# Evaluation of Gender Bias in Masked Face Recognition with Deep Learning Models

Mustafa Atay
Department of Computer Science
Winston-Salem State University
Winston-Salem, NC USA
ataymu@wssu.edu

Megh Poudyel
Department of Computer Science
Winston-Salem State University
Winston-Salem, NC USA
mpoudyel119@rams.wssu.edu

Saul Evora
Department of Computer Science
Winston-Salem State University
Winston-Salem, NC USA
sevora120@rams.wssu.edu

***Abstract-* We explore gender bias in the presence of facial masks in automated face recognition systems using various deep learning algorithms in this research study. The paper focuses on an experimental study using an imbalanced image database with a smaller percentage of female subjects compared to a larger percentage of male subjects and examines the impact of masked images in evaluating gender bias. The conducted experiments aim to understand how different algorithms perform in mitigating gender bias in the presence of face masks and highlight the significance of gender distribution within datasets in identifying and mitigating bias. We present the methodology used to conduct the experiments and elaborate the results obtained from male only, female only, and mixed-gender datasets. Overall, this research sheds light on the complexities of gender bias in masked versus unmasked face recognition technology and its implications for real-world applications.**

***Keywords—* gender, bias, fairness, masked, face recognition, deep learning**

## I. INTRODUCTION

Automated facial recognition has been used for various tasks such as user identification, user authentication, gender classification and facial expression recognition. Federal and state government offices such as law enforcements, homeland security, customs control, transportation security administration, courts, and so many others utilize automated facial recognition systems. These automated systems are also used by work places for employee tracking or by schools for keeping track and recording attendance. Due to the broad and critical use of automated facial recognition systems, perfect accuracy of such systems is utmost important all the time.

It is reported in the media that a number of people wrongfully arrested by the law enforcement forces due to false facial recognition matches. The aggrieved people are mainly reported to be among the people of color. In August 2023, a pregnant black mother was wrongfully detained due to mistaken identity because of a false positive match by an automated face recognition system [1].

Several researchers studied and evaluated demographic bias including gender bias in automated facial recognition systems [2-6]. Majority of the researchers reported a degree of demographic bias in facial recognition systems which is said to be mainly originated from the imbalanced datasets.

Mask wearing has become very common and mandated at some public places especially after the start of Covid-19 pandemic. Faces with masks adversely affected the accuracy of automated facial recognition systems which are usually trained with unmasked face images. Thus, mask wearing introduced a challenge for face recognition systems [7] besides the challenges introduced due to demographic biases such as gender bias.

It is a fact that both demographic bias and mask wearing are challenges to overcome in automated face recognition systems. Although several researchers studied gender bias in face recognition [2-6], no research study has elaborated impact of gender in the presence of facial masks in automated face recognition systems, to our knowledge. We aim to address this research problem in this study.

We conduct an experimental study to evaluate gender bias in masked face recognition using six selected Deep Learning (DL) models which are VGG16, AlexNet, GoogleNet, LeNet, FaceNet and ResNet50. We analyze accuracies, F1 scores and gender-based miss rates of the selected DL models using male only, female only, and mixed gender datasets. We use the MaskTheFace [8] software to generate synthesized masks for the masked counterparts of the male only, female only, and mixed gender datasets for training and testing the models. We then conducted three experiments for masked face recognition and three for unmasked face recognition with those datasets for each DL model.

We aim to address the followings with our research study:

- Does there exist any degree of gender bias in masked or unmasked face recognition?
- If so, how does gender bias in masked face

recognition compare to the one in unmasked face recognition?

- Which models excel and mitigate the gender bias most, if bias exists?
- Which models suffer from the gender bias most and degrade, if bias exists?

The rest of the paper is organized as follows: Section II gives a brief overview of the related work while Section III describes the deep learning models and the image database used in the experimental study. The methodology of the experimental study is introduced in Section IV and the experimental results are presented in Section V. Section VI concludes the paper and shares the future work plans.

## II. RELATED WORK

Impact of demographic factors such as race, ethnicity, gender and age are studied by several researchers in the literature. To the best of our knowledge, there is no study to shed light on the impact of gender in face recognition in the presence of facial masks.

In a recent study, researchers conducted experiments using machine learning algorithms on unmasked and masked images. Authors did not study impact of gender. They reported that out of all machine learning algorithms, LR (Logistic Regression) performed the best while DT (Decision Tree) perform poorly with masked faces. The accuracy was higher for unmasked images compared to masked images [7].

Gender bias in facial recognition is elaborated in [2]. The study is conducted with five machine learning algorithms (LDA, LR, SVC, DT and KNN) and with three datasets. They reported a visible gap of miss rates between female and male subjects. Authors did not evaluate impact of wearing facial masks in this study.

Researchers deal with images that have poor resolutions and illuminations which leads to difficulties on facial recognition in [9]. LBPH (Learning Binary Patterns Histogram) is used for not only identifying faces on images, but tightly controlling the environment of the images, especially the illuminations. The clearer the images, the more likely the faces in those images could be identified. They elaborate neither gender nor masked face recognition in this study.

In another study [3], authors elaborate gender bias and demographic unfairness while focusing on face presentation attacks which involves spoofed faces. This study uses ResNet50 and VGG16 in its experiments and leads to a conclusion that the gender bias was found to be not significant since the male and female subjects had similar performances. They did not elaborate masked face recognition in this study.

In some research work, researchers study gender bias along with another demographic attribute such as race or ethnicity. Authors report in [4] that there is significant bias against subjects with darker skins, especially darker skinned females in the tested systems. The impact of mask wearing is not studied in this work.

Age and gender bias towards pedestrians is studied in [5]. The authors report that they are able to mitigate gender and age bias using Multi-Task Convolution Neural Network (MTCNN) in [6]. Neither one of these studies elaborated facial coverings in face recognition unlike our study in this paper.

## III. PRELIMINARIES

We use six different deep learning (DL) algorithms with one facial image database (ORL) for conducting our experimental study. We briefly explain the selected DL algorithms and the image database in the followings.

### *A. Deep Learning Algorithms*

AlexNet, VGG, ResNet, LeNet, GoogLeNet, and FaceNet are all different convolutional neural network (CNN) architectures designed for various computer vision tasks. Here are the key differences between each of them:

- *LeNet* – A simple CNN designed for handwritten digit recognition. It consists of two convolutional layers followed by max-pooling layers and fully connected layers to classify digits [10].
- *AlexNet* – A deep convolutional neural network (CNN) model that revolutionized computer vision tasks. Its key function is to perform visual object recognition [10].
- *VGG (Visual Geometry Group)* – Is known for its uniform and straightforward architecture. It uses 3x3 convolutional filters, max-pooling, and fully connected layers, achieving competitive results on various computer vision tasks [10].
- *ResNet (Residual Network)* – It addresses the vanishing gradient problem in deep networks by introducing residual blocks. It can be much deeper than previous architectures, leading to better performance and easier training [10].
- *GoogLeNet* – It introduced inception modules, which use multiple filter sizes within the same layer, reducing the number of parameters while capturing features at different scales [10].
- *FaceNet* – It is designed to identify and verify a person based on a photograph of their face. FaceNet achieved state-of-the-art results on various face recognition benchmark datasets at the time of its release [11].

### *B. The ORL Image Database*

We picked ORL database for this study. ORL stand for Our Database of Faces. The database was used for a face recognition project at the Cambridge University Engineering Department [12]. There are 41 distinct subjects and ten different images of each subject. The distribution of genders in this database is not balanced. There are 36 male and 5 female subjects.

All of the images were taken in a controlled environment against a dark homogeneous background with the subjects primarily in an upright frontal position. The image files are in PGM format. Each image file is in gray scale with a size of 92x112 pixels [13].

## IV. METHODOLOGY

Our study focuses on digging deeper into DL models to see if they have any gender bias while recognizing masked male and female faces. We perform our experiments with subjects wearing mask and as well as without wearing mask. We divide the database into three different subsets, namely, male only (Ma), female only (Fe), and mixed gender (MG) having both male and female subjects included. We then split each subject's images into separate folders to create training, testing, and validation datasets for each subset of images. We use MaskTheFace software to synthesize masked faces out of the training, validation, and testing datasets making six different datasets for each subset to include the masked versions of those images. We have created a total of 18 datasets for our experiment including masked and unmasked images for all three groups.

Our three subsets consist of 36 males for Ma group, 5 females for Fe group, and 10 subjects with selected 5 males and 5 females for MG . For each subject in each subset, we have 10 images. Out of the 10 images, we use the first 8 for training, the last image for testing and the remaining one for the validation. We repeat this step for the masked images.

Before we begin our experiments, we start hyper tuning the models' parameters, namely, epoch number and batch size to optimize the models' performance. An epoch is a single pass through the entire training dataset. During training, the dataset is divided into several batches, and each batch is used to update the model's weights. Once all batches have been processed, one epoch is completed. The number of epochs determines how many times the model will go through the entire dataset during training. Batch size refers to the number of data samples processed in each iteration (forward and backward pass) of the training process. Instead of updating the model's weights after each individual data point, batches are used to efficiently parallelize the computations and make use of hardware optimizations. We use PyCharm IDE to develop and run our project code.

Table 1 illustrates different datasets built from ORL database with their description. We have a total of 18 datasets including 6 validation datasets.

TABLE I. DATASETS FROM ORL DATABASE

| DATASETS | Description | # of Subjects | # of Images/Subject | Total # of Images |
|---|---|---|---|---|
| Training_Ma_UM | Unmasked male training images | 36 | 8 | 288 |
| Training_Ma_M | Masked male training images | 36 | 8 | 288 |
| Training_GM_UM | Unmasked mixed-gender training images | 10 | 8 | 80 |
| Training_GM_M | Unmasked mixed-gender training images | 10 | 8 | 80 |
| Training_Fe_UM | Masked female training images | 5 | 8 | 40 |
| Training_Fe_M | Masked female training images | 5 | 8 | 40 |
| Testing_Ma_UM | Unmasked male testing images | 36 | 1 | 36 |
| Testing_Ma_M | Masked male testing images | 36 | 1 | 36 |
| Testing_GM_UM | Unmasked mixed-gender testing images | 10 | 1 | 10 |
| Testing_GM_M | Masked mixed-gender testing images | 10 | 1 | 10 |
| Testing_Fe_UM | Unmasked female testing images | 5 | 1 | 5 |
| Testing_Fe_M | Masked female testing images | 5 | 1 | 5 |

The names of the datasets with UM extension are for unmasked and with M extension are for masked face images that are contained in the datasets. We perform 6 experiments for each one of 6 DL algorithms resulting in 36 experiments in total. These experiments are based on training the model with unmasked images, then validating it with unmasked images itself, and finally testing it with unmasked images for each one of our 3 subsets. Similarly, its counterpart version is training with masked images, validating it with masked images, and testing it with masked images.

*Experiment 1: Training with unmasked and testing with unmasked images for Male only group.*

We performed these experiments by training 6 DL models using ORL database to observe the performance of unmasked face recognition with DL models when the system is completely trained with male only (Ma) unmasked images of 36 subjects. We used 8 unmasked images for training, 1 unmasked image for validation for each 36 subjects and then tested each of the DL models with 36 unmasked images, 1 for each one of 36 individuals.

*Experiment 2: Training with masked and testing with masked images for Male only group.*

We performed these experiments by training 6 DL models using ORL database to observe the performance of masked face recognition with DL models when the system is completely trained with male only (Ma) masked images of 36 subjects. We used 8 masked images for training, 1 masked image for validation for each one of 36 subjects and then tested each of the DL models with 36 masked images, 1 for each one of 36 individuals.

*Experiment 3: Training with unmasked and testing with unmasked images for mixed gender group (MG).*

We performed these experiments by training 6 DL models using ORL database to observe the performance of unmasked face recognition. The system is trained with mixed gender (MG) group including unmasked images of 10 subjects with 5 males and 5 females. We used 8 unmasked images for training, 1 unmasked image for validation for each one of 10 subjects and then tested each of the DL models with 10 unmasked images, 1 for each one of 10 individuals.

*Experiment 4: Training with masked and testing with masked images for mixed gender group (MG).*

We performed these experiments by training 6 DL models using ORL database to observe the performance of masked face recognition with DL models when the system is trained with mixed gender (MG) group including masked images of 10 subjects with 5 males and 5 females. We used 8 masked images for training, 1 masked image for validation for each one of 10 subjects, and then, tested each of the DL models with 10 masked images, 1 for each one of 10 individuals.

*Experiment 5: Training with unmasked and testing with unmasked images for Female only group.*

We performed these experiments by training 6 DL models using ORL database to observe the performance of unmasked face recognition with DL models when the system is trained with female only (Fe) unmasked images of 5 subjects. We used 8 unmasked images for training, 1 unmasked image for validation for each one of 5 subjects and then tested each of the DL models with 5 unmasked images, 1 for each one of 5 individuals.

*Experiment 6: Training with masked and testing with masked images for Female only group.*

We performed these experiments by training 6 DL models using ORL database to observe the performance of masked face recognition with DL models when the system is completely trained with female only (Fe) masked images of 5 subjects. We used 8 masked images for training, 1 masked image for validation for each one of 5 subjects and then tested each of the DL models with 5 masked images, 1 for each one of 5 individuals.

## V. EXPERIMENTAL RESULTS

We created three different group of subsets out of original ORL database, namely, Male only (Ma), Female only (Fe) and Mixed Gender (MG) groups. Male only dataset consists of 36 male subjects, Female only dataset consists of 5 female subjects, and Mixed Gender group consists of 10 subjects with 5 female and 5 male subjects. For each subset, we created 6 datasets that includes training, validation, and testing groups for both masked and unmasked face images. For the Male only (Ma) group, the datasets are Training_Ma_UM, Training_Ma_M, Validation_Ma_UM, Validation_Ma_M, Testing_Ma_UM, and Testing_Ma_M. For the Female only (Fe) group, we have Training_Fe_UM, Training_Fe_M, Validation_Fe_UM, Validation_Fe_M, Testing_Fe_UM, and Testing_Fe_M. Similarly, for the Mixed Gender (MG) group, we have Training_MG_UM, Training_MG_M, Validation_MG_UM, Validation_MG_M, Testing_MG_UM, and Testing_MG_M. These datasets are illustrated in the Table I.

We selected 6 deep learning (DL) algorithms which are used in face recognition studies. These algorithms are VGG16, AlexNet, GoogleNet, LeNet, FaceNet, and ResNet50. We trained, validated, and tested these selected DL models. The experiments were performed in PyCharm environment. The results are recorded using accuracy, precision, recall an F1 scores [14] and reported with only accuracy score in the following tables and charts due to the space limitation.

Table II. THE EXPERIMENTAL RESULTS SHOWING ACCURACY OF 6 DL MODELS WITH ORL

| Accuracy Table | Experiments | | | | | | AVERAGES |
|---|---|---|---|---|---|---|---|
| DL Algorithm | Male/UM | Male/M | MG/UM | MG/M | Female/UM | Female/M | |
| VGG16 | 100% | 97% | 100% | 90% | 100% | 100% | 98% |
| AlexNet | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| GoogleNet | 97% | 100% | 100% | 70% | 100% | 100% | 95% |
| LeNet | 97% | 97% | 60% | 60% | 100% | 100% | 86% |
| FaceNet | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| ResNet50 | 97% | 92% | 100% | 100% | 80% | 40% | 85% |
| AVERAGES | 99% | 98% | 93% | 87% | 97% | 90% | 94% |

TABLE III. THE EXPERIMENTAL RESULTS SHOWING F1 SCORES OF 6 DL MODELS WITH ORL DATABASE

| F1-Score Table | Experiments | | | | | | AVERAGES |
|---|---|---|---|---|---|---|---|
| DL Algorithm | Male/UM | Male/M | MG/UM | MG/M | Female/UM | Female/M | |
| VGG16 | 100.00% | 98% | 100% | 90% | 100% | 100% | 98% |
| AlexNet | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| GoogleNet | 97% | 100% | 100% | 70% | 100% | 100% | 95% |
| LeNet | 97% | 97% | 60% | 60% | 100% | 100% | 88% |
| FaceNet | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| ResNet50 | 97% | 92% | 100% | 100% | 80% | 40% | 85% |
| AVERAGES | 99% | 98% | 93% | 87% | 97% | 90% | 94% |

### A. Experiments with Male Only Group

As shown in Table II for the male only group without mask, three out of six used deep learning algorithms, namely, VGG16, AlexNet, and FaceNet, show excellent performance with an accuracy of 100%, while GoogleNet, LeNet and ResNet50 showed an accuracy of 97%. The average accuracy across all DL models is 99%. These results are shown in Figure 1.

The performance of masked face recognition for Male only group slightly degrades when compared to its unmasked counterpart. However, the only differences are that VGG16 degraded by 3%, GoogleNet enhanced by 3%, LeNet remains at 97%, and ResNet50 degraded by 5%. The average accuracy across all DL models is 98%. These results are shown in Figure2.

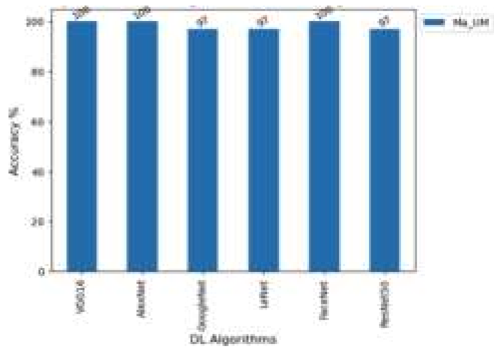

Fig. 1. Accuracies of the DL models on unmasked male dataset

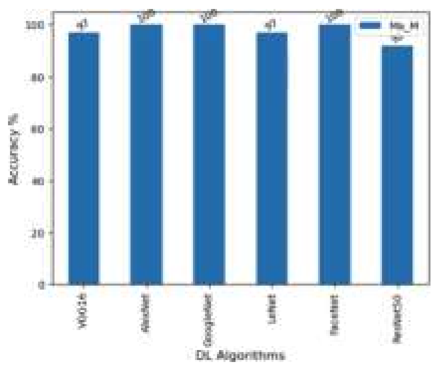

Fig. 2. Accuracies of the DL models on masked male dataset

TABLE IV. MISS RATES FOR MALE ONLY GROUP

| Miss Rate Table | Dataset: Male Only | | | |
|---|---|---|---|---|
| DL Algorithm | Group | Misses | Out Of | Percent |
| VGG16 | Unmasked | 0 | 36 | 0% |
| | Masked | 1 | 36 | 3% |
| AlexNet | Unmasked | 0 | 36 | 0% |
| | Masked | 0 | 36 | 0% |
| GoogleNet | Unmasked | 1 | 36 | 3% |
| | Masked | 0 | 36 | 0% |
| LeNet | Unmasked | 1 | 36 | 3% |
| | Masked | 1 | 36 | 3% |
| FaceNet | Unmasked | 0 | 36 | 0% |
| | Masked | 0 | 36 | 0% |
| ResNet50 | Unmasked | 1 | 36 | 3% |
| | Masked | 3 | 36 | 8% |

TABLE V. OVERALL MISS RATE FOR MALES

| | |
|---|---|
| Average Unmasked Miss Rate | 1% |
| Average Masked Miss Rate | 2% |

According to the Table IV, models trained with masked faces experienced higher miss rates than the unmasked ones. AlexNet and FaceNet both performed the best towards both unmasked and masked subjects, while ResNet50 degraded the most. The AlexNet and FaceNet models seemed to be robust models with or without masks for male subjects.

### *B. Experiments with Female Only Group*

As shown in Table II, the Female only group without mask shows very high performance in all first five deep learning algorithms with 100% accuracy except the ResNet50 algorithm which performed at 80% accuracy. The bar graph in Figure 3 illustrates these results.

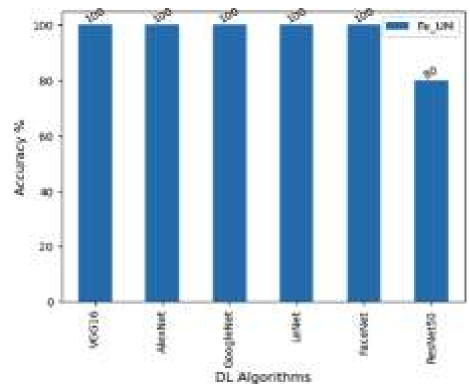

Fig. 3. Accuracies of the DL models on unmasked female dataset

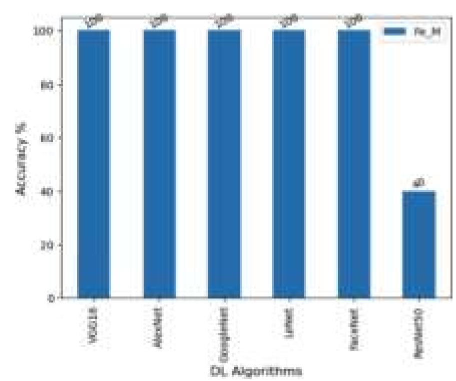

Fig. 4. Accuracies of the DL models on masked female dataset

The results for the masked female only group shows the same as the unmasked Female only group for the first five algorithms except ResNet50 that heavily degraded by half of the original amount from 80% to 40% accuracy. These results are displayed in Figure 4.

The Table VI shows that the first five deep learning algorithms perform much better with zero miss rates than ResNet50 algorithm which performed poorly with lesser accuracies for both unmasked and masked datasets by 20% and 60% respectively. This shows that all five deep learning algorithms mitigate impacts of both gender and masked face recognition, while ResNet50 is struggling with both masked and unmasked face recognition for female only group.

TABLE VI. MISS RATES FOR FEMALE ONLY GROUP

| Miss Rate Table | Dataset: Female Only | | | |
|---|---|---|---|---|
| DL Algorithm | Group | Misses | Out Of | Percent |
| VGG16 | Unmasked | 0 | 5 | 0% |
| | Masked | 0 | 5 | 0% |
| AlexNet | Unmasked | 0 | 5 | 0% |
| | Masked | 0 | 5 | 0% |
| GoogleNet | Unmasked | 0 | 5 | 0% |
| | Masked | 0 | 5 | 0% |
| LeNet | Unmasked | 0 | 5 | 0% |
| | Masked | 0 | 5 | 0% |
| FaceNet | Unmasked | 0 | 5 | 0% |
| | Masked | 0 | 5 | 0% |
| ResNet50 | Unmasked | 1 | 5 | 20% |
| | Masked | 3 | 5 | 60% |

TABLE VII. OVERALL MISS RATE FOR FEMALES

| | |
|---|---|
| Average Unmasked Miss Rate | 3% |
| Average Masked Miss Rate | 10% |

### C. Experiments with Mixed Gender Group

As shown in Table II, the Mixed Gender (MG) group without mask shows that VGG16, AlexNet, GoogleNet, FaceNet and ResNet50 all performed perfect at 100% accuracies while LeNet performed poorly at 60% accuracy. Figure 5 illustrates these results in a bar graph.

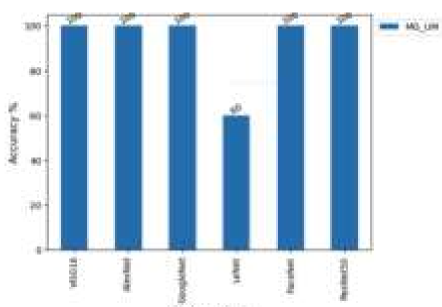


Fig. 5. Accuracies of the DL models on unmasked mixed gender dataset

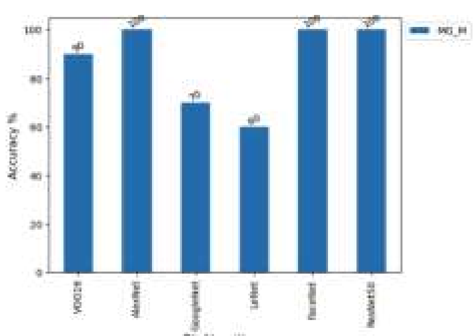


Fig. 6. Accuracies of the DL models on masked mixed gender dataset

The result for masked Mixed Gender group shows that three deep learning algorithms, namely, AlexNet, FaceNet, and ResNet50 all performed excellent with 100% accuracy without being impacted from masks while the other two were degraded. VGG16 degraded down to 90% and GoogleNet down to 70% accuracies. LeNet stayed at 60% accuracy. These results are illustrated in Figure 6.

TABLE VIII. MISS RATES FOR MIXED GENDER GROUP

| Miss Rate Table I | Dataset: Mixed Gender | | | |
|---|---|---|---|---|
| DL Algorithm | Group | Misses | Out Of | Percent |
| VGG16 | Unmasked | 0 | 10 | 0% |
| | Masked | 1 | 10 | 10% |
| AlexNet | Unmasked | 0 | 10 | 0% |
| | Masked | 0 | 10 | 0% |
| GoogleNet | Unmasked | 0 | 10 | 0% |
| | Masked | 3 | 10 | 30% |
| LeNet | Unmasked | 4 | 10 | 40% |
| | Masked | 4 | 10 | 40% |
| FaceNet | Unmasked | 0 | 10 | 0% |
| | Masked | 0 | 10 | 0% |
| ResNet50 | Unmasked | 0 | 10 | 0% |
| | Masked | 0 | 10 | 0% |

TABLE IX. OVERALL MISS RATE FOR MG

| | |
|---|---|
| Average Unmasked Miss Rate | 7% |
| Average Masked Miss Rate | 13% |

TABLE X. OVERALL SEPARATE MISS RATE PER GENDER IN MG

| | |
|---|---|
| Average Unmasked Female Miss Rate | 7% |
| Average Masked Female Miss Rate | 17% |
| Average Unmasked Male Miss Rate | 7% |
| Average Masked Male Miss Rate | 10% |

As shown in Table VIII, AlexNet, FaceNet and ResNet50 all performed perfect towards mixed gender dataset by having no miss rates for both unmasked and masked subjects while LeNet performed poorly by having 40% miss rate for both masked and unmasked subjects.

In Tables V, VII, and IX, we observe that the models suffer from masked face recognition and give more miss rates for masked face recognition than for unmasked face recognition. This brings up some attention to hyper tuning these DL algorithms when we use them for masked face recognition.

In the accuracy Table II, we observe that the average performance for Male only (Ma) group exceeds the average performance for Female only (Fe) group by 2% in unmasked face recognition. Similarly, we see that the average performance for Male only (Ma) group exceeds the average performance for Female only (Fe) group by 8% in masked face recognition. These differences in performance indicate that these DL models when used for recognizing female faces are not as effectively performing as they do for recognizing male faces. In other words, there are some signs of gender bias in both unmasked and masked face recognition systems which is more apparent with masked face recognition.

The Table X shows overall separate miss rates for male and female in Mixed Gender (MG) group. We observe that in masked face recognition, female subjects have more miss rate than male subjects by 7%. However, both male and female subjects have equal number of miss rates for unmasked face recognition.

Table XI for average miss rates of all 3 groups shows that unmasked subjects have lesser miss rates than masked subjects by 5% in general. We see that female subjects have more miss rates than male subjects by 4.8%.

TABLE XI. DL MISS RATES FOR ALL DATASETS

| DL Algorithms Overall Miss Rates for All Datasets | Male | MG | Female | Average |
|---|---|---|---|---|
| Average Unmasked Miss Rate | 1.4% | 7% | 3% | 4% |
| Average Masked Miss Rate | 2.3% | 13% | 10% | 9% |
| Averages | 1.9% | 10.0% | 6.7% | |

## VI. CONCLUSIONS

This study is aimed to evaluate gender bias issues in masked face recognition using deep learning algorithms. We analyze accuracies, F1 scores and miss rates of various DL models using male only, female only, and mixed gender datasets for both masked and unmasked face recognition.

Overall results show that the masked only face recognition performance degrades considerably when compared to unmasked only face recognition. We observe that while female subjects have overall 7.7% more miss rates than male subjects in masked face recognition, the difference of miss rates between female and male subjects still exists but reduced to 1.6% in unmasked face recognition. These findings reveal that there are indications of bias against female subjects in face recognition models which becomes higher and more visible in the presence of masked face recognition.

In our study, we employed a facial image database with photos taken in a controlled environment to train the CNN models for masked face recognition. We acknowledge the gender distribution imbalance within the tested database particularly with fewer female images. To address this imbalance, we plan to incorporate a larger and more diverse database to balance the training datasets with higher number of images and equal gender distribution. Furthermore, we intend to incorporate genuine images of users with masks to reduce potential sources of error. In future research, we also aim to study a comprehensive assessment of bias across multiple demographics like race and ethnicity enriching our study's depth.


## ACKNOWLEDGMENT

This research is funded by NSF Award #1900087. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.


## REFERENCES


[1] Eight Months Pregnant and Arrested After False Facial Recognition Match, retrieved at August 6, 2023. https://www.nytimes.com/2023/08/06/business/facial-recognition-false-arrest.html

[2] M. Atay, H. Gipson, T. Gwyn and K. Roy, "Evaluation of Gender Bias in Facial Recognition with Traditional Machine Learning Algorithms," 2021 IEEE Symposium Series on Computational Intelligence (SSCI), Orlando, FL, USA, 2021, pp. 1-7.

[3] Norah Alshareef, Xiaohong Yuan, Kaushik Roy, and Mustafa Atay. 2021. "A Study of Gender Bias in Face Presentation Attack and Its Mitigation" Future Internet 13, no. 9: 234.

[4] A. Krishnan, A. Almadan and A. Rattani, "Understanding Fairness of Gender Classification Algorithms Across Gender-Race Groups," 2020 19th IEEE ICMLA, Miami, FL, USA, 2020, pp. 1028-1035.

[5] Brandao, Martim. “Age and gender bias in pedestrian detection algorithms.” ArXiv abs/1906.10490 (2019).

[6] Das, Abhijit, Antitza Dantcheva and François Brémond. “Mitigating Bias in Gender, Age and Ethnicity Classification: A Multi-task Convolution Neural Network Approach.” ECCV Workshops (2018).

[7] M. Pudyel and M. Atay, "An Exploratory Study of Masked Face Recognition with Machine Learning Algorithms," IEEE SoutheastCon 2023, Orlando, FL, USA, 2023, pp. 877-882.

[8] Anwar, Aqeel, and Arijit Raychowdhury. "Masked face recognition for secure authentication." arXiv preprint arXiv:2008.11104 (2020)

[9] T. Ahonen, A. Hadid and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 12, pp. 2037-2041, Dec. 2006.

[10] Z. Li, F. Liu, W. Yang, S. Peng and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 12, pp. 6999-7019, Dec. 2022.

[11] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 815-823.

[12] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," Proceedings of 1994 IEEE Workshop on Applications of Computer Vision, Sarasota, FL, USA, 1994, pp. 138-142, doi: 10.1109/ACV.1994.341300.

[13] El-Sayed, M.A. and Hamed, K. (2015) Study of Similarity Measures with Linear Discriminant Analysis for Face Recognition. Journal of Software Engineering and Applications, 8, 478-488.

[14] Goutte, C., Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation, in Lecture Notes in Computer Science 3408:345-359.