, i \in [1, k]$ **do**
  compute the privacy budget $\epsilon(\Psi_i)$ with Eq. 3
end
// Pixel sampling
foreach $\Psi_i, i \in [1, k]$ **do**
  compute maximum x_i with Eq. 4
  randomly select x_i pixels from $\mathcal{I}$ with intensity Ψ_i to preserve in output image $\hat{\mathcal{I}}$
end
// Interpolation
linear interpolate non-sampled pixels in $\hat{\mathcal{I}}$

A recent study [10] introduced a pixel-sampling technique designed to safeguard visual elements, such as people and objects, in videos. In this research, we develop a new method called *DP-Samp*, which adapts the previously mentioned technique to protect up to m pixels in a source image. *DP-Samp* consists of the following four stages:

5.2.1 Pixel Clustering. The clustering's objective is to identify pixel intensities useful for image reconstruction. A simple approach involves selecting the most frequent intensities in the image; however, this method may not capture the structures of images with large regions containing slightly varying intensities. The original study employed multi-scale analysis to divide each visual element in a video into k cells. Since this method is not applicable to a single image, we propose generating k pixel clusters using K-means. The most frequent intensity in each cluster, $\psi_{1:k}$, will be candidates for pixel sampling. This clustering step is performed in a public setting, as is the multi-scale analysis in the original study. Integrating differentially private clustering techniques is possible but beyond this study's scope.

5.2.2 Budget Allocation. The privacy budget ϵ is divided among all selected intensities in ψ. DP-Samp assigns higher privacy budgets to more frequently occurring intensities. Let Freq(ψ_i) be the number of pixels in the source image with intensity ψ_i, then the privacy budget for ψ_i can be computed as:

$$\epsilon(\psi_i) = \frac{\epsilon \cdot Freq(\psi_i)}{\sum_{j=1}^{k} Freq(\psi_j)}$$

5.2.3 Pixel Sampling. From each intensity ψ_i, we randomly sample x_i pixels from the source image, maintaining their location and intensity. The value of x_i is determined by:

$$max(x_i), s.t. \binom{c_i}{x_i} \Big/ \binom{c_i - m}{x_i} \leq e^{\epsilon(\psi_i)}$$

where c_i is the count of pixels with intensity ψ_i in the input, and m is the number of pixels allowed to differ between neighbouring images. Selecting x_i according to this equation satisfies ϵ-DP, with an analysis similar to the original study.

5.2.4 Interpolation. DP-Samp performs linear interpolation on the sampled pixels to estimate the values of non-sampled pixels. Leveraging the post-processing property of Differential Privacy[1], interpolation does not incur additional privacy loss in the output image ^I

5.3 DP-SNOW

DP-SNOW is a de-identification method that combines differential privacy with the concept of k-anonymity. The algorithm groups similar faces together and then applies a noise-adding mechanism to ensure privacy. By maintaining the overall structure of the face while adding noise, DP-SNOW produces de-identified faces that are visually similar to the original images, while still protecting individual privacy.

Algorithm 3: Snow

Input : Input image $\mathcal{I}$, Privacy budget δ
Output: Obfuscated image satisfying $(0, \delta)$-DP

p $\leftarrow (1 - \delta)$
$\mathcal{S} \leftarrow$ random subset of $p \cdot \mathcal{I}_{width} \cdot \mathcal{I}_{height}$ pixels in $\mathcal{I}$
$\hat{\mathcal{I}} \leftarrow \mathcal{I}$
foreach pixel *in* $\mathcal{S}$ **do**
  set intensity of pixel to 127 in $\hat{\mathcal{I}}$
end

This method [6] proposes the addition of pixel-level noise ("snow") to the RGB images via randomly re-assigning the pixel intensities to a constant value i.e. 127 for each channel. The method is outlined in **Algorithm 3**. Let us denote an image as I(x) where x is the index of each pixel of value of image I. For example an image of size 10 x 10, the value of x = [1,2,3,...100]. And the intensity of each pixel in x is represented by I(x) where I(x) ∈ [0, 255]. Then, a subset S of pixels from the image is selected randomly of size $p \cdot I_{rows} \cdot I_{cols}$ where p = 1 - δ, I_{rows}, I_{cols} are the image dimensions. Based on this a new image I'(x) is constructed such that:

$$I'(x) = \begin{cases} 127, & \text{if } x \in S. \\ I(x), & \text{if } x \notin S. \end{cases}$$

Intuitively, as the value of parameter p changes, the noisy content in the image relatively changes. If p increases so does the noise and the same is true if the value of p decreases. As shown in [6], the method achieves (0, δ) differential privacy with δ = 1- p.

5.4 CNN-based

A convolutional autoencoder, a deep learning model created primarily for image processing tasks [4], is used in this study's obfuscation procedure. An encoder and a decoder are the two basic parts of an autoencoder. The decoder reconstructs the input image from this representation after the encoder has compressed it into a lower-dimensional form known as latent space. The autoencoder learns to capture key face features during the obfuscation process while retaining a specific level of obfuscation to protect the privacy of the individuals in the images. We collected a dataset of facial images representing five demographic groups: African, Asian, Caucasian, and Indian. The dataset consists of a balanced number of images for each demographic group, ensuring equal representation. Each image was cropped and resized to a standard size to facilitate further processing. Convolutional layers are used in a convolutional autoencoder, a variation on the traditional autoencoder design, to enhance the ability to detect local patterns and spatial information in images. In order to gradually lower the spatial dimensions, the encoder consists of several convolutional layers with tiny filters and pooling layers. The decoder uses upsampling layers and transposed convolutional layers, commonly referred to as deconvolutional layers, to recover the spatial dimensions and recreate the original image.

5.4.1 Training the Encoder. The autoencoder is trained on a large dataset of facial images using unsupervised learning. The objective is to minimize the reconstruction error between the input images and the images reconstructed by the autoencoder. The training process adjusts the model's weights to capture the most relevant features of the facial images while retaining the obfuscation necessary to protect privacy [4]. Once the training is complete, the autoencoder can effectively obfuscate facial images across a wide range of demographics.

5.4.2 Obfuscation Process. We use the pre-trained autoencoder to process the original images for each demographic group to produce obfuscated versions. The input images are compressed by the autoencoder's encoder component into lower-dimensional latent space representations, which keep the most important facial details while excluding potentially sensitive information. The images are then rebuilt from these representations by the decoder, producing obfuscated images with a certain level of privacy protection. By evaluating these four methods, we aim to assess their effectiveness in terms of obfuscation success, fairness across demographic, race, and gender groups, and their overall performance.

6 Evaluation

In this section, we intend to show the different metrics and experiments we used to evaluate the above given methods.

6.1 DP Methods

In this section, we discuss the evaluation of the DP methods used in our experiments. We will cover the metrics used for assessment and the experiments carried out with different parameters for each DP method.

6.1.1 Input Parameters.

- DP-PIX — *Privacy Budge (δ) = 5, Number of Pixels (m) = 16, Number of blocks (b) = 12*
- DP-SAMP — *Privacy Budge (δ) = 25, Number of Pixels (m) = 12, Number of clusters (k) = 24*
- DP-SNOW — *Privacy Budge (δ) = 0.5*

6.1.2 Metrics. The primary metric used to evaluate the performance of these DP methods is obfuscation success confidence. This metric is determined by running the Face-PlusPlus (FPP) [7] compare tool on the obfuscated images. By comparing obfuscation success confidence across datasets and pairs, we can interpret the effectiveness of each method and understand their strengths and weaknesses in different scenarios.

6.1.3 Experiments. We conducted experiments on the three datasets mentioned earlier (BFW[9], DemogPairs[5], and RFW[11]) using the DP methods with the specified parameters. For each method, we applied the respective parameters and processed the images. Afterward, we ran the FPP compare tool to evaluate the obfuscation success confidence for each method. By plotting the *Obfuscation Success Rate*, *True False rate* and *False Negative rate* across datasets and pairs, we were able to analyse the performance of the DP methods in terms of their ability to protect individuals' privacy while maintaining the utility of the images. This analysis provides insights into the strengths and limitations of each method and informs the development of more effective and fair face de-identification techniques.

6.2 CNN-based

The evaluation methodology employed in this study aims to assess the performance of the obfuscation process and identify potential biases across different demographic groups. To do this, we compare the original and obfuscated photos using a variety of quantitative parameters. With the help of the specified metrics, we can evaluate many facets of the obfuscation procedure, such as the size of the discrepancies created, the preservation of structural data, and the general effectiveness of the obfuscated photos. The obfuscated images are compared to the original images using a set of evaluation metrics, including *Mean Absolute Error (MAE)*, *Mean Squared Error (MSE)*, *Structural Similarity Index (SSIM)*, and *Peak Signal-to-Noise Ratio (PSNR)*. These metrics allow us to assess the quality of the obfuscation process and measure the difference between the original and obfuscated images. By comparing these metrics across different demographic

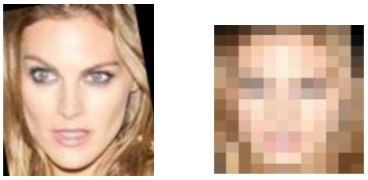
groups, we can analyse the obfuscation process's effectiveness and fairness.

7 Results

7.1 DP Methods

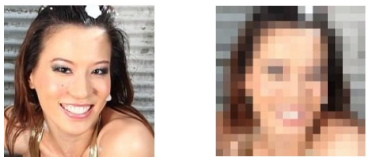
In this section, we present and interpret the results obtained from our experiments using the DP methods on the three datasets (*BFW, DemogPairs, and RFW*). We have applied the DP methods, generated obfuscated images, and evaluated their obfuscation success confidence using the FacePlusPlus compare tool [7].

7.1.1 DP-Pix. In this section, we present the results obtained from DP-PIX [2] based on the original and obfuscated image samples for each dataset. The first image is the input image and second is the resulting output image generated from the method as shown in **Algorithm 1**. The parameters used for the method to yeild the obfuscated images from the original source images are: *Privacy Budget (δ) = 5*, *Number of Pixels (m) = 16* and *Block Size (b) = 12*.



(a) Source **(b)** Obfuscated Image

Figure 1. BFW Dataset



(a) Source **(b)** Obfuscated Image

Figure 2. DemogPairs Dataset

(a) Source **(b)** Obfuscated Image

Figure 3. RFW Dataset

For the *BFW dataset*[9], we have evaluated the algorithm's performance using three different choices of parameters.The confidence scores obtained from *Face*$^{++}$[7] are in the range as follows:

	indian_females	white_males
Max Confidence	85.902	92.597
Min Confidence	82.452	92.597

Table 1. Choice 1

	indian_females	asian_females
Max Confidence	48.188	76.586
Min Confidence	48.188	41.884

Table 2. Choice 2

	indian_females	white_males
Max Confidence	49.045	37.101
Min Confidence	49.045	37.101

Table 3. Choice 3

For the *RFW Dataset*[11], we have evaluated the algorithm's performance using a single choice of parameters. The confidence scores obtained from *Face*$^{++}$[7] are in the range as follows:

	african	asian	indian	caucasian
Max Confidence	90.698	87.691	88.762	90.126
Min Confidence	36.994	34.839	38.468	36.354

Table 4. Choice 1

	african	asian	indian	caucasian
Max Confidence	87.653	84.062	87.744	86.43
Min Confidence	32.343	31.667	33.576	28.161

Table 5. Choice 2

	african	asian	indian	caucasian
Max Confidence	81.981	76.24	80.414	81.816
Min Confidence	17.068	12.328	15.499	14.872

Table 6. Choice 3

For the *DemogPairs dataset*[5], we have evaluated the algorithm's performance using three different choices of parameters.The confidence scores obtained from *Face*$^{++}$[7] are in the range as follows:

	Max Confidence	Min Confidence
Asian_female	96.446	38.154
Asian_male	96.252	40.782
Black_female	96.552	46.753
Black_male	96.363	52.063
White_female	94.608	72.196
White_male	95.771	65.229

Table 7. Choice 1

	Max Confidence	Min Confidence
Asian_female	86.626	29.193
Asian_male	89.413	34.545
Black_female	87.832	40.715
Black_male	92.065	42.909
White_female	78.41	70.942
White_male	79.561	64.353

Table 8. Choice 2

	Max Confidence	Min Confidence
Asian_female	74.546	9.685
Asian_male	76.221	12.499
Black_female	75.66	24.063
Black_male	74.908	27.542
White_female	58.054	36.636
White_male	61.193	37.251

Table 9. Choice 3

Refering to the **Image 4** which shows the obfuscation success rate, it takes a sigmoid curve for *Asian_Females, Asian_Males, Black_Females*, and *Black_Males* suggesting that the obfuscation technique is initially not very effective in reducing the accuracy of the facial recognition system, but becomes increasingly effective as the confidence threshold increases. This could be due to the fact that the facial recognition system is initially highly confident in its identifications and therefore more resistant to the obfuscation technique, but becomes less confident as the threshold is raised and therefore more susceptible to obfuscation.
The obfuscation success rate takes a stepwise curve for white individuals (*White_Females, White_Males*), it suggests that the obfuscation technique is effective at reducing the accuracy of the facial recognition system at certain thresholds, but less effective at others. This could be due to differences in the facial features of White individuals compared to other groups, or differences in the way that the obfuscation technique affects different types of faces.
DemogPairs dataset for different photos varies depending on the demographic group. The sigmoid curve for Asian and Black individuals suggests that the obfuscation method used is more effective in hiding their identity as the confidence threshold increases. However, for White individuals, the obfuscation success rate appears to increase in a stepwise manner. This suggests that the method may not be as effective in hiding the identity of White individuals, as the obfuscation success rate increases only after a certain threshold is reached. Overall, this information may indicate that the effectiveness of the obfuscation method may vary depending on the demographic group of the individuals being obfuscated.

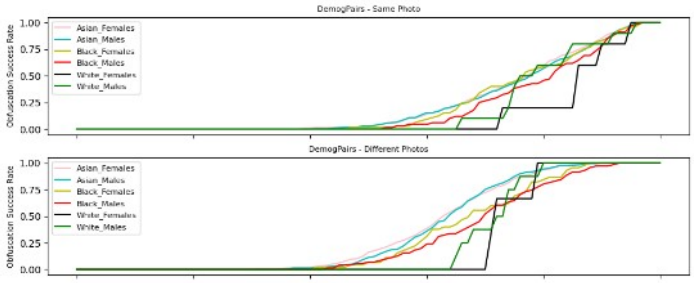


Figure 4. BFW vs DemogPairs : Obfuscation Success Rate

In the case of the *DemogPairs dataset*[5], where different photos are used for each demographic group, the FNR vs Confidence Threshold plot (referring to Image 5) shows a sigmoid curve for all groups, indicating that the system's ability to recognize individuals after obfuscation improves as the confidence threshold increases. This means that the system is more accurate in identifying individuals as the confidence threshold is raised. Furthermore, the fact that the obfuscation success rate is low or almost nil until a certain threshold (60% in this case) suggests that the facial recognition system is not able to accurately identify individuals when the obfuscation level is high. However, as the obfuscation level is reduced (i.e., confidence threshold is increased), the system's ability to recognize individuals improves and the FNR decreases. Overall, this plot provides valuable insights into the performance of the facial recognition system and how well it can identify individuals after obfuscation, as well as how the system's accuracy is affected by changes in the confidence threshold.
This graph indicates that for all the different photos of the same demographic group, the model's True False Rate (TFR) varies as a function of the confidence threshold. The TFR is a measure of how often the model incorrectly classifies an image (false positive) as well as how often it correctly classifies an image (true negative). For Asian females, Asian males, Black females, and Black males, the TFR starts off low and then increases as the confidence threshold increases. This indicates that at lower confidence thresholds, the model is making more errors (false positives and true negatives) than at higher confidence thresholds. The sigmoid curve suggests

that there is a threshold beyond which the model is making fewer errors and becomes more confident in its predictions. For White females and White males, the TFR starts off low and then jumps up at a certain confidence threshold before leveling off. This suggests that the model is more confident in its predictions for this demographic group, but there is a threshold at which it becomes less accurate in its predictions. The stepwise curve indicates that the TFR remains relatively constant once the threshold is passed.

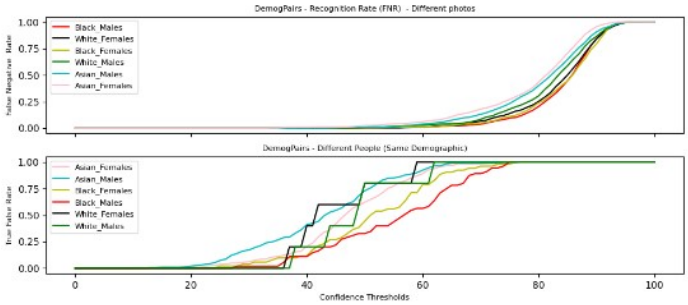


Figure 5. DemogPairs

Referring to **Image 6**, the obfuscation success rate on the y-axis and confidence threshold on the x-axis for the *BFW dataset*[9] for the Same Photo setting indicates how effective the obfuscation method is in concealing the identities of the individuals in the photos at different levels of confidence. In this case, for most demographic groups, the obfuscation method does not seem to be effective in concealing their identities, as the obfuscation success rate is nil for all confidence thresholds, indicating that the identities can be easily recognized. However, for Indian Females and White Males, the obfuscation method seems to be effective, as there is a sudden growth in obfuscation success rate at around 82% and 90% confidence threshold, respectively. This indicates that the method is successful in concealing their identities at these higher confidence thresholds.

The obfuscation success rate on the y-axis and the confidence threshold on the x-axis for the BFW dataset for different photos of the same demographic group show how effective the obfuscation technique is at different confidence thresholds. For Indian Females, the sudden growth in obfuscation success rate at around 50% confidence threshold suggests that the obfuscation technique works well for this group when the confidence level of the face recognition algorithm is around 50%. This means that the obfuscation technique is effective in hiding the identity of Indian Females from face recognition algorithms that have confidence levels around 50%. For Asian Females, the stepwise curve at around 40% confidence threshold suggests that the obfuscation technique has an initial effect on the algorithm's ability to recognize this demographic group's faces. However, this effect remains constant at 25% obfuscation success rate until around 78% confidence threshold, after which it takes another step and reaches maximum obfuscation success rate. This indicates that the obfuscation technique has a limited initial effect on the algorithm's ability to recognize Asian Females' faces, which improves only when the confidence level of the face recognition algorithm reaches around 78%.

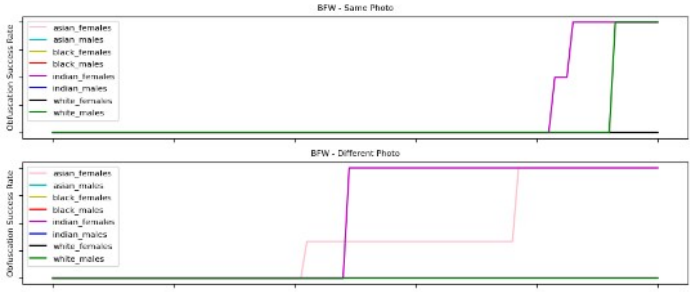


Figure 6. BFW

In the *BFW dataset*[9] for Different Photos, all demographic groups (*Asian_Females, Asian_Males, Black_Females, Black_Males, White_Females, and White_Males*) show a similar trend in terms of FNR and confidence threshold as seen in **Image 7**. At low confidence thresholds (below 50%), the obfuscation success rate is very low or almost nil, which means that the model is not able to correctly identify the faces in the dataset. However, as the confidence threshold increases, the obfuscation success rate gradually increases and follows a sigmoid curve, indicating that the model becomes more accurate in identifying faces as the level of certainty required for a prediction increases. This trend suggests that the model's performance improves as the confidence threshold increases.

In the case of the *BFW dataset*[9] for different photos of the same demographic, the graphs show that for most demographic groups, the true false rate is low or almost nil for all confidence thresholds. This suggests that the algorithm has difficulty accurately identifying individuals from these groups in different photos of the same demographic. However, for White Males and Indian Females, the true false rate suddenly increases and reaches the highest confidence threshold at around 38% and 50% confidence threshold respectively. This suggests that the algorithm is able to more accurately identify individuals from these groups in different photos of the same demographic compared to other groups in the dataset. It is important to note that the reasons behind this difference in performance could be due to various factors such as lighting, pose, or the level of variability within the demographic group.

In the context of Attack 1 on the *BFW dataset*[9] (*first plot of Image 8*), the plot shows that for both female and male photos, the obfuscation success rate is very low or almost nil

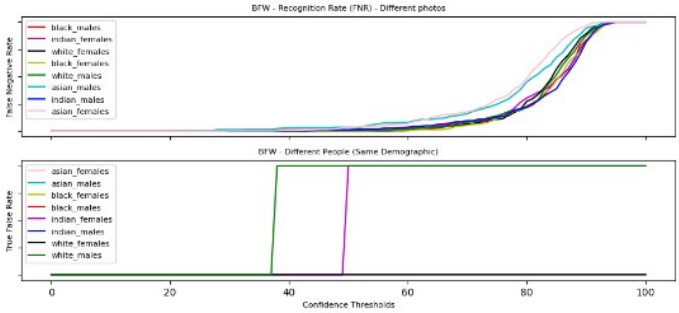


Figure 7. BFW - TFR - FNR

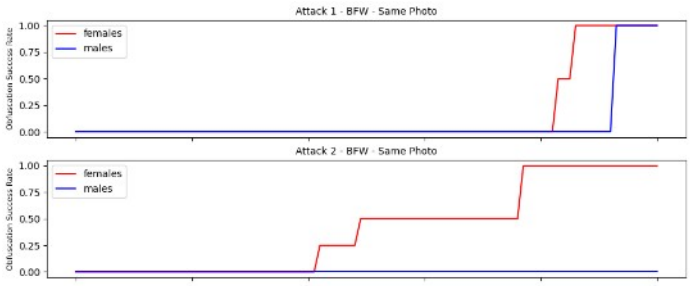


Figure 8. Attack - BFW - Same Photo

until the confidence threshold reaches a certain level (82% for females and 90% for males). This indicates that the generated faces were not successful in hiding the identity of the person until a certain level of confidence was reached. However, after this threshold, there is a sudden growth in the obfuscation success rate, which means that the generated faces were able to successfully hide the identity of the person from the recognition system. This sudden growth in the success rate may be due to the generated faces becoming more similar to the original faces as the confidence threshold increases.

In Attack 2 - BFW - Same photo(*second plot of Image 8*), the obfuscation success rate is plotted against different confidence thresholds for the same photo attack on the BFW dataset for both males and females separately. For males in the BFW dataset, the obfuscation success rate is almost nil across all confidence thresholds, which means that the facial recognition algorithm is highly accurate in recognizing the male faces in the dataset, and it is difficult to fool the system with the same photo attack. For females in the BFW dataset, the obfuscation success rate is nil or very low until 40% confidence threshold, indicating that the facial recognition algorithm is still highly accurate in recognizing the female faces in the dataset. However, after the 40% confidence threshold, the obfuscation success rate suddenly increases, and the graph takes a stepwise curve, reaching the maximum obfuscation success rate. This means that the facial recognition algorithm is less accurate in recognizing the female faces with low confidence scores, and the algorithm can be fooled with the same photo attack by generating low confidence scores for female faces. The reason for this difference between males and females could be due to the different facial features, expressions, and lighting conditions present in the photos of the dataset.

In the case of the Recognition- BFW - Different Photo attack(*first plot of Image 9*), we see that both the Male and Female datasets have a very low FNR until a confidence threshold of 60%, after which there is a sharp increase in the FNR as the confidence threshold increases. This means that the algorithm is able to correctly identify matches with a high degree of accuracy until a certain confidence threshold is reached. After this threshold, the algorithm begins to make more false negative identifications, which means that it is becoming less accurate. The reason for this trend could be that as the confidence threshold increases, the algorithm becomes more conservative and less likely to make positive identifications. This may result in some matches being missed, leading to an increase in the FNR. Another possibility is that as the confidence threshold increases, the algorithm becomes more susceptible to errors due to image noise, changes in lighting conditions, or other factors that may affect the quality of the input photo.

In Attack 3 - BFW - Different Faces - Same Demo(*second plot of Image 9*), the true false rate is plotted against the confidence threshold for the attack where different faces of the same demographic group are used. The dataset includes females and males. The true false rate represents the percentage of times when the attacker can correctly identify a different photo of the same individual as belonging to that individual. Therefore, a lower true false rate indicates a higher obfuscation success rate.In this plot, we can see that both females and males show a very low or nil obfuscation success rate until a certain confidence threshold is reached. For females, this threshold is around 48%, while for males, it is around 38%. After reaching this threshold, both females and males show a sudden growth in the obfuscation success rate, reaching the maximum value at the highest confidence threshold. The plot of obfuscation success rate on the y-axis against confidence threshold on the x-axis for Attack 1 - DemogPairsPairs - Same Photo (*first plot of Image 10*), for both Females and Males, shows the success rate of an obfuscation attack on the DemogPairsPairs dataset[5] when using the same photo as the original. The sigmoid curve indicates that at lower confidence thresholds, the attack is not very successful, i.e., the obfuscated images are still being recognized as the original face with a high probability. However, as the confidence threshold increases, the obfuscation success rate

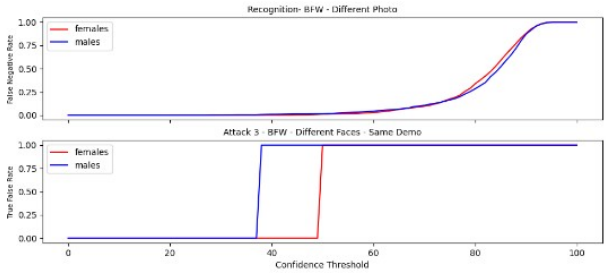


Figure 9. BFW - Different Photo

also increases. This suggests that the obfuscated images become less recognizable as the original face as the confidence threshold is increased.

In Attack 2 - DemogPairsPairs - Same Photo (*second plot of Image 10*), the obfuscation success rate is plotted against confidence thresholds for the same photo attack on the DemogPairsPairs datase[5]t, both Females and Males show almost no obfuscation success until 40% confidence threshold. This suggests that the models are not able to successfully obfuscate the images at lower confidence thresholds. However, after 40% confidence threshold, the obfuscation success rate shows a sigmoid curve and reaches maximum obfuscation success rate. This means that the models are able to obfuscate the images effectively when the confidence threshold is increased beyond a certain value. The reason behind this behavior can be attributed to the fact that the models used for the attack are more confident in their predictions when the confidence threshold is increased beyond a certain value. At lower confidence thresholds, the models are not very certain about their predictions and hence are not able to effectively obfuscate the images. However, when the confidence threshold is increased beyond a certain value, the models become more certain about their predictions and are able to successfully obfuscate the images.

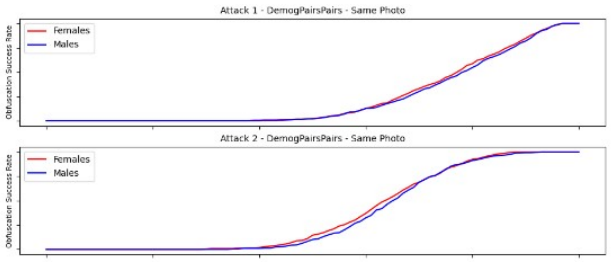


Figure 10. Attack - DemogPairs

Referring to the *first plot Image 11* the False negative rate vs confidence threshold plot for Recognition - DemogPairsPairs - Different Photo shows how a face recognition system performs when an attacker has access to different photos of the same individual from two different demographic groups, in this case, females and males. The plot shows that for both females and males, the false negative rate is low or almost zero until the confidence threshold reaches 60%, beyond which it starts to increase. The system is capable of accurately recognizing individuals up to a certain threshold, after which its performance begins to degrade. The sudden rise in the false negative rate at higher confidence thresholds could be due to the system becoming more cautious in its decision-making and requiring a greater degree of confidence to make a positive match. This can result in the rejection of genuine matches and an increase in the false negative rate.

The plot of True false rate on y-axis and confidence threshold on x-axis for Attack 3 - DemogPairsPairs - Different Faces - Same Demo (*second plot of Image 11*)represents the performance of an obfuscation attack on a face recognition system using a dataset that includes individuals from different demographic groups with different photos. The true false rate represents the fraction of positive matches that are incorrectly identified as negative matches by the system. A low true false rate is desirable as it implies that the system can correctly identify positive matches even when presented with different photos and demographic groups. The plot shows that for both females and males, the true false rate is very low or almost nil until a confidence threshold of 18%, after which it starts to increase and reaches a maximum value at around 80% confidence threshold. This means that the obfuscation attack is successful in reducing the recognition accuracy of the system, as it is able to produce false negatives at a high rate once the confidence threshold is crossed.

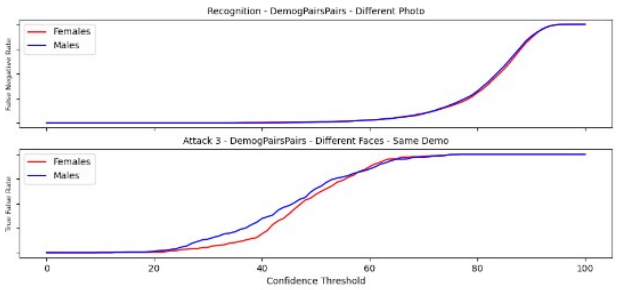


Figure 11. DemogPairs - Different Photo

The plot of obfuscation success rate on the y-axis and confidence threshold on the x-axis(*first plot of Image 12*) for the RFW (Racial Faces in-the-Wild) dataset[11] with different demographic groups represents the effectiveness of the obfuscation attack on the face recognition system. The plot shows that for all the different demographic groups in

the dataset (African, Asian, Caucasian, and Indian), the obfuscation success rate is very low or nil until a confidence threshold of 40%. After this threshold, the success rate starts to increase and eventually reaches its maximum value. This means that the attacker can successfully obfuscate the facial features of the individuals in the dataset and prevent the face recognition system from recognizing them with high accuracy. The sudden increase in the obfuscation success rate at higher confidence thresholds may be due to the fact that the perturbation applied to the faces becomes more effective and starts to significantly alter the facial features. As a result, the face recognition system becomes less reliable and less accurate in identifying individuals.

Referring to the plot of obfuscation success rate on y-axis and confidence threshold on x-axis for the dataset RFW - Different Demo (*Image 12, Plot 2*) represents the performance of the obfuscation attack on a face recognition system trained on a dataset that includes faces from four different demographic groups: African, Asian, Caucasian, and Indian. In this case, the attack is performed by generating synthetic faces from different demographic groups to obfuscate the original face and prevent the system from correctly recognizing the individual. The obfuscation success rate represents the fraction of synthetic faces generated by the attacker that can successfully fool the face recognition system into making an incorrect match. A high obfuscation success rate indicates that the attacker is able to effectively obfuscate the original face and prevent the system from correctly recognizing the individual. The plot shows that for all the demographic groups, the obfuscation success rate is very low or almost nil until a confidence threshold of 40%, after which it starts to increase and reaches a maximum value at around 80% confidence threshold. This means that the obfuscation attack is not very effective until a certain threshold, beyond which it starts to become more successful in fooling the face recognition system. The sudden increase in the obfuscation success rate at higher confidence thresholds may be due to the fact that the system becomes more confident in its decision-making and is more likely to make incorrect matches when presented with synthetic faces that are designed to look similar to the original face but belong to a different demographic group.

The plot of False negative rate on y-axis and confidence threshold on x-axis for the dataset Recognition - RFW - Different Photos(*first plot of Image 13*) represents the performance of the face recognition system on a dataset where the attacker has access to different photos of the same individual from four different demographic groups (African, Asian, Caucasian, Indian). The plot shows that for all four demographic groups, the false negative rate is very low or almost nil until a confidence threshold of 60%, after which it starts to increase and reaches a maximum value at around 90% confidence threshold. This means that the system is able to correctly recognize the individuals with a high accuracy

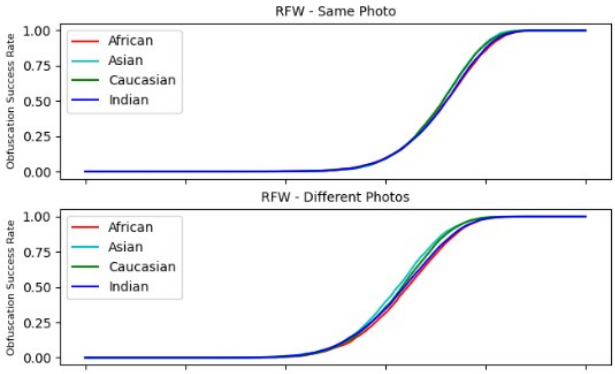


Figure 12. RFW - Obfuscation Success Rate

until a certain threshold, beyond which the performance starts to degrade. The sudden increase in the false negative rate at higher confidence thresholds may be due to the fact that the system becomes more conservative in its decision-making and requires a higher degree of confidence to make a positive match. This can lead to rejecting genuine matches and increasing the false negative rate.

The plot of True false rate on y-axis and confidence threshold on x-axis for the dataset RFW - Different Faces - Same DemogPairs(*second plot of Image 13*) represents the performance of an attack on a face recognition system using different photos of individuals belonging to the same demographic group. In this case, the dataset includes individuals from four different demographic groups: African, Asian, Caucasian, and Indian.The plot shows that for all the demographic groups, the true false rate is very low or almost nil until a confidence threshold of 20%, after which it starts to increase and reaches a maximum value at around 80% confidence threshold. This means that the attack is not very successful until a certain threshold, beyond which the performance starts to degrade. The sudden increase in the true false rate at higher confidence thresholds may be due to the fact that the system becomes less conservative in its decision-making and is more likely to accept fake identities.

Referring to the Obfuscation success rate versus confidence threshold plot [*Plot 1 , Image 14*] for the BFW-Same Photo dataset represents the performance of an obfuscation method on a dataset of individuals from different demographic groups. The plot shows that the obfuscation method has a low success rate at low confidence thresholds for all demographic groups. As the confidence threshold increases, the success rate also increases, indicating that the obfuscation method becomes more effective at higher confidence thresholds. However, the degree of effectiveness varies depending on the demographic group. For Indian individuals, the obfuscation method shows a sudden increase in success rate at around 80% confidence threshold, indicating that the

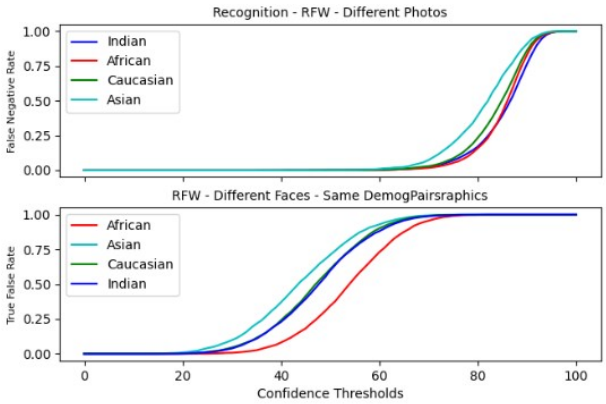


Figure 13. RFW - Different Faces

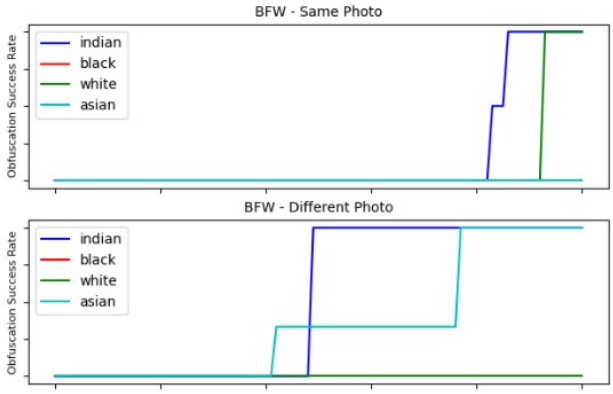


Figure 14. BFW - Obfuscation Success Rate

method is more effective in obfuscating Indian individuals compared to other demographic groups. This could be due to the specific facial features or characteristics of Indian individuals that make them more difficult to recognize or distinguish. For White individuals, the plot shows a sudden growth in the obfuscation success rate at around 90% confidence threshold, indicating that the method is more effective in obfuscating White individuals at that point. This could be due to the fact that the facial features or characteristics of White individuals are more distinct and easier to recognize compared to other demographic groups.

The plot of Obfuscation success rate on y-axis and confidence threshold on x-axis [*Plot 2 , Image 14*]for the dataset BFW - Different Photo represents the performance of the obfuscation method on a dataset where the attacker has access to different photos of an individual from different demographic groups. In this case, the dataset includes Indian, Black, White, and Asian individuals. For Indian individuals, the plot shows a sudden growth in obfuscation success rate at around 50% confidence threshold and a stepwise curve at around 80% confidence threshold, indicating that the obfuscation method is more effective in obfuscating Indian individuals at those thresholds. This could be due to the specific facial features or characteristics of Indian individuals that make them more difficult to recognize or distinguish. For Asian individuals, the plot shows a stepwise curve at around 40% confidence threshold, indicating that the obfuscation method is more effective in obfuscating Asian individuals at that point. The success rate remains constant at 25% until it takes another step at around 78% confidence threshold and reaches the maximum obfuscation success rate. This could be due to the fact that the facial features or characteristics of Asian individuals are different from other demographic groups, making them more difficult to recognize or distinguish.

The plot of False Negative rate on y-axis and confidence threshold on x-axis for the dataset Recognition - BFW - Different Photos, shown in *Plot 1 , Image 15* represents the performance of a face recognition system on a dataset where the individuals are from different demographic groups, including Indian, Black, White, and Asian. The plot shows that for all demographic groups, the false negative rate is very low or nil for confidence thresholds below 40%. This means that the face recognition system is able to correctly identify the individuals from different demographic groups with high accuracy at those confidence thresholds. However, as the confidence threshold increases beyond 40%, the false negative rate starts to increase gradually, indicating that the face recognition system becomes less accurate in identifying the individuals from different demographic groups. The reason for this increase in false negative rate at higher confidence thresholds could be due to the fact that the face recognition system is less certain about the identity of the individuals, especially for individuals from different demographic groups whose facial features and characteristics may be different from those of the individuals in the training set. As a result, the face recognition system may be more likely to incorrectly reject the identity of individuals from different demographic groups at higher confidence thresholds, leading to a higher false negative rate.

The plot of True false rate on y-axis and confidence threshold on x-axis for the dataset BFW - Different Faces - Same DemogPairs [*Plot 2 , Image 15*]represents the performance of a face recognition system on a dataset where the faces are from different photos but belong to the same demographic group. The dataset includes Indian, Black, White, and Asian individuals. For White and Indian individuals, the plot shows a sudden growth in the true false rate at around 38% and 50% confidence thresholds, respectively. This indicates that the face recognition system is less accurate in correctly identifying these individuals at those confidence thresholds. This

could be due to the fact that the facial features or characteristics of White and Indian individuals are more similar or less distinct, making it more difficult for the system to differentiate between them. For all other demographic groups, the plot shows a nil value for all confidence thresholds, which means that the face recognition system is not able to identify them correctly. This could be due to the fact that the faces are from different photos but belong to the same demographic group, making them more similar and harder to distinguish. Overall, the plot suggests that the face recognition system has a lower accuracy in correctly identifying individuals from different photos but belonging to the same demographic group, especially for White and Indian individuals.

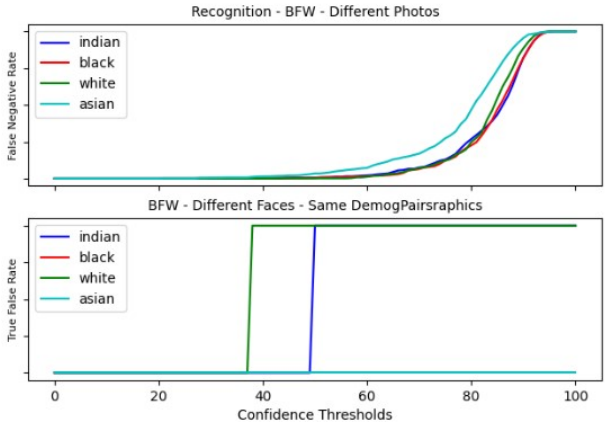


Figure 15. BFW - Different Faces

Referring to the **Image 16**, the plot of obfuscation success rate on y-axis and confidence threshold on x-axis for DemogPairs - Same Photo represents the effectiveness of an obfuscation method on a dataset where the attacker has access to the same photo of individuals from different demographic groups. In this case, the dataset includes Black, White, and Asian individuals. The plot shows that for Asian and Black individuals, the obfuscation success rate increases gradually and reaches a maximum value at around 40% confidence threshold, as indicated by the sigmoid curve. This means that the obfuscation method is moderately effective in reducing the recognizability of faces from these two demographic groups at that threshold. The reason for the gradual increase in the success rate could be due to the fact that the facial features or characteristics of Asian and Black individuals are less distinct or easier to confuse than those of other groups, making it easier to obfuscate their faces. For White individuals, the plot shows a stepwise curve at around 40% confidence threshold, where the obfuscation success rate suddenly increases and reaches the maximum value. This indicates that the obfuscation method is highly effective in reducing the recognizability of White individuals at that threshold. The reason for the sudden increase in the success rate could be due to the fact that the facial features or characteristics of White individuals are more distinct or easier to recognize than those of other groups, making it more challenging to obfuscate their faces.

The plot of obfuscation success rate on the y-axis and confidence threshold on the x-axis for the DemogPairs - Different Photo dataset, in **Image 16** shows the performance of an obfuscation method on a dataset where the attacker has access to different photos of individuals from the same demographic group. The dataset includes Black, White, and Asian individuals. For Asian and Black individuals, the plot shows a sigmoid curve at around 30% confidence threshold, indicating that the obfuscation method has a gradual increase in success rate at that point. This means that the method is somewhat effective in obfuscating these demographic groups. However, the effectiveness of the obfuscation method is still relatively low for these groups, as the success rate does not reach its maximum. For White individuals, the plot shows a stepwise curve at 60% confidence threshold, indicating that the obfuscation method has a sudden increase in success rate at that point. This means that the method is more effective in obfuscating White individuals compared to the other demographic groups. The reason for this could be due to the specific facial features or characteristics of White individuals that make them more distinct and easier to recognize, which require a higher degree of obfuscation to make them less recognizable.

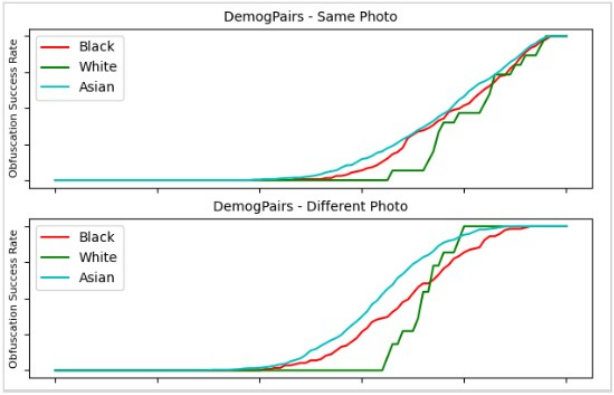


Figure 16. DemogPairs - Obfuscation Success Rate

The False Negative (FN) rate on y-axis and confidence threshold on x-axis for the Recognition - DemogPairs - Different Photos dataset in the **Image 17** shows the performance of a face recognition system when recognizing individuals from different demographic groups (Black, White, and Asian) with different photos. The plot shows that the FN rate is low or nil at low confidence thresholds for all demographic groups, which means that the face recognition system is good at

correctly recognizing individuals at those thresholds. However, as the confidence threshold increases, the FN rate also increases, indicating that the face recognition system starts making more false negatives. The sudden increase in the FN rate at around 50% confidence threshold for all demographic groups suggests that the face recognition system has difficulty in recognizing individuals correctly when the confidence threshold is high. The sigmoid curve shows that as the confidence threshold increases, the FN rate increases rapidly, which means that there is a higher chance of the face recognition system making a false negative. This could be due to various reasons such as illumination changes, facial expressions, occlusions, or other factors that affect the face recognition performance.

Referring to the plot of True False rate on y-axis and confidence threshold on x-axis in **Image** 17 for the DemogPairs - Different Faces dataset shows the performance of a face recognition system on images of individuals from different demographic groups. The dataset includes Black, White, and Asian individuals. For Asian and Black individuals, the plot shows a sigmoid curve at around 18% threshold, indicating that the system has a gradual increase in the True False rate at that point. This means that the system becomes less accurate in recognizing Asian and Black individuals as the confidence threshold decreases. The reason for this could be due to the specific facial features or characteristics of these individuals that are more difficult to distinguish or recognize. For White individuals, the plot shows a stepwise curve at around 30% threshold, indicating that the system is more accurate in recognizing White individuals at that point. This means that the system is more effective in distinguishing White individuals from other demographic groups compared to Asian and Black individuals. The reason for this could be due to the fact that the facial features or characteristics of White individuals are more distinct and easier to recognize compared to other demographic groups.

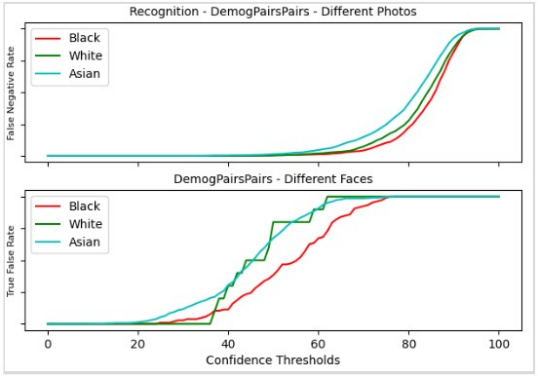
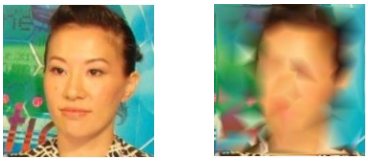


Figure 17. DemogPairs - Different Photos

7.1.2 DP-SAMP. In this section, we present the results obtained from DP-SAMP [10] based on the original and obfuscated image samples for each dataset. The first image is the input image and second is the resulting output image generated from the method as shown in **Algorithm 2**. The parameters used for the method to yeild the obfuscated images from the original source images are: *Privacy Budget (δ) = 25*, *Number of Pixels (m) = 12* and *Number of Clusters (k) = 24*

(a) Source **(b)** Obfuscated Image

Figure 18. BFW Dataset



(a) Source **(b)** Obfuscated Image

Figure 19. DemogPairs Dataset

(a) Source **(b)** Obfuscated Image

Figure 20. RFW Dataset

For the *BFW dataset*[9], we have evaluated the algorithm's performance using three different choices of parameters.The confidence scores obtained from *Face*$^{++}$[7] are in the range as follows:

	Max Confidence	Min Confidence
Asian_females	84.426	33.881
Asian_males	84.673	24.402
Black_females	87.055	37.483
Black_males	85.078	36.077
Indian_females	84.127	21.741
Indian_males	85.166	40.721
White_females	84.557	34.941
White_males	86.449	37.117

Table 10. BFW - Choice 1

	Max Confidence	Min Confidence
Asian_females	80.576	26.559
Asian_males	78.889	31.225
Black_females	78.263	31.457
Black_males	79.3	32.907
Indian_females	81.783	30.414
Indian_males	84.204	28.942
White_females	79.735	38.276
White_males	77.571	34.306

Table 11. BFW - Choice 2

	Max Confidence	Min Confidence
Asian_females	77.946	19.229
Asian_males	79.961	17.648
Black_females	78.115	28.795
Black_males	82.696	33.36
Indian_females	79.212	25.345
Indian_males	78.406	24.958
White_females	75.626	27.398
White_males	77.468	23.326

Table 12. BFW - Choice 3

For the *RFW Dataset*[11], we have evaluated the algorithm's performance using a single choice of parameters. The confidence scores obtained from *Face*$^{++}$[7] are in the range as follows:

	african	asian	indian	caucasian
Max Confidence	79.769	83.74	83.763	83.195
Min Confidence	28.171	28.154	29.675	22.781

Table 13. RFW - Choice 1

	african	asian	indian	caucasian
Max Confidence	79.804	80.72	80.224	76.979
Min Confidence	27.174	25.733	26.15	20.0197

Table 14. RFW - Choice 2

	african	asian	indian	caucasian
Max Confidence	79.697	79.011	85.008	78.164
Min Confidence	30.163	20.507	24.042	20.172

Table 15. RFW - Choice 3

For the *DemogPairs dataset*[5], we have evaluated the algorithm's performance using three different choices of parameters.The confidence scores obtained from *Face*$^{++}$[7] are in the range as follows:

The *first plot* of **Image 21** shows the performance of an obfuscation method on a dataset of different demographic groups and genders where the attacker has access to the same photo of an individual from different demographic pairs. The plot measures the obfuscation success rate on the y-axis and the confidence threshold on the x-axis. The sigmoid curve indicates that the obfuscation method becomes more effective as the confidence threshold increases. The sudden growth in obfuscation success rate at around 40% threshold indicates that the method is more effective in obfuscating the faces of the individuals in the dataset. This means that the obfuscation method can make it difficult for an attacker to recognize and identify individuals from the dataset.
Referring to the second plot of Obfuscation success rate on the y-axis and confidence threshold on the x-axis in **Image ??** for DemogPairs - Different Photo dataset shows the performance of an obfuscation method on a dataset where the attacker has access to different photos of individuals from different demographic groups. The dataset includes Asian and Black females and males, as well as White females and males. The plot shows that the obfuscation method has a low success rate at low confidence thresholds, indicating that the method is not effective in obfuscating faces at those thresholds. As the confidence threshold increases, the success rate also increases, indicating that the obfuscation method becomes more effective at higher confidence thresholds. The sigmoid curve observed in all demographic groups indicates that the obfuscation method becomes increasingly effective

	Max Confidence	Min Confidence
Asian_female	75.565	40.409
Asian_male	75.876	41.082
Black_female	79.176	32.217
Black_male	78.168	34.23
White_female	83.765	36.874
White_male	81.988	42.157

Table 16. Demog - Choice 1

	Max Confidence	Min Confidence
Asian_female	70.56	30.679
Asian_male	70.774	38.777
Black_female	75.127	35.13
Black_male	74.492	36.957
White_female	78.584	32.816
White_male	77.838	28.801

Table 17. Demog - Choice 2

	Max Confidence	Min Confidence
Asian_female	71.171	32.756
Asian_male	75.984	23.989
Black_female	79.62	30.907
Black_male	76.306	31.236
White_female	76.863	29.271
White_male	81.77	29.943

Table 18. Demog - Choice 3

as the confidence threshold approaches a certain value, after which the success rate plateaus. This could be due to the fact that the method becomes more effective at obfuscating facial features as the confidence threshold increases

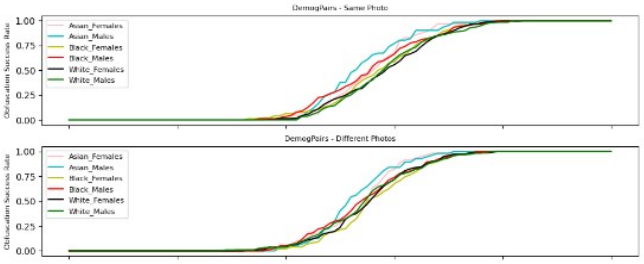


Figure 21. DemogPairs - Obfuscation Rate

The false negative rate on y-axis and confidence threshold on x-axis in **Image 22 Plot 1** for the DemogPairs - Recognition Rate (FNR) - Different Photos dataset indicates the performance of a face recognition system in identifying individuals of different demographics when they appear in different photographs. In this dataset, the performance of the system is initially low, with a negligible or very low recognition rate for all demographic groups at low confidence thresholds. As the confidence threshold increases, the system begins to recognize more individuals correctly, and the false negative rate decreases. The sigmoid curve indicates that there is a critical point in the confidence threshold beyond which the false negative rate increases rapidly. This suggests that the system is more likely to make errors when identifying individuals from different demographics as the confidence threshold decreases. Overall, the dataset highlights the need for face recognition systems to be tested and evaluated across diverse demographic groups and different photographic conditions to ensure equitable performance.
The True False rate (TFR) in **Image 22 Plot 2** represents the percentage of times the system makes a correct identification and an incorrect identification simultaneously. In the context of the DemogPairs - Different People (Same Demographic) dataset. The sigmoid curve in this dataset indicates that as the confidence threshold increases, the TFR also increases, reaching its maximum at around 30% confidence threshold. This means that at lower confidence thresholds, the system is able to correctly identify people from the same demographic with a very low TFR. However, as the confidence threshold increases, the system starts making more false positive identifications, resulting in a higher TFR. The reason for the sigmoid shape of the curve is because the system becomes increasingly conservative as the confidence threshold increases, resulting in fewer identifications overall. At a certain point, the system becomes so conservative that it starts to miss correct identifications, resulting in an increase in false negatives and a higher TFR.

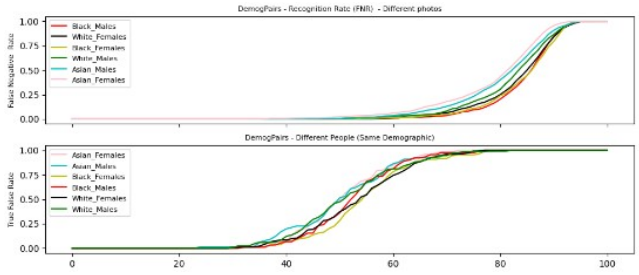


Figure 22. DemogPairs - Different Photos

The plot of **Figure 23** shows the relationship between obfuscation success rate on the y-axis and the confidence threshold on the x-axis for the BFW-Same Photo dataset. The plot suggests that for all demographic groups (Asian females, Asian males, Black females, Black males, Indian females, Indian males, White females, and White males), the model's confidence in correctly identifying the individual decreases as the confidence threshold increases from low values. However, there is a sudden growth in the obfuscation success

rate at around 30% confidence threshold, which indicates that obfuscation is more effective at fooling the model as the confidence threshold increases. This means that obfuscation techniques are more effective at preventing facial recognition algorithms from accurately recognizing individuals at lower confidence thresholds. At higher confidence thresholds, the model is more confident in its predictions, and obfuscation is less effective.

The graph of obfuscation success rate on the y-axis and confidence threshold on the x-axis for *BFW (Different Photo)* dataset indicates the performance of the face recognition system in terms of the ability to obfuscate (or hide) the identity of individuals belonging to different demographics, including Asian, Black, Indian, and White of both genders. The sigmoid curve indicates that at low confidence thresholds, the system is unable to reliably identify individuals, but as the confidence threshold increases, the system's performance in terms of obfuscation success rate improves and approaches the maximum possible obfuscation success rate. The fact that all demographic groups show a similar sigmoid curve indicates that the face recognition system's performance in terms of obfuscation success rate is relatively consistent across different demographic groups. Overall, a higher obfuscation success rate implies that the face recognition system has a lower probability of correctly identifying an individual's identity, which can be useful for privacy and security purposes.

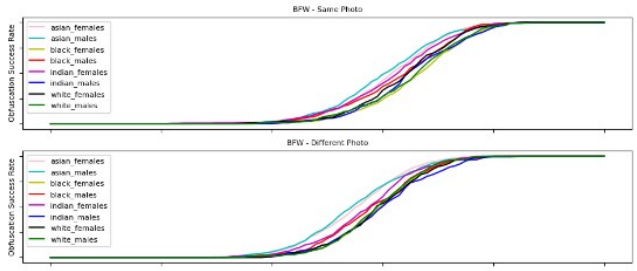


Figure 23. BFW - Obfuscation Success Rate

In the case of the BFW - Recognition Rate (FNR) - Different Photos dataset [**Figure 24**], we can see that for all demographic groups, the false negative rate is initially very low or nil at low confidence thresholds. However, as the confidence threshold increases, the false negative rate gradually increases, and the curve eventually reaches a maximum at some point. This means that when the model is very selective about recognizing faces, it tends to miss some faces (false negatives), which can be particularly problematic in security or surveillance applications. The sigmoid shape of the curve is likely due to the fact that the model's performance tends to improve rapidly as the confidence threshold increases up to a certain point, after which the improvement slows down and eventually reaches a plateau. The exact shape and location of the curve may depend on various factors such as the size and quality of the dataset, the complexity of the model, and the choice of the confidence threshold.

The plot of true false rate on the y-axis and confidence threshold on the x-axis for BFW - Different People - Same Demographics [**Figure 24**] shows the performance of a face recognition system when identifying individuals of the same demographic group. The dataset includes different demographic groups such as Asian, Black, Indian, and White, and different genders such as Females and Males. The plot shows that the face recognition system has a low true false rate for low confidence thresholds. This means that the system correctly identifies most of the individuals in the dataset. As the confidence threshold increases, the true false rate also increases, indicating that the system becomes less accurate in identifying individuals. The sigmoid curve observed in the plot indicates that the system performs well for most of the individuals in the dataset, but there are a few individuals for whom the system fails to recognize. As the confidence threshold increases, the system becomes more conservative and starts rejecting more faces. This leads to a higher true false rate for higher confidence thresholds. The low true false rate at low confidence thresholds can be attributed to the fact that the individuals in the dataset belong to the same demographic group. Since the system is trained to recognize faces of a particular demographic group, it performs well when the individuals in the dataset belong to the same group. However, as the confidence threshold increases, the system becomes more prone to making errors, which leads to a higher true false rate

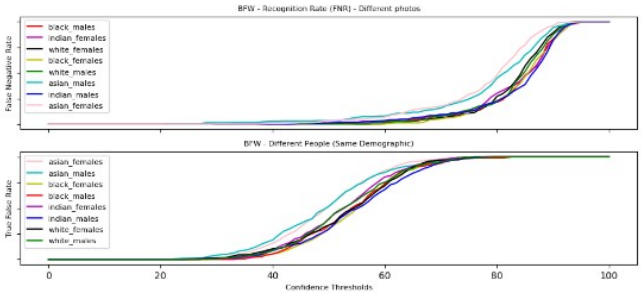


Figure 24. BFW - Different Photos

In the context of *Attack 1 - BFW - Same photo* [**Figure 25**], the obfuscation success rate on the y-axis represents the percentage of times an attacker was able to successfully obfuscate a given individual's identity. The sigmoid curve in both females and males indicates that as the confidence threshold increases, the obfuscation success rate also increases. This means that when the facial recognition model has a higher confidence score, it becomes more difficult for the attacker to obfuscate the individual's identity. At lower confidence

scores, however, the attacker has a higher success rate in obfuscating the individual's identity. The sudden increase in obfuscation success rate at around 40% threshold indicates that below this threshold, the facial recognition model may not have enough confidence in its identification to prevent obfuscation attacks. Beyond this threshold, the model becomes more certain of its identification and therefore more resistant to obfuscation attacks.

In the context of the graph showing obfuscation success rate on the y-axis and confidence threshold on the x-axis [**Figure 25**] for *Attack 2 on the BFW dataset with same photo pairs*, the sigmoid curve represents the effectiveness of the attack at different levels of confidence threshold. A low obfuscation success rate means that the attack is not able to successfully alter the facial recognition system's output, while a high success rate means that the attack is successful in obfuscating the identity of the individual. The fact that both females and males show nil or very low confidence thresholds until 40% indicates that the facial recognition system is not very confident in its identification until it reaches that threshold. The sigmoid curve that follows shows that the obfuscation attack becomes more effective as the confidence threshold increases, which means that the facial recognition system is more confident in its identification and the attack has to work harder to obfuscate the identity. The fact that both males and females show similar curves suggests that the attack is equally effective for both genders.

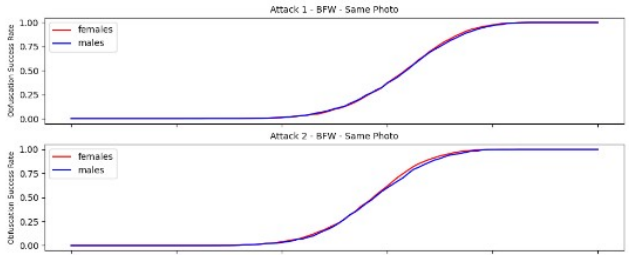


Figure 25. BFW - Attack - Same Photos

The plot of False Negative Rate (FNR) vs confidence threshold on the x-axis for the Recognition task on the BFW dataset with different photos [**Figure 26**] shows how well the model is able to correctly identify a person from their face image. In this case, both the female and male datasets show that the model has very low FNR at low confidence thresholds, which means the model is able to correctly identify most of the people in the dataset. However, as the confidence threshold increases, the FNR increases and reaches a maximum value at around 60%, which means the model is more likely to make false negative errors and fail to recognize people who are actually in the dataset. This behavior may be due to the fact that the model is becoming more conservative in its predictions as the confidence threshold increases, which leads to more false negatives. It could also be due to limitations in the dataset or the model's architecture, which may struggle with certain types of faces or image variations. The exact reason would require further investigation.

In the context of the dataset *Attack 3 - BFW - Different Faces - Same Demo* [**Figure 26**], the true false rate on the y-axis and confidence threshold on the x-axis represents the trade-off between correctly identifying a person and incorrectly identifying a different person of the same demographic group. At very low confidence thresholds, the true false rate is very low because the system is not making any confident decisions. As the confidence threshold increases, the true false rate also increases, indicating that the system is making more confident decisions. At around 30% threshold, the true false rate reaches its maximum value, indicating that the system is making a trade-off between correctly identifying a person and incorrectly identifying a different person of the same demographic group. Beyond this point, as the confidence threshold increases further, the true false rate begins to decrease, indicating that the system is becoming more conservative in its decisions, and is more likely to reject a person even if they are correctly identified.

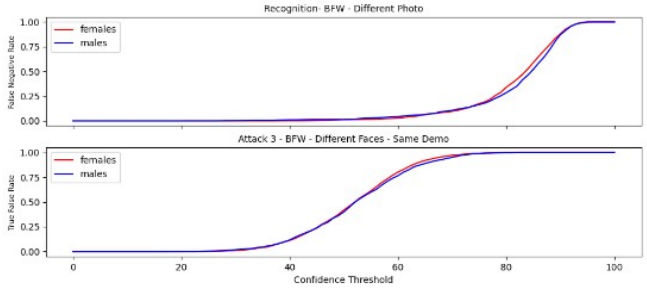


Figure 26. BFW - Different Photos

The plot of obfuscation success rate on the y-axis and confidence threshold on the x-axis for *Attack 1 - DemogPairsPairs - Same Photo [**Figure 27**]*, indicates the effectiveness of obfuscation method on the dataset. In this case, both females and males datasets show negligible obfuscation success rate when the confidence threshold is low, but as the confidence threshold increases, the obfuscation success rate increases gradually and reaches its maximum value. This indicates that the obfuscation method becomes more effective as the confidence threshold increases. The sigmoid curve represents the gradual increase in obfuscation success rate, which indicates that there is a correlation between confidence threshold and obfuscation success rate. The reason behind this behavior is that the obfuscation method relies on reducing the confidence level of the machine learning model in recognizing the identity of the person in the image. Therefore, when the confidence level of the model is low, the obfuscation method is more effective, and as the confidence level increases, the

effectiveness of the obfuscation method decreases.
The plot shows the obfuscation success rate on the y-axis and the confidence threshold on the x-axis for the *Attack 2 - DemogPairsPairs - Same Photo [**Figure** 27]* dataset. The dataset includes females and males. The sigmoid curve indicates the relationship between the confidence threshold and the obfuscation success rate. At lower confidence thresholds (less than 30%), the obfuscation success rate is low or nil, which means that the model is not able to obfuscate the identities of the individuals in the photos. As the confidence threshold increases, the obfuscation success rate increases as well, indicating that the model is becoming more effective at obfuscating the identities. Once the confidence threshold reaches a certain point, the obfuscation success rate reaches its maximum value, indicating that the model is now highly effective at obfuscating the identities of the individuals in the photos. The reason for the sigmoid curve is likely because the model's ability to obfuscate identities is dependent on the confidence threshold. When the model is less confident, it may be more likely to misidentify individuals and therefore less effective at obfuscating their identities. However, as the confidence threshold increases, the model becomes more accurate and is better able to obfuscate identities, which results in the sigmoid curve pattern.

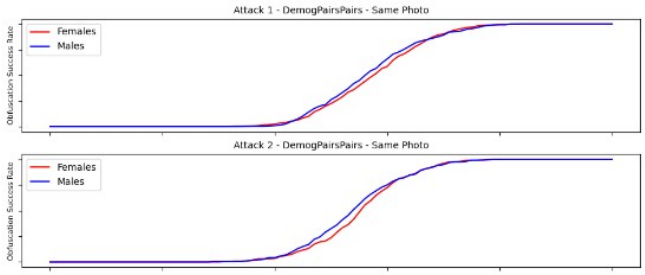


Figure 27. DemogPairs - Attack - Same Photos

In the *Recognition - DemogPairsPairs - Different Photo dataset [**Figure** 28]*, we have two groups - females and males - and the plot shows the False Negative Rate (FNR) on the y-axis and the confidence threshold on the x-axis. The low confidence threshold indicates that the face recognition system is more likely to make false positive identifications. However, as the confidence threshold increases, the system becomes more cautious, leading to a decrease in false positives and an increase in true negatives. The sigmoid curve in the plot shows that the FNR remains low until a threshold of around 60%, after which it starts to increase rapidly. This indicates that the face recognition system starts to make more false negative identifications (i.e., it fails to identify the target individual) as the confidence threshold increases beyond a certain point. The plot suggests that the face recognition system is more accurate at lower confidence thresholds, but as the threshold increases, the system becomes more conservative and cautious, which leads to an increase in false negatives.
In the context of *Attack 3 - DemogPairsPairs - Different Faces - Same Demo*, the true false rate represents the probability of the model making a wrong prediction given a certain confidence threshold [**Figure** 28]. When the dataset shows nil or very low confidence thresholds until 30%, it means that the model is not very confident in its predictions, and therefore is making many errors. As the confidence threshold increases, the model becomes more confident and the true false rate begins to increase, meaning the model is making more incorrect predictions. The sigmoid curve indicates that there is a threshold at which the model is confident enough to make accurate predictions while avoiding too many incorrect predictions. Beyond this threshold, the model becomes overconfident and makes more errors. The reason for this trend could be due to the difficulty of the task - identifying different faces of people from the same demographic - which may be inherently challenging for the model, particularly when it is not very confident in its predictions. Additionally, it could be due to biases in the training data, which may lead to the model being less accurate for certain demographic groups.

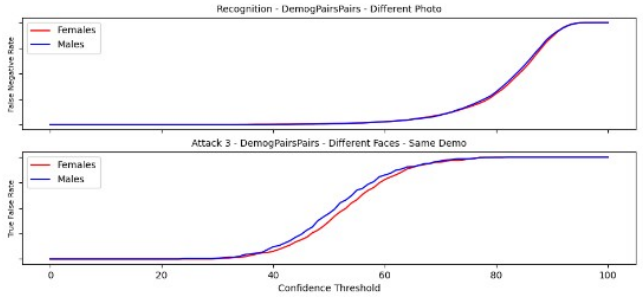


Figure 28. DemogPairs - Different Photos

In this case[**Figure** 29], the graph shows the performance of an obfuscation attack on a dataset of faces with the same demographic (RFW). The obfuscation success rate is plotted on the y-axis against the confidence threshold on the x-axis. The sigmoid curve indicates that the obfuscation success rate initially increases slowly as the confidence threshold is lowered, but then accelerates rapidly as the threshold approaches 40%. This means that at a confidence threshold of 40%, the adversarial attack is most effective at obfuscating the identities of the individuals in the dataset. The fact that the success rate remains low until the confidence threshold reaches 40% suggests that the facial recognition model is relatively robust to simple modifications or perturbations to the input image. However, beyond a certain threshold, the adversarial attack is able to modify the input image in a way that causes the model to fail, resulting in a high obfuscation

success rate. This suggests that the model is vulnerable to more sophisticated attacks that can modify the input image in a more subtle way.

In the context of the dataset RFW (RFW - Same Photo)[**Figure 29**], the obfuscation success rate on the y-axis and confidence threshold on the x-axis shows the performance of an algorithm in obfuscating faces belonging to different ethnic groups in the same photo. The four ethnic groups in the dataset are African, Asian, Caucasian, and Indian. The obfuscation success rate represents the percentage of faces that the algorithm successfully obfuscated, while the confidence threshold represents the level of confidence required by the algorithm to consider a face as a match to a particular ethnicity. The plot shows that the algorithm is initially not confident in its ability to classify the ethnicity of the faces, as the confidence threshold is low and the obfuscation success rate is also low or negligible. As the confidence threshold increases, the algorithm becomes more confident in its predictions, and the obfuscation success rate gradually increases until it reaches a maximum value.

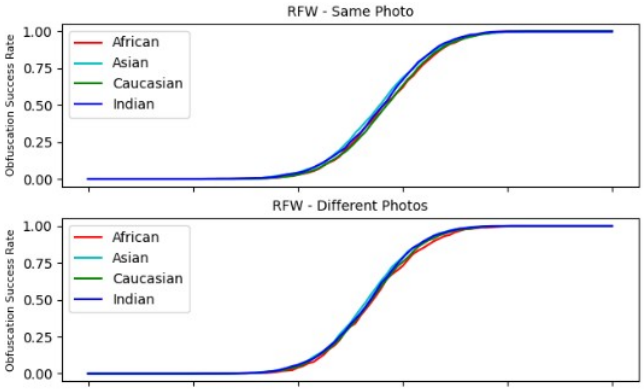


Figure 29. RFW - Obfuscation Success Rate

The *False negative rate vs Confidence Threshold plot*[**Figure 30**] for the Recognition - RFW - Different Photos dataset represents the performance of a face recognition system where an attacker has access to different photos of the same individual from four demographic groups (African, Asian, Caucasian, and Indian). The plot shows that the false negative rate is negligible or very low for all four demographic groups until a confidence threshold of 60%. After this threshold, the false negative rate starts increasing and reaches its maximum at around 80% confidence threshold. This suggests that the system can accurately recognize individuals with high confidence until a certain threshold, beyond which its performance starts to degrade. The sudden increase in the false negative rate at higher confidence thresholds can be attributed to the system becoming more cautious in its decision-making and requiring higher confidence to make a positive match. This can result in the rejection of genuine matches and an increase in the false negative rate.

The plot of *True false rate on y-axis and confidence threshold* on x-axis for the dataset RFW - Different Faces - Same DemogPairs represents the performance of a face recognition system on a dataset where the attacker has access to different photos of individuals from four different demographic groups (African, Asian, Caucasian, Indian) but they are all from the same demographic pair. The plot shows that for all four demographic groups, the true false rate is very low or almost nil until a confidence threshold of 20%, after which it starts to increase and reaches a maximum value at around 60% confidence threshold. This means that the system is able to correctly distinguish between genuine matches and impostors with a high accuracy until a certain threshold, beyond which the performance starts to degrade. The sudden increase in the true false rate at higher confidence thresholds may be due to the fact that the system becomes more lenient in its decision-making and requires a lower degree of confidence to make a positive match. This can lead to accepting impostors and increasing the true false rate.

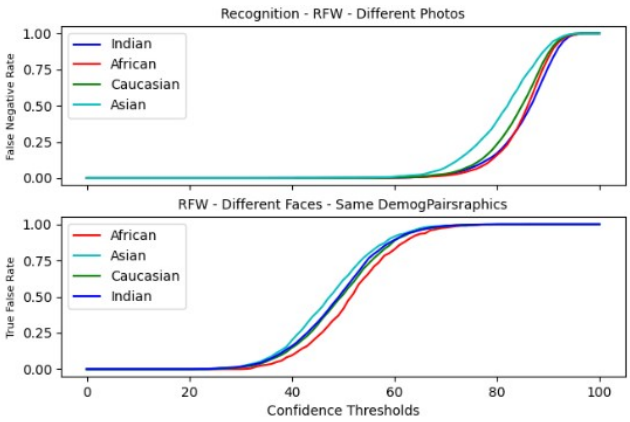


Figure 30. RFW - Different Photos

The plot of Obfuscation success rate on y-axis and confidence threshold on x-axis for the dataset BFW - Same Photo[**first plot , Figure 31**] represents the performance of an obfuscation algorithm on a dataset consisting of individuals from four different demographic groups (Indian, Black, White, Asian). The plot shows that for all four demographic groups, the obfuscation success rate is very low or almost nil until a confidence threshold of 30%, after which it starts to increase and reaches a maximum value at around 80% confidence threshold. This means that the obfuscation algorithm is effective in protecting the identities of the individuals only when the confidence threshold is above a certain value. At low confidence thresholds, the algorithm is not able to effectively obfuscate the identities, possibly due to the high variability in the facial features among individuals in the dataset.

The plot of Obfuscation success rate on y-axis and confidence threshold on x-axis for the dataset BFW - Different Photo[**second plot, Figure 31**] represents the performance of an obfuscation technique on a face recognition system. The plot shows that for all four demographic groups (Indian, Black, White, and Asian), the obfuscation success rate is very low or almost nil until a confidence threshold of 40%, after which it starts to increase and reaches a maximum value at around 70-80% confidence threshold. This means that the obfuscation technique is effective in hiding the identity of individuals from the face recognition system until a certain threshold, beyond which the effectiveness starts to degrade. The sudden increase in the obfuscation success rate at higher confidence thresholds may be due to the fact that the obfuscation technique becomes more effective in hiding the individual's identity as the confidence of the recognition system decreases. The plot of False negative rate on y-axis and

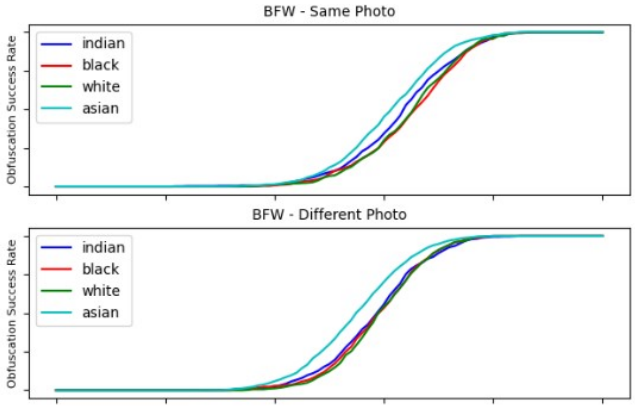


Figure 31. BFW -Obfuscation Success Rate

confidence threshold on x-axis[**first plot , Figure 32**] for the dataset *Recognition - BFW - Different Photos* represents the performance of a face recognition system on a dataset where the attacker has access to different photos of the same individual from four different demographic groups (Indian, Black, White, and Asian). The plot shows that for all four demographic groups, the false negative rate is very low or almost nil until a confidence threshold of 40%, after which it starts to increase and reaches a maximum value at around 80% confidence threshold. This means that the system is able to correctly recognize the individuals with a high accuracy until a certain threshold, beyond which the performance starts to degrade. The sudden increase in the false negative rate at higher confidence thresholds may be due to the fact that the system becomes more conservative in its decision-making and requires a higher degree of confidence to make a positive match. This can lead to rejecting genuine matches and increasing the false negative rate.
Referring to the [**second plot of Figure 32**] of True false rate on y-axis and confidence threshold on x-axis for the dataset *BFW - Different Faces - Same DemogPairs* represents the performance of an attack on a face recognition system. The plot shows that for all four demographic groups, the true false rate is very low or almost nil until a confidence threshold of 20%, after which it starts to increase and reaches a maximum value at around 60-80% confidence threshold. This means that the face recognition system is able to correctly identify the genuine individuals with high accuracy until a certain threshold, beyond which the system becomes vulnerable to the attack.

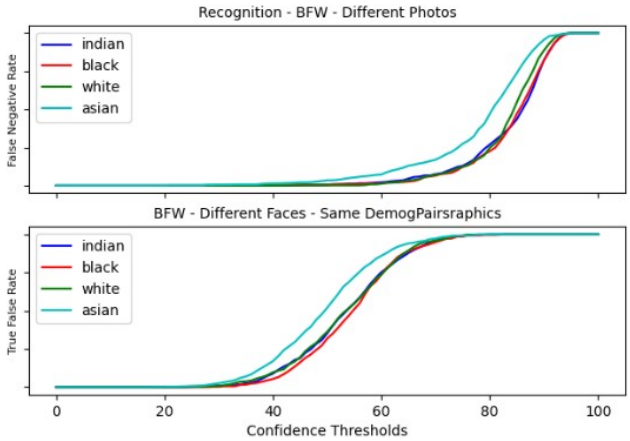


Figure 32. BFW - Different Photos

The plot of obfuscation success rate on y-axis and confidence threshold on x-axis for *DemogPairs - Same Photo*[**first plot , Figure 33**] represents the performance of an obfuscation technique on a dataset where the attacker has access to the same photo of an individual from three different demographic groups (Black, White, Asian).The plot shows that for all three demographic groups, the obfuscation success rate is very low or almost nil until a confidence threshold of around 30%, after which it starts to increase and reaches a maximum value. This means that the obfuscation technique is not effective in protecting privacy at low confidence thresholds, but becomes more effective as the confidence threshold increases. The sudden increase in obfuscation success rate at higher confidence thresholds may be due to the fact that the face recognition system becomes more conservative in its decision-making and requires a higher degree of confidence to make a positive match. This can lead to rejecting more genuine matches and decreasing the recognition accuracy, making it more difficult for the attacker to recognize the individual.
The plot of obfuscation success rate on the y-axis and confidence threshold on the x-axis for the *DemogPairs - Different Photo*[**first plot , Figure 33**] dataset indicates the effectiveness of obfuscation techniques in preventing face recognition

systems from correctly identifying individuals. This dataset consists of three demographic groups: Black, White, and Asian, and the plot shows that for all of them, the obfuscation success rate is initially low or nil until a confidence threshold of around 30%. At this threshold, the obfuscation techniques become effective in confusing the face recognition system, and the success rate starts to increase rapidly, reaching a maximum value. This indicates that the system is unable to correctly recognize the individuals in the images once they have been obfuscated beyond a certain degree. The sudden increase in the obfuscation success rate at around 30% confidence threshold suggests that the face recognition system is sensitive to changes in the images and can be easily confused by even small degrees of obfuscation. This makes it difficult for the system to accurately identify individuals in images where obfuscation techniques have been applied. The

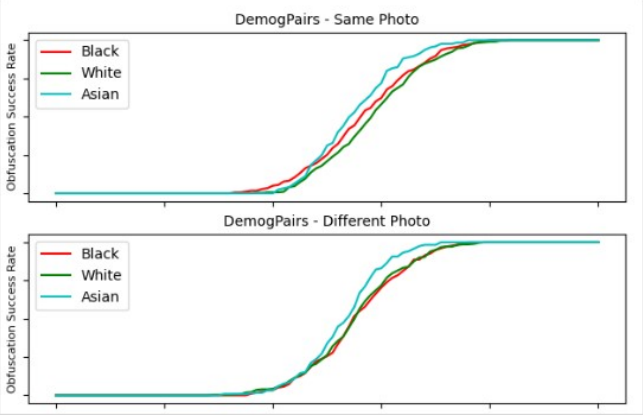


Figure 33. DemogPairs - Obfuscation Success Rate

plot of False negative rate on y-axis and confidence threshold on x-axis[**first plot , Figure 34**] shows the performance of a face recognition system on a dataset where the attacker has access to different photos of the same individual from three different demographic groups (Black, White, and Asian). The plot indicates that the false negative rate for all three demographic groups is very low or almost nil until a confidence threshold of 50%, after which it starts to increase and reaches a maximum value. This suggests that the face recognition system can accurately recognize individuals with a high level of confidence until a certain threshold, beyond which its performance starts to degrade. The sudden increase in the false negative rate at higher confidence thresholds may be due to the system becoming more conservative in its decision-making and requiring a higher degree of confidence to make a positive match. This can lead to rejecting genuine matches and increasing the false negative rate.

The plot of True False rate on the y-axis and confidence threshold on the x-axis for the dataset *DemogPairs - Different Faces [**first plot , Figure 33**]* represents the performance of a face recognition system on a dataset where the attacker has access to different faces of individuals from three demographic groups (Black, White, Asian).The plot shows that for all three demographic groups, the TFR is very low or almost nil until a confidence threshold of 30%, after which it starts to increase and reaches a maximum value at around 80% confidence threshold. This means that the system is able to correctly distinguish between genuine and impostor matches with a high accuracy until a certain threshold, beyond which the performance starts to degrade. The sudden increase in TFR at higher confidence thresholds may be due to the fact that the system becomes more conservative in its decision-making and requires a higher degree of confidence to reject impostor matches. This can lead to accepting impostor matches as genuine and increasing the TFR.

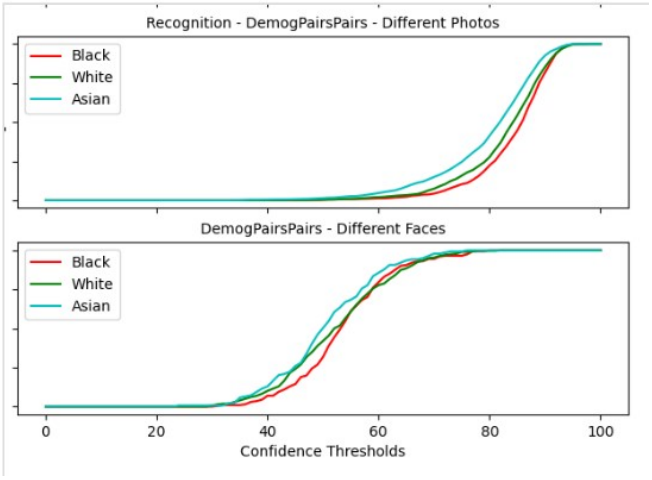
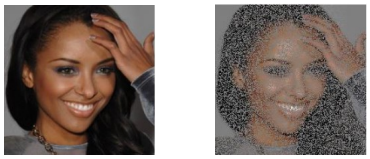


Figure 34. DemogPairs - Different Photos

7.1.3 DP-SNOW. In this section, we present the results obtained from DP-SNOW [6] based on the original and obfuscated image samples for each dataset[*Figure: 35, 36, 37*]. The first image is the input image and second is the resulting output image generated from the method as shown in **Algorithm 3**. The parameters used for the method to yeild the obfuscated images from the original source images are: *Privacy Budget (δ) = 0.5* For the *BFW dataset*[9], we have

(a) Source **(b)** Obfuscated Image

Figure 35. BFW Dataset



(a) Source **(b)** Obfuscated Image

Figure 36. DemogPairs Dataset

(a) Source **(b)** Obfuscated Image

Figure 37. RFW Dataset

evaluated the algorithm's performance using three different choices of parameters.The confidence scores obtained from *Face*$^{++}$[7] are in the range as follows:

	Max Confidence	Min Confidence
Asian_females	93.362	52.063
Asian_males	94.484	69.049
Black_females	95.155	65.549
Black_males	95.347	62.536
Indian_females	94.092	69.926
Indian_males	94.721	59.513
White_females	93.492	65.456
White_males	94.199	64.323

Table 19. BFW - Choice 1

	Max Confidence	Min Confidence
Asian_females	85.648	27.259
Asian_males	88.616	30.256
Black_females	90.43	43.681
Black_males	89.601	36.119
Indian_females	90.917	37.753
Indian_males	91.23	31.408
White_females	88.197	43.274
White_males	89.048	40.702

Table 20. BFW - Choice 2

	Max Confidence	Min Confidence
Asian_females	74.614	14.6
Asian_males	70.535	14.624
Black_females	82.838	20.636
Black_males	82.177	28.746
Indian_females	80.054	20.122
Indian_males	81.065	23.083
White_females	83.405	24.048
White_males	78.638	15.799

Table 21. BFW - Choice 3

For the *RFW Dataset*[11], we have evaluated the algorithm's performance using a single choice of parameters. The confidence scores obtained from *Face*$^{++}$[7] are in the range as follows:

	african	asian	indian	caucasian
Max Confidence	94.344	94.536	94.893	94.095
Min Confidence	39.495	57.263	48.632	36.22

Table 22. RFW - Choice 1

	african	asian	indian	caucasian
Max Confidence	91.042	92.111	92.3	91.914
Min Confidence	40.596	45.946	31.391	32.82

Table 23. RFW - Choice 2

	african	asian	indian	caucasian
Max Confidence	80.989	72.888	77.575	79.308
Min Confidence	20.66	7.776	12.679	12.361

Table 24. RFW - Choice 3

For the *DemogPairs dataset*[5], we have evaluated the algorithm's performance using three different choices of parameters.The confidence scores obtained from *Face*$^{++}$[7] are in the range as follows:

	Max Confidence	Min Confidence
Asian_female	93.862	48.383
Asian_male	93.87	51.247
Black_female	95.074	47.802
Black_male	94.519	45.785
White_female	92.858	57.64
White_male	93.954	56.414

Table 25. Demog - Choice 1

	Max Confidence	Min Confidence
Asian_female	92.149	37.206
Asian_male	89.685	35.354
Black_female	92.221	40.931
Black_male	92.913	38.513
White_female	89.819	29.686
White_male	90.1	41.061

Table 26. Demog - Choice 2

	Max Confidence	Min Confidence
Asian_female	74.391	15.41
Asian_male	70.3	14.348
Black_female	79.323	20.658
Black_male	81.234	27.369
White_female	74.686	17.816
White_male	79.226	17.138

Table 27. Demog - Choice 3

For the DemogPairs same photo, the graph of obfuscation success rate vs Confidence threshold [*first plot , Figure 38*] takes a sigmoid like curve for the different labels. Till the confidence threshold hits around 57%, the graph is a straight line indicating that the obfuscation technique is still not very effective in reducing the accuracy of the facial recognition system. But as the threshold hits 58%, it is observed that the curve starts to the shape for*Asian_males* first and then for other remaining labels. This could be due to the fact that the system is highly confident in recognition and is therefore more resistant to the obfuscation technique, but it starts to fail as the confidence threshold hits to around 58%.
As the confidence threshold increases, it is observed that the obfuscation rate for different labels are around the same with some slight difference in obfuscation rate. But the gap is large for *Asian_females* as compared to *Black_females* for the recognition system. This might be due to the difference in facial features or skin tone such that the recognition systems fail to identify them.
DemogPairs dataset[5] for different photos varies depending on the demographic group. For every data label or demographic group, the plot starts taking the sigmoid curve at around 43%. For Asian individuals, the success rate increases steeply, indicating that the obfuscation method used is more effective in hiding their identity as the confidence threshold increases. Whereas for Black individuals, the curve gradually increases even though the confidence threshold in comparison to Asian groups. On reaching the confidence threshold of 90%, the success rate of obfuscation is at the peak. Overall, this information may indicate that the obfuscation success rate may vary depending on the demographic group of the individual's image being obfuscated.

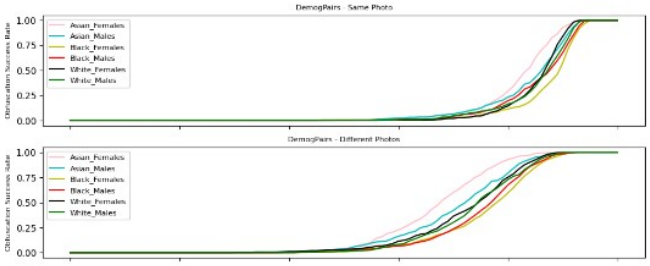


Figure 38. DemogPairs - Obfuscation Success Rate

In this case of the DemogPairs dataset[5], where different photos are used for each demographic group, the False Negative Rate(FNR) vs Confidence Threshold shows a sigmoid curve for all the groups[*first plot , Figure 39*]. This shows that the system's performance in recognizing the person after the image is obfuscated improves as the confidence threshold increases. Also, the confidence threshold is almost 0% indicating that the facial recognition system is not able to identify the person until a certain threshold is reached (50% in this case). So as the obfuscation level is reduced (i.e. confidence threshold increases), the system can correctly recognize the individual from the obfuscated images and FNR decreases. Overall, this plot provides some important insights as to how well the system is performing and how well it can recognize the individuals after obfuscation as well as the effect on accuracy based on the changes in confidence threshold.
For the second case [*second plot , Figure 39*], where different photos but of the same demographic group is considered, the True False Rate (TFR) varies as a function of the confidence threshold. The TFR measures how often the model incorrectly classifies (false positive) an image along with the correct classification of the image (true negative). For *Asian_males*, the graph starts low from TFR of 0.00 till confidence threshold is 17% and then rises to TFR of 1.00 at 62%.

This indicates that for the *Asian_males* group, the model is making more errors in classification at a low confidence threshold of 17% and then as soon as it reaches to around 60%, it starts to correctly classify the images. As for other demographic groups, the model makes classification errors if the threshold is in the range of 20 to 37%. Once the threshold crosses the mark of 40%, the errors are less and the model is able to make correct classification of the images. Once the threshold hits 75%, the models is levelled suggesting that the model is more confident in its predictions for the demographic group

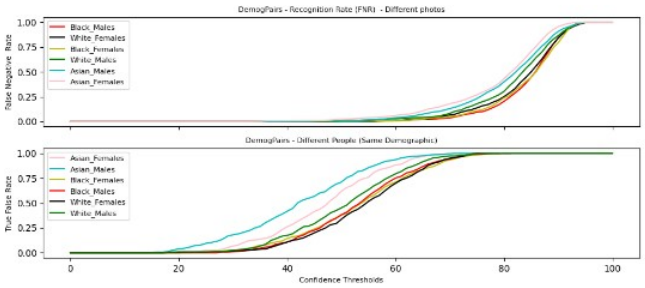


Figure 39. DemogPairs - Different Photos

The plot for obfuscation success rate vs confidence threshold for the BFW dataset[9] for the same photo [*Figure 40*] indicates how effective the obfuscation method is in concealing the identities of the individuals in the photos for different confidence thresholds. In this case, the obfuscation method does not seem to be effective up until the threshold reaches 75%, indicating the identities of the person can be easily recognized. Once it crosses that mark, it becomes more difficult for the model to recognize the person in the image.
Whereas for the case of different photos being used, the threshold of 40% marks the point from where the images start getting recognized by the model. For the *Asian* group of people it is a bit easy for the model to recognize as compared to other demographic groups of people for whom the model correctly recognizes people at near the 55% confidence threshold. The reason for such a bias might be the facial features or skin tone of the images of people labelled as *Asian* in the dataset.

In the BFW dataset[9] for Different Photos scenario[*Figure 41*], all the demographic groups (*Asian_females, Asian_males, Black_females, Black_males, White_females, White_males*) show a similar trend in terms of FNR and confidence threshold. For *Asian_males*, the FNR starts to slightly increase at the confidence threshold of 30%. Whereas for other groups the FNR is almost 0 till the threshold is around 58% and then starts to gradually increase. This indicates that until the confidence threshold hits the mark of 58%, the model is not able to correctly identify the faces in the dataset. However, as the threshold increases, the FNR also increases and follows a

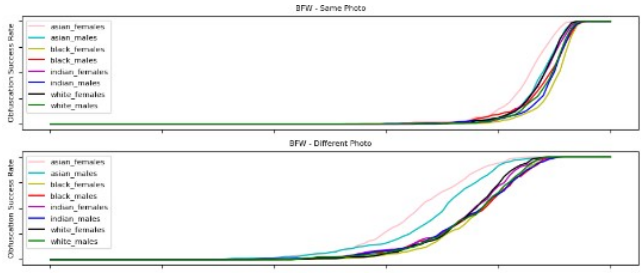


Figure 40. BFW - Obfuscation Success Rate

sigmoid curve. This trend indicates that the model becomes more accurate in identifying faces as the level of certainty required for a prediction increases.

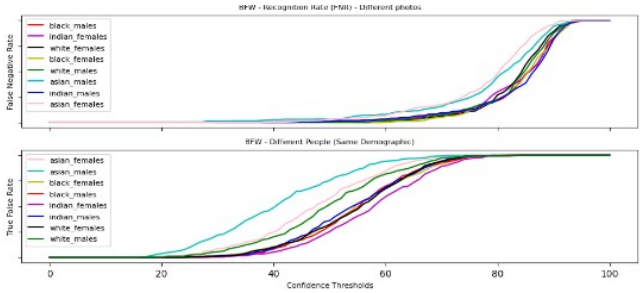


Figure 41. BFW - Different Photos

In the context of *Attack 1 on the BFW dataset[9]*, the plot shows that for both female and male same photos (the original image and the obfuscated image of the same photo)[*Figure 42*], the obfuscation success rate is very low or almost nil until the confidence threshold reaches 75%. This indicates that the generated obfuscated faces were not successful in hiding the identity of the person until a certain level of confidence was reached. However, after this there is a gradual increase in the obfuscation success rate till 83% and then suddenly skyrocketed at 92% which stayed the same till the end. This sudden increase in the success rate may be due to the generated faces becoming more similar to the original faces as the confidence threshold increases.
For *Attack - 2 on the BFW dataset*, the obfuscation success rate is plotted against the confidence threshold for the same photo attack for both males and females separately. Both the data groups have a similar trend behaviour in terms of the success rate given the confidence threshold. It is observed that till 40% confidence threshold the success rate of obfuscation is almost 0 which then gradually rises following a sigmoid curve till the very end. This means that the facial recognition algorithm is less accurate in recognizing the faces with low confidence scores and can be fooled with the same photo attack by generating a low confidence score for the dataset.

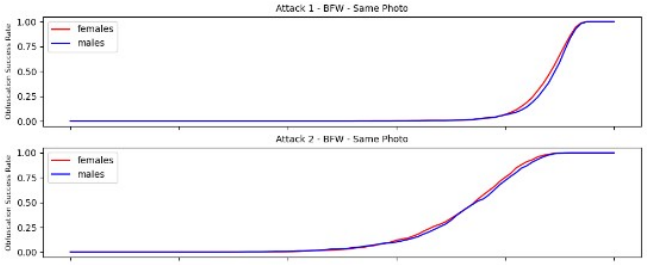


Figure 42. BFW - Attack - Same Photos

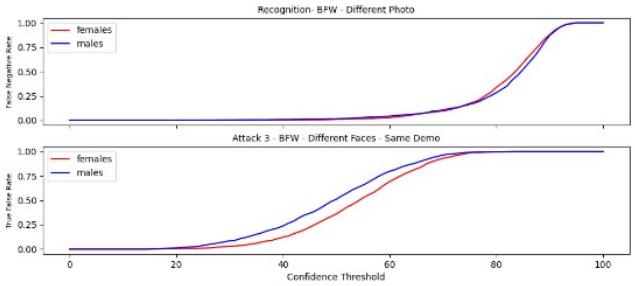


Figure 43. BFW - DIfferent Photos

In the case of *Recognition - BFW with different photos* [*Figure 43, Plot 1*], we see that both male and female data groups have very low FNR (False Negative Rate) until a confidence threshold of 60% is reached. After this, there is a sharp increase in the FNR as the confidence threshold increases. This means that the algorithm is able to correctly identify matches with a high degree of accuracy until a certain confidence threshold is reached. After this threshold, the algorithm begins to make more false negative identifications resulting in the model being less accurate. The reason for this trend might be that as the confidence threshold increases, the algorithm becomes more conservative and is less likely to make positive classifications. This may result in some matches being missed leading to an increase in FNR. Another possibility is that as the confidence threshold increases, the algorithm becomes more susceptible to errors due to image noise, changes in lighting conditions, or other factors that may affect the quality of the input photo.

In *Attack 3 - BFW - Different Faces - Same Demo*[*Figure 43, Plot 2*], the True False Rate is plotted against the confidence threshold for the attack where different faces of the same demographic group are used. The TFR represents the percentage of times when the attacker can correctly identify a different photo of the same individual as belonging to that person. Thus, a lower TFR indicates higher obfuscation success rate. In the plot we can see that the True False Rate is almost 0 for both the data groups till the confidence threshold of 20% is reached, whereas it is 35% for the males data group. After this both females and males follow a sigmoid curve trend reaching the maximum True False Rate at nearly 80% of the threshold and remaining stable till 100.

For the *first attack on DemogPairs dataset considering the same photo*[*Figure 44, Plot 1*], it is observed that the plot follows sigmoid trend indicating that at lower confidence thresholds, the attack is not very successful i.e. the obfuscated images are still being recognized by the system as the original face. However as the confidence threshold of 75% is reached, the obfuscation rate also increases indicating that the obfuscated images became more and more unrecognisable then the original image.

In *Attack 2 - Demogpairs - Same Photo*[*Figure 43, Plot 2*], the behaviour of the plot is similar to that of a sigmoid curve. This indicates that the model is not able to correctly recognize the obfuscated image until a threshold of 42% is reached and then it gradually starts to identify the images until the maximum value is reached. The reason for this type of model behaviour might be due to some facial features, environment lightings or skin tone of the person in the image.

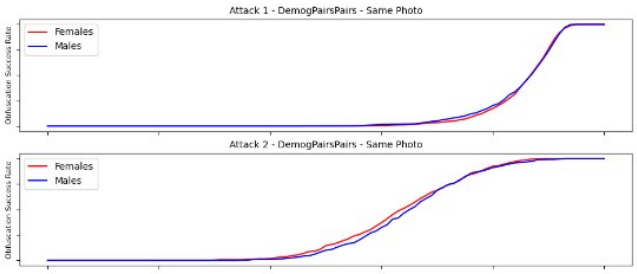


Figure 44. DemogPairs - Attack - Same Photos

The graph *Recognition - DemogPairs with Different Photo*[*Figure 45, Plot 1*] but same labels is plotted with False Negative Rage(FNR) on y-axis with Confidence Threshold as x-axis. The plot exhibits a similar trend as that of DP-Pix method. Here, the FNR is almost 0 till the confidence reaches a threshold of 60% beyond which it begins to rise. The algorithm can reliably recognize persons up to a certain point, after which its performance begins to deteriorate. The abrupt increase in the false negative rate at higher confidence thresholds might be attributed to the system becoming more conservative in its decision-making and requiring a higher level of certainty to establish a positive match. As a result, real matches may be rejected, and the false negative rate may rise.

In the case of *Attack 3 - DemogPairs with Different Faces but Same demographic*[*Figure 45, Plot 2*], the graph of True False Rate vs Confidence Threshold follows a sigmoid trend.The true false rate represents the fraction of positive matches that are incorrectly identified as negative matches by the system. A low true false rate is desirable as it implies that the system can correctly identify positive matches even when

presented with different photos and demographic groups. The plot shows that for both females and males, the true false rate is very low or almost nil until a confidence threshold of 20%, after which it starts to increase and reaches a maximum value at around 75% confidence threshold. This means that the obfuscation attack is successful in reducing the recognition accuracy of the system, as it is able to produce false negatives at a high rate once the confidence threshold is crossed.

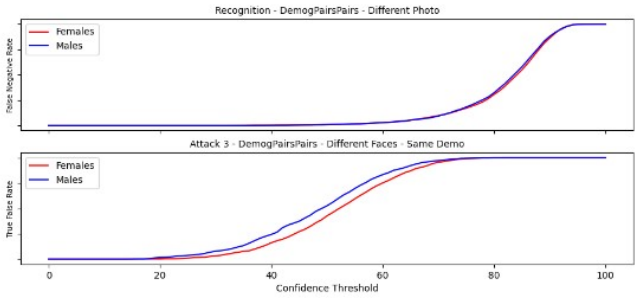


Figure 45. DemogPairs - Different Photos

The plot of obfuscation success rate on the y-axis and confidence threshold on the x-axis for the RFW (Racial Faces in the Wild) dataset[11][*Figure 46, Plot 1*] with different demographic groups demonstrates the efficacy of the facial recognition system obfuscation assault. The plot reveals that until a confidence level of 62%, the obfuscation success rate is very low or nonexistent for all demographic groups in the sample (*African, Asian, Caucasian, and Indian*). After this point, the success rate begins to rise and finally reaches its peak. This means that the attacker can successfully conceal the facial traits of the persons in the dataset, preventing the face recognition system from accurately detecting them. A sharp rise in obfuscation success rate at higher confidence threshold might be explained by the fact that the perturbation given to the faces becomes more effective and begins to drastically modify facial characteristics. As a result, the facial recognition system's ability to identify persons becomes less dependable and accurate.

The plot of the obfuscation success rate on the y-axis and the confidence threshold on the x-axis for the dataset RFW - Different Demo[*Figure 46, Plot 2*] represents the performance of the obfuscation attack on a face recognition system trained on a dataset with faces from four different demographic groups: *African, Asian, Caucasian, and Indian*. The assault in this example is carried out by producing synthetic faces from various demographic groups in order to obscure the original face and hinder the system from accurately detecting the subject. The obfuscation success rate is the proportion of the attacker's synthetic faces that successfully deceive the facial recognition system into producing an inaccurate match. A high obfuscation success rate suggests that the attacker was successful in obscuring the original face and preventing the system from accurately recognizing the subject. The plot reveals that until a confidence level of 48%, the obfuscation success rate is very low or nearly nil for all demographic groups, after which it begins to climb and reaches a maximum value at about 90% confidence barrier. This means that the obfuscation approach is ineffective until a specific level is reached, after which it becomes increasingly effective at deceiving the facial recognition system.

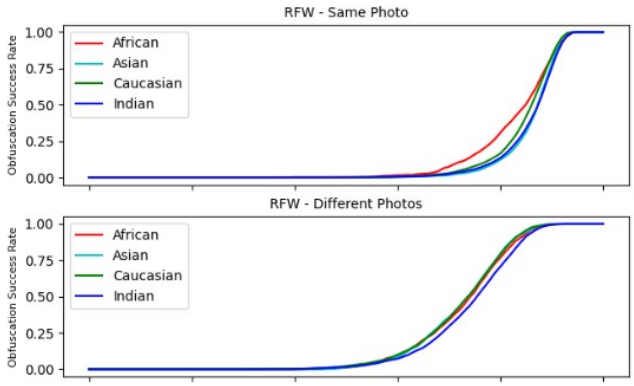


Figure 46. RFW - Obfuscation Success Rate

The plot of False negative rate on the y-axis and confidence threshold on the x-axis for the dataset Recognition - RFW - Different Photos[*Figure 47*] represents the performance of the face recognition system on a dataset where the attacker has access to multiple photos of the same individual from four different demographic groups (*African, Asian, Caucasian, and Indian*). The plot indicates that until a confidence level of 60%, the false negative rate is very low or practically zero for all four demographic groups, after which it begins to climb and reaches a maximum value around 95%. This indicates that the system can accurately distinguish persons with great accuracy up to a certain point, beyond which performance begins to deteriorate. This abrupt increase in the false negative rate at higher confidence levels might be attributed to the system becoming more cautious in its decision-making and requiring a greater level of confidence to make a positive match. This may result in the rejection of real matches and an increase in the false negative rate.

Referring to **Figure 48**, in the case of *BFW dataset with the Same photo*, the graph is plotted with Obfuscation Success Rate on the y-axis with Confidence threshold on the x-axis. The plot shows that for all demographic groups namely *Indian, Black, White and Asian* the success rate is extremely low (almost 0) till the confidence reaches a threshold of 75%. This indicates that the method has a very low success rate for obfuscation at confidence levels < 75%. But then it gets skyrockets and rate achieves a maximum value when the confidence crosses the threshold of 92%. But for a fact, the

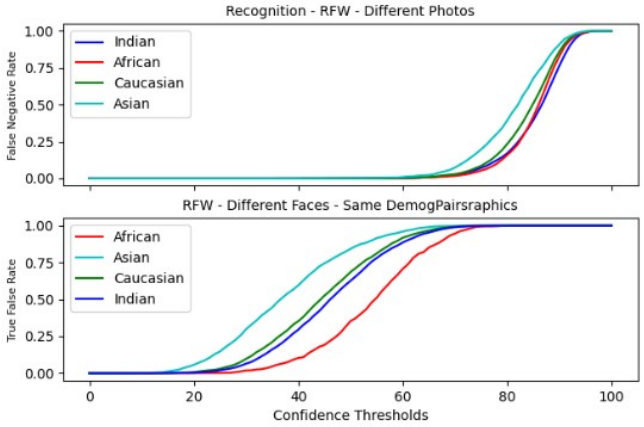


Figure 47. RFW - Different Photos

curve for *Asian* data group is more steep as compared to that of other data groups may be because of specific facial features or characteristics of *Asian* people making them more difficult to recognize or distinguish when the same photo is considered.

Whereas, in the case of Different photos condition, the plot behaves similarly to a sigmoid and also to that of Same Photo case. But, till the confidence threshold of 40%, the success rate is almost 0 indicating that the obfuscation method fails to hide the identities of the person and the recognition system is able to recognize the images. But as soon as the threshold crosses the 40% mark, the plot gradually increases and attains the maximum value for success rate for *Asian* data groups at 82% and 87 for the other remaining groups. This may indicate that for *Asian* group, the algorithm is able to fully hide the necessary details in the image at a lower confidence threshold as compared to that of other groups resulting in a bias towards the Asian data group.

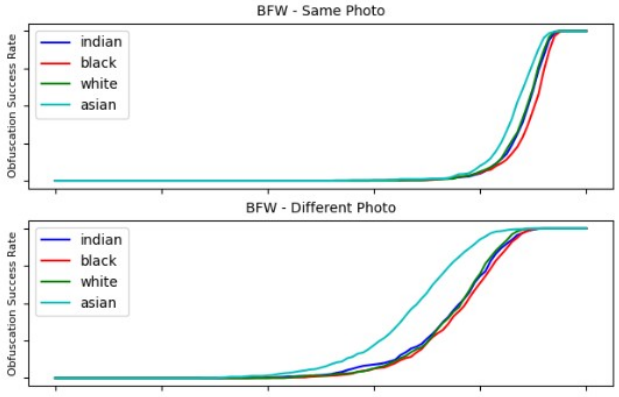


Figure 48. BFW - Obfuscation Success Rate

The plot of the False Negative rate on the y-axis and the confidence threshold on the x-axis for the dataset *Recognition - BFW - Different Photos [Figure 49, Plot 1]* represents the performance of a face recognition system on a dataset with individuals from various demographic groups, such as *Indian, Black, White, and Asian.* The plot indicates that for confidence criteria below 40%, the false negative rate is very low or negligible for all demographic categories. This indicates that at certain confidence levels, the face recognition system can accurately identify individuals from various demographic categories with high accuracy. However, once the confidence level above 40%, the false negative rate steadily rises, demonstrating that the face recognition system becomes less reliable in recognizing individuals from various demographic groups.The reason for this increase in false negative rate at higher confidence thresholds could be that the face recognition system is less certain about the identity of the individuals, particularly those from different demographic groups whose facial features and characteristics may differ from those in the training set. As a result, at higher confidence levels, the face recognition system may be more likely to mistakenly reject the identification of persons from diverse demographic groups, resulting in a larger false negative rate.

Considering the case of *BFW dataset with different faces but same demographic information [Figure 49, Plot 2]*, the graph is plotted as True False Rate(TFR) vs Confidence threshold which represents the performance of a face recognition system on a dataset where the faces are from different photos but belong to the same demographic group. It is observed that till 18% confidence the TFR was 0 for *Asian* groups which means that the face recognition system is less accurate in correctly classifying these individuals at those confidence intervals. But as soon as the confidence breaks the threshold of 18%, the graph exhibits a sigmoid behaviour and attains a maximum value of TFR at only 70% confidence. But for other data groups the TFR is 0 till the confidence of 22% but attains the maximum value at 70%, the same as that of the *Asian* data group. This could be due to the fact that the faces are from different photos but belong to the same demographic group, making them more similar and harder to distinguish.Overall, the plot indicates that the face recognition system has a lower accuracy in properly recognizing persons from multiple photographs that belong to the same demographic category, particularly for *Asian* people.

DemogPairs - Same plot of obfuscation success rate on y-axis and confidence threshold on x-axis.The efficacy of an obfuscation strategy on a dataset where the attacker has access to the same photo of persons from various demographic groups is represented by Figure 50, Plot 1. The dataset in this scenario contains *Black, White, and Asian* individuals. The plot shows that for all the data groups, the obfuscation success rate is extremely low (almost 0) till the confidence

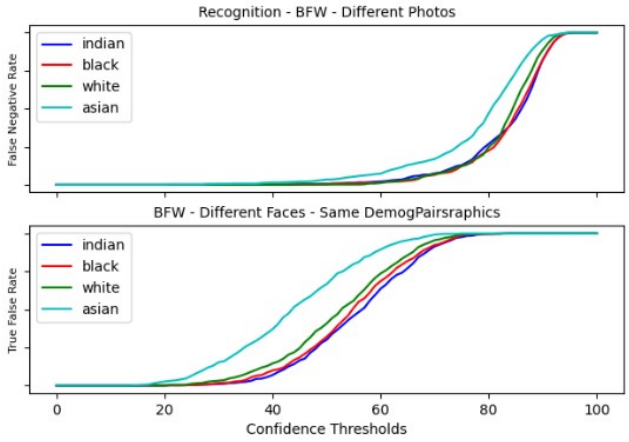


Figure 49. BFW - Different Photos

threshold of 60%. Then the graph has a sigmoidal increase with a steep curve for Black data group as compared to that of White and Asian group and then achieves the maximum value of success rate at confidence of 91%. The reason for the gradual increase in the success rate could be due to the fact that the facial features or characteristics of *Black* individuals are less distinct or easier to confuse than those of other groups, making it easier to obfuscate their faces.
On the other hand, for the case of *Different Photo on the same dataset [Figure 50, Plot 2]*, the graph follows a proper sigmoidal behaviour. The success rate is still 0 but at a lower confidence than the Same Photo case. The success rate remains 0 till the threshold of 42% and then gradually increases to attain the maximum value when the threshold crosses the mark of 92%

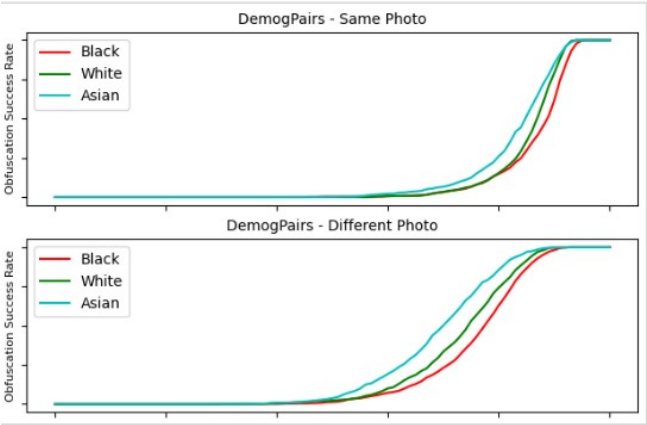


Figure 50. DemogPairs - Obfuscation Success Rate

The *Recognition - DemogPairs [Figure 51, Plot 1]* - distinct images dataset's False Negative (FN) rate on the y-axis and confidence threshold on the x-axis indicate the performance of a face recognition system while recognizing persons from distinct demographic groups (*Black, White, and Asian*) using different images. The plot demonstrates that the FN rate is low or negligible for all demographic categories at low confidence thresholds, indicating that the face recognition system is adept at properly detecting persons at such levels. However, when the confidence threshold rises, so does the FN rate, suggesting that the face recognition system begins to produce more false negatives. The dramatic increase in the FN rate for all demographic groups at roughly the 50% confidence level implies that the face recognition system has problems accurately recognizing individuals when the confidence threshold is high. The sigmoid curve illustrates that when the confidence threshold increases, the FN rate climbs fast, indicating that the face recognition system is more likely to make a false negative. This might be due to a variety of causes such as changes in lighting, facial expressions, occlusions, or other factors affecting face recognition performance.
The DemogPairs - Different Faces dataset's True False rate plot on the y-axis and confidence threshold plot on the x-axis displays the performance of a face recognition system on photos of persons from various demographic groups[Figure 51, Plot 2]. Individuals in the sample areemph Black, White, and Asian. For *Asian* people, the figure exhibits a sigmoid curve around the 18% threshold, indicating that the system's True False rate is gradually increasing around that point. This indicates that when the confidence threshold drops, the algorithm gets less accurate in distinguishing *Asian*. This might be because these people have distinct face traits or qualities that make them difficult to differentiate or recognize. Whereas for *Black and White* data groups, the graph has almost similar traits of following a sigmoidal trend at the confidence threshold of 28%. This suggests that the method is more effective at differentiating *Asian* people from other demographic groupings. This might be because *Asian* people's face traits or attributes are more unique and simpler to distinguish when compared to other demographic groupings.

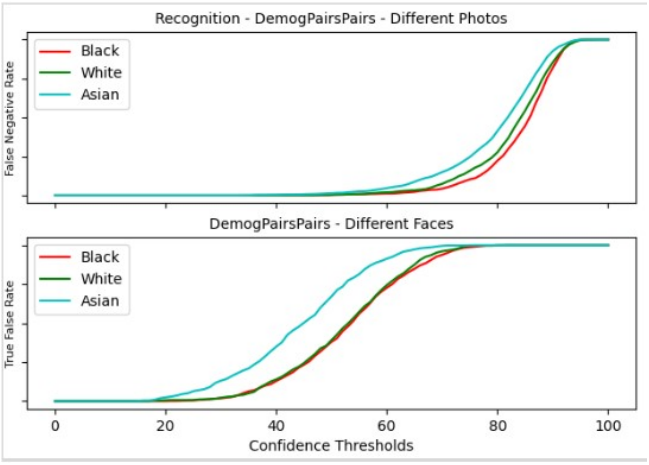


Figure 51. DemogPairs - Different Photos

7.1.4 CNN-based Live Video De-identification. The obfuscation process is achieved by training the model to learn a compressed representation of the images, which inherently introduces some loss of information. After the photos have been obfuscated, we compare the original and obfuscated images using a number of assessment measures, such as *Mean Absolute Error (MAE), Mean Squared Error (MSE), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR).* These measurements, which take into account several facets of image comparison, offer a thorough evaluation of the obfuscation process. To perform evaluation in this research

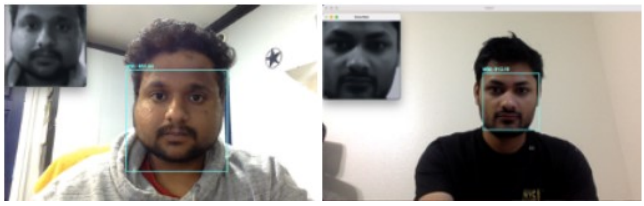


Figure 52. Model detection in Live video

snapshots from live video are taken for the 5 demographics (*African, Indian, Asian, Caucasian, and American*) which can seen *Figure 53* gives us the real and obfuscated images.

Figure 53. Real vs obfuscated images from Live video

The *Indian* demographic has the highest MAE (17.71), indicating that the difference between the original and obfuscated images is the largest among all demographics. The *Caucasian* demographic has the lowest MAE (13.20), meaning that the difference between the original and obfuscated images is the smallest among all demographics. Similar to the MAE, the *Indian* demographic has the highest MSE (583.64), which suggests that the obfuscation has caused significant changes in the pixel values compared to the original images. The *Caucasian* demographic has the lowest MSE (346.10), which means the obfuscation has the least impact on the pixel values for this demographic. The *Caucasian and American* demographics have the highest SSIM values (0.83), which means that the structural similarity between the original and obfuscated images is the highest among all demographics. This suggests that the obfuscation process has preserved the structure of the images well for these demographics. The *African and Indian* demographics have the lowest SSIM values (0.80), indicating that the structural similarity between the original and obfuscated images is the lowest among all demographics. The *Caucasian* demographic has the highest PSNR (23.24), meaning that the quality of the obfuscated images is the highest compared to other demographics. The*Indian* demographic has the lowest PSNR (20.81), which suggests that the quality of the obfuscated images is the lowest among all demographics.

Demographic	*MAE*	*MSE*	*SSIM*	*PSNR*
African	15.34	429.35	0.80	21.91
Asian	16.86	507.18	0.81	21.40
Caucasian	13.20	346.10	0.83	23.24
Indian	17.71	583.64	0.80	20.81
American	14.22	362.02	0.83	22.68

Table 28. Metrics

Based on the evaluation metrics in the provided table, it appears that there might be a bias in the obfuscation process for different demographic groups. The bias becomes evident when we analyze the differences in the MAE, MSE, SSIM, and PSNR values for each demographic.

The Indian demography has the highest Mean Absolute Error (MAE) and Mean Squared Error (MSE) values, which may indicate that the obfuscation process is changing the pixel values for this demographic more significantly than for other demographics. As a result, it's possible that the quality of obfuscated images will decrease or that the obfuscation method will be more easily recognized. On the other side, the MAE and MSE values of the Caucasian demographic are least affected, suggesting that the obfuscation process is more successful for this group.

Additionally, the demographic disparity is evident in the Structural Similarity Index Measure (SSIM) values. The highest SSIM values are found in the Caucasian and American demographics, indicating that these groups have the greatest structural similarity between the original and obfuscated images. The SSIM scores for Indian and African demographics are the lowest, indicating that the obfuscation process may not be as effective at preserving the images' structural integrity for these populations.

The Peak Signal-to-Noise Ratio (PSNR) figures lastly offer additional proof of probable bias. The demography that is Caucasian has the highest PSNR, which suggests that the quality of the obscured images is the highest of all the demographics. The Indian demography, on the other hand, has the lowest PSNR score, indicating that the quality of the obscured images is the lowest among all the demographics.

These disparities in the evaluation metrics across demographics may suggest the presence of a bias in the obfuscation process. The underlying cause of this bias could be rooted in the algorithm's design, the training data, or both. For instance, if the algorithm was primarily trained on images belonging to a specific demographic, it might not generalize well to other demographics. This could result in a higher error rate and lower image quality for underrepresented groups, as evidenced by the evaluation metrics.

Addressing such bias is crucial to ensure fairness and effectiveness across different demographics. Potential solutions to mitigate the bias could include retraining the algorithm with a more diverse dataset, adjusting the obfuscation process to account for demographic-specific features, or utilizing additional pre-processing techniques to improve the algorithm's performance for underrepresented demographics.

8 Conclusion

DP-Pix aims to assess the performance of obfuscation methods and face recognition systems across various demographic groups and datasets. The results indicated that the success rate of obfuscation varies across different demographic groups and datasets. For instance, the success rate for Asian and Black individuals in the *DemogPairs dataset[5]* followed a sigmoid curve, while it increased stepwise for White individuals. Similarly, the success rates in the *BFW dataset*[9] were low for most groups, except for Indian Females and White Males, who experienced a sudden growth in obfuscation success rate at specific confidence thresholds. The study also revealed that confidence thresholds played a critical role in obfuscation success rates, particularly for certain demographic groups like White and Indian individuals. Additionally, face recognition systems struggled to accurately recognize individuals from different photos belonging to the same demographic group, especially for White and Indian individuals, possibly due to the similarities in facial features or characteristics within these groups. The observed disparities in obfuscation success rates and face recognition performance across different demographic groups raise concerns about fairness and effectiveness, as they may lead to biases in the application of such technologies and have consequences on individuals from specific demographic groups. *DP SAMP* obfuscation technique aims to protect the privacy of individuals in face recognition systems. The method's performance varies depending on the confidence threshold, with higher thresholds resulting in more effective obfuscation. Demographic and gender-related variations also affect the method's effectiveness, emphasizing the need for accounting for diversity in evaluating privacy-enhancing techniques. In the context of the BFW dataset, *DP SAMP*'s success rate is relatively low for both males and females at lower confidence thresholds, but it increases as the threshold increases. Optimizing the confidence threshold is essential to strike a balance between accuracy and error rates in face recognition systems while providing effective privacy protection. Overall, *DP SAMP* has the potential to provide privacy protection at higher confidence thresholds, but further research is necessary to address disparities and optimize the method's performance

The *DP Snow* method provides valuable insights into the performance and vulnerabilities of facial recognition systems across different datasets. The method highlights the importance of understanding obfuscation success rates, false negative rates (FNR), and true false rates (TFR) when analyzing the performance of facial recognition systems. Obfuscation success rate follows a sigmoid-like curve across various demographic groups, increasing when the confidence threshold is higher. Some demographic groups, such as Asian individuals, experience higher obfuscation success rates at lower confidence thresholds, potentially due to distinct facial features or characteristics of these groups. False negative rates increase at higher confidence thresholds as the system becomes more conservative in its decision-making, leading to the rejection of genuine matches and a rise in false negatives. The true false rate plot reveals that facial recognition systems' accuracy varies when recognizing individuals from different demographic groups and photos, underscoring the need to optimize these systems' balance between accuracy and conservativeness. Overall, these findings emphasize the importance of refining facial recognition systems to improve their accuracy and robustness, particularly when dealing with diverse datasets and various demographic groups, to mitigate the impact of obfuscation attacks and enhance the performance of facial recognition technology.

9 Limitations & Future Work

The current study on DP Pix, DP SAMP, and DP SNOW methods provides valuable insights into the performance of obfuscation techniques and face recognition systems across

various demographic groups and datasets. However, there are limitations and areas for future research:

Demographic Disparities: The study reveals disparities in obfuscation success rates and face recognition performance across different demographic groups. Future work should focus on investigating the underlying causes of these disparities and developing strategies to address them, ensuring fairness and effectiveness in the application of such technologies.

Confidence Thresholds: Confidence thresholds play a critical role in obfuscation success rates and face recognition performance. Further research is needed to optimize these thresholds for different demographic groups, striking a balance between accuracy and error rates while providing effective privacy protection.

Feature Analysis: The current study suggests that similarities in facial features or characteristics within certain demographic groups may lead to face recognition systems struggling to accurately recognize individuals. Future work should delve into a more detailed feature analysis to understand how these similarities impact system performance and develop methods to improve the accuracy of recognition for diverse datasets.

Robustness against Obfuscation Attacks: The study highlights the need to refine facial recognition systems to improve their robustness, particularly when dealing with diverse datasets and various demographic groups. Further research should explore techniques that can mitigate the impact of obfuscation attacks and enhance the overall performance of facial recognition technology.

Evaluation Metrics: The study focuses on obfuscation success rates, false negative rates (FNR), and true false rates (TFR) as evaluation metrics. Future work may benefit from incorporating additional metrics or developing new evaluation methods to provide a more comprehensive understanding of the performance of obfuscation techniques and face recognition systems.

Generalizability: The findings in this study are based on specific datasets (DemogPairs and BFW). Additional research on other datasets and different face recognition systems is required to confirm the generalizability of these findings and ensure the conclusions drawn are applicable to a wider range of technologies and situations.

The goal of the current study is to assess how well obfuscation methods operate in a controlled environment. Future studies could examine how these methods are used in realistic circumstances, including privacy-preserving image sharing or facial recognition systems, to gauge their usefulness and efficacy. Additionally, recognizing potential biases in the obfuscation procedure is a crucial initial step; nevertheless, additional study is required to create strategies for minimizing these biases. Future research could examine methods for modifying the obfuscation process to be more equal across demographic groups or look into the effects of various data distributions for the training set on the performance of the model.

References

[1] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* 9, 3–4 (aug 2014), 211–407. https://doi.org/10.1561/0400000042

[2] Liyue Fan. 2018. Image Pixelization with Differential Privacy. In *Database Security*.

[3] Liyue Fan. 2019. Differential Privacy for Image Publication.

[4] Oran Gafni, Lior Wolf, and Yaniv Taigman. 2019. Live Face De-Identification in Video. arXiv:1911.08348 [cs.LG]

[5] Isabelle Hupont Torres and Carles Fernández. 2019. DemogPairs: Quantifying the Impact of Demographic Imbalance in Deep Face Recognition. 1–7. https://doi.org/10.1109/FG.2019.8756625

[6] Brendan John, Ao Liu, Lirong Xia, Sanjeev Koppal, and Eakta Jain. 2020. Let It Snow: Adding Pixel Noise to Protect the User's Identity. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 43, 3 pages. https://doi.org/10.1145/3379157.3390512

[7] Megvii. [n. d.]. Face++. https://www.faceplusplus.com/

[8] Dominick Reilly and Liyue Fan. [n. d.]. A Comparative Evaluation of Differentially Private Image Obfuscation. *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)* ([n. d.]). https://doi.org/10.1109/TPSISA52974.2021.00009

[9] Joseph Robinson. 2022. Balanced Faces in the Wild. https://doi.org/10.21227/nmsj-df12

[10] Han Wang, Shangyu Xie, and Yuan Hong. 2020. VideoDP: A Flexible Platform for Video Analytics with Differential Privacy. *Proceedings on Privacy Enhancing Technologies* 2020 (10 2020), 277–296. https://doi.org/10.2478/popets-2020-0073

[11] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. 2019. Racial Faces in-the-Wild: Reducing Racial Bias by Information Maximization Adaptation Network. arXiv:1812.00194 [cs.CV]

Acknowledgments

Firstly, we would like to thank Dr. Shirin Nilizadeh, our professor in this course who gave us the oppurtunity to work on this project and also we would like to express our gratitude to, Mr. Sadegh Moosavi, who guided us throughout this project.

A Datasets

A.1 Racial Faces in the Wild (RFW)

Total 40,609 images and 11,430 different identities.

Race	No. of Identities	No. of Images
African	2995	10,416
Asian	2492	9688
Caucasian	2959	10,916
Indian	2984	10,908

Table 29. RFW Metadata

A.2 DemogPairs

Total 10,800 images and 600 different identities.

Race	No. of Identities	No. of Images
Asian_females	100	1800
Asian_males	100	1800
Black_females	100	1800
Black_males	100	1800
White_females	100	1800
White_males	100	1800

Table 30. DemogPairs Metadata

A.3 Balanced Faces in the Wild (BFW)

Total 20,000 images and 800 different identities.

Race	No. of Identities	No. of Images
Asian_females	100	2500
Asian_males	100	2500
Black_females	100	2500
Black_males	100	2500
White_females	100	2500
White_males	100	2500
Indian_females	100	2500
Indian_males	100	2500

Table 31. BFW Metadata

B CNN Metrics Values Plots

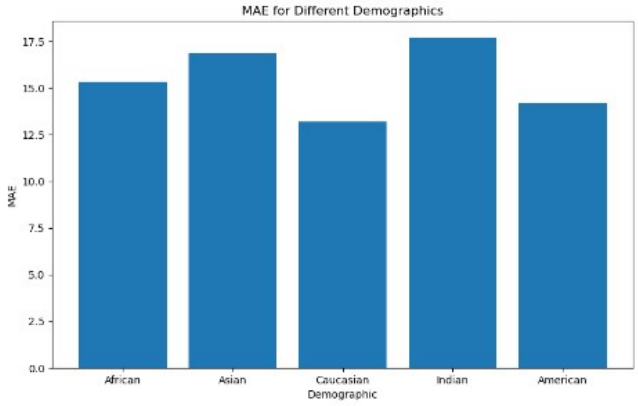


Figure 54. Mean Absolute Error

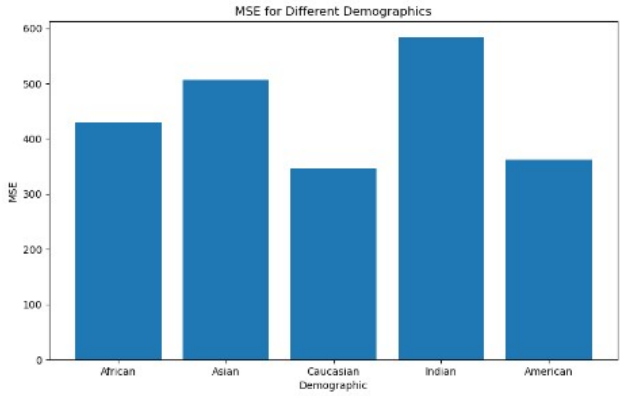


Figure 55. Mean Squared Error

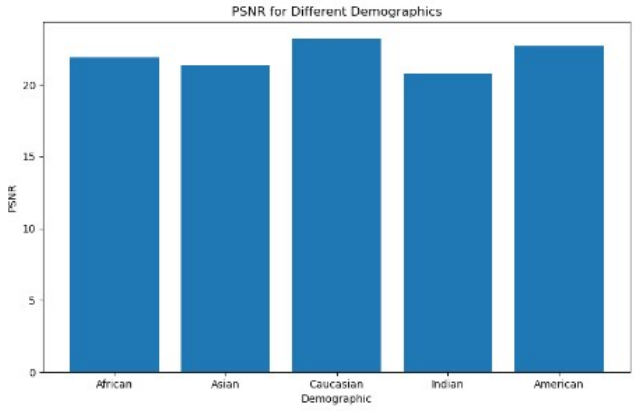


Figure 56. Peak Signal-to-Noise Ratio

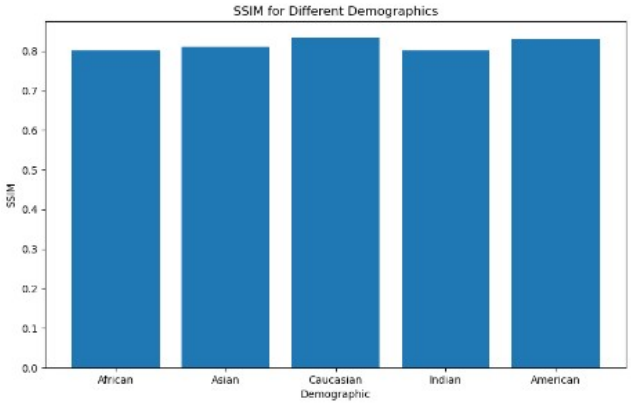


Figure 57. Structural Similarity Index Measure