Evaluation of Gender Bias in Facial Recognition with Traditional Machine Learning Algorithms

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

Hailey Gipson
Department of Computer Science
Winston-Salem State University
Winston-Salem, NC USA
hgipson118@rams.wssu.edu

Tony Gwyn
Department of Computer Science
NC A&T State University
Greensboro, NC USA
tgwyn@aggies.ncat.edu

Kaushik Roy
Department of Computer Science
NC A&T State University
Greensboro, NC USA
kroy@ncat.edu

***Abstract*—The prevalent commercial deployment of automated facial analysis systems such as face recognition as a robust authentication method has increasingly fueled scientific attention. Current machine learning algorithms allow for a relatively reliable detection, recognition, and categorization of face images comprised of age, race, and gender. Algorithms with such biased data are bound to produce skewed results. It leads to a significant decrease in the performance of state-of-the-art models when applied to images of gender or ethnicity groups. In this paper, we study the gender bias in facial recognition with gender balanced and imbalanced training sets using five traditional machine learning algorithms. We aim to report the machine learning classifiers which are inclined towards gender bias and the ones which mitigate it. Miss rates metric is effective in finding out potential bias in predictions. Our study utilizes miss rates metric along with a standard metric such as accuracy, precision or recall to evaluate possible gender bias effectively.**

Keywords—facial recognition, machine learning, gender, bias, fairness, race, equality, inclusivity, diversity

I. Introduction

Automated gender classification has essential applications in many domains, such as demographic research, law enforcement, online advertising, and human-computer interaction. Recent research has questioned the fairness of this technology across gender and race. Specifically, several studies raised the concern of higher error rates of the face-based gender classification system for darker-skinned people like African Americans and women [12].

Several studies found that the commercial face gender classification systems all perform better on male and light faces, mainly caused by the biases in the data set. This bias can lead to inconsistent model accuracy, limit the applicability of face analytic systems to non-white race groups, and adversely affect research findings based on such skewed data [11].

Pedestrian detection algorithms are essential components of mobile robots, such as autonomous vehicles, which directly relate to human safety [10]. Performance disparities in these algorithms could translate into disparate impacts in the form of biased accident outcomes. More and more research is coming out every day about the problematic biases in computer systems [17], artificial intelligence [18], and robotics [19]. These numerous studies, along with the ethical and social studies of disparate impact, the nature of algorithm discrimination, and concrete algorithm audits, have shown a fairness and justice dimension of algorithms.

An example of these problems is the several commercial computer vision systems like Microsoft, IBM, or Face++, which have been criticized for their asymmetric accuracy across sub-demographics in recent studies. The studies with these commercial systems reported that the accuracies dropped down for dark-skinned female faces [3]. Biases in their training data mainly cause this difference. Various unwanted biases in image datasets can quickly occur due to limited selection, capture, and negative sets.

A proposed solution is to divide facial images into local regions for better recognition rates, and texture descriptors are extracted independently from each region. The descriptors are then concatenated to form a global description of the face. This approach is described in [1].

This study will evaluate the effectiveness of the facial recognition algorithms to see how well they can accurately recognize images of different genders in balanced and imbalanced training sets. The purpose of this study is to evaluate the gender bias in facial recognition using traditional machine learning algorithms and report our findings. We tested a large group of traditional machine learning algorithms in this study. Miss rates metric is effective and to the point and in finding out potential bias in predictions. Our study utilizes miss rates metric along with a standard metric such as accuracy, precision or recall to evaluate possible gender bias effectively.

II. Related work

There have been several sources of errors within the automated facial recognition systems across different gender, race, and age groups [15]. The latest report on gender classification released by National Institute for Standards and Technology reflects those algorithms performed worse for females than males [16].

Sources of errors in automated face recognition algorithms are generally attributed to the well-studied variations in pose, illumination, and expression collectively known as PIE. Other factors such as image quality, time-lapse, and occlusion also contribute to face recognition errors. Previous studies have also shown that specific cohorts are more susceptible to errors in the face matching process within a specific demographic group. However, there has yet to be a comprehensive study

This research is funded by NSF Award #1900087.

investigating whether or not we can train face recognition algorithms to exploit knowledge regarding the demographic cohort of a probe subject. This study presents a large-scale analysis of face recognition performance on three different demographics, one being gender. For each of the demographics, the study had the performance of three different types of systems. Those being, three commercial off-the-shelf (COTS) face recognition systems (FRS), face recognition algorithms that do not utilize training data, and a trainable face recognition algorithm. The study was enabled by a collection of over one million mug shot face images from the Pinellas County Sheriff's Office, but a total of 102,942 images were used in this study. As far as the results for the gender demographic, the performance was consistent in that it exhibited lower recognition accuracies on the female cohort. The results of this study should motivate the design of algorithms that specifically target different demographic cohorts within the race/ethnicity, gender, and age demographics [2]. While authors report the performance differences of COTS FRS's due to demographically imbalanced datasets, they do not disclose the underlying machine learning models in [2]. Unlike [2], we present our findings with five traditional machine learning models and compare their performances for demographically imbalanced and balanced datasets.

Automated facial analysis (FA) includes many applications, including face detection, visual attribute classification such as gender and age prediction, and actual face recognition. Among other visual attributes, gender is an essential demographic attribute. Automated gender classification has drawn significant interest in numerous applications such as surveillance, human-computer interaction, anonymous customized advertisement system, and image retrieval neural network. The organization of the visual cortex inspires the connectivity pattern between its neurons that have learnable weights and biases. Buolamwini and Gebru evaluated the fairness of the gender classification system using three commercial SDKs from Microsoft, Face++, and IBM on Pilot Parliaments Benchmark (PPB) developed by the authors. The CNN models used include VGG, ResNet, InceptionNet, and Network Implementation and Fine-tuning. The datasets used include the UTKFace and FairFace [13] datasets. They investigated the source of bias of the gender classification algorithms across gender-race groups. Experimental investigations with architectural differences suggested that algorithms with architectural differences may vary in performance even when trained on race and gender-balanced sets. Therefore, the bias of the gender classification system is not due to a particular algorithm [3]. We evaluate impact of gender imbalance in facial recognition unlike the study in [3] where they focus on gender imbalance in gender classification problem.

III. Preliminaries

We briefly describe the Machine Learning (ML) algorithms and the image dataset we will use in the experimental study in the following.

A. Machine Learning Algorithms

Classification algorithms play an integral role with the facial recognition algorithm, helping to categorize the images and determine their relationship to each other. Our proposed study uses a total of five different traditional classification methods that are found in the literature to enhance the function. They include Support Vector Classifier (SVC), Linear Discriminant Analysis(LDA), K-Nearest Neighbors (KNN), Decision Trees (DT), and Logistic Regression (LR).

- *Support Vector Classifier (SVC)* – Support Vector Classifier is an image classifier designed for binary problems and can also be extended to handle multi-class problems. SVC ensures high generalization by mapping inputs non-linearly to high-dimensional feature spaces, constructing linear decision surfaces.

Originally, SVC was to have data separable by a hyperplane without error. However, later versions include a 'soft-margin' which allows for a permittableminimal subset of error [4]. In the literature, SVC performed admirably compared to both Decision Tree (DT) and Histogram of Gradients (HoG) [4].

- *Linear Discriminant Analysis (LDA)* – Linear Discriminant Analysis (LDA), as well as its counterpart, Principal Component Analysis (PCA), is a very well-known classification technique. While PCA is an unsupervised technique, therefore not of use to us in our study, LDA is supervised.

LDA is both a dimensionality reduction technique and a linear classifier, but we focus on the latter for our purposes. Linear Discriminant Analysis focuses on projecting the data to maximize class separability. This technique works well for our study due tothe previously mentioned fact that our dataset has a normal distribution.

In the research, LDA is often used for dimensionality reduction. However, many of the papers where LDA was used for image classification did see some rather impressive results, especially with larger sizes of datasets [5].

- *K-Nearest Neighbors (KNN)* – K-nearest neighbors (KNN) is a standard classification method used for data mining and image classification. While KNN can be used for both classification and regression, we focus on KNN classification for our study.

In KNN classification, the output is a class membership. An object is classified by a plurality vote of its nearest neighbors. The object is assigned to the class most common among those k nearest neighbors. K is a positive integer and typically small; through the research, we have found that setting $k = 3$ or $k = 5$ both give good results.

We found many examples of KNN classification in the literature. Most studies determined that KNN, when paired with a reasonable k value, produced good accuracy results [6]. As for our testing, KNN performed worse than DT and SVC.

- *Decision Trees (DT)* – Decision Trees are classifiers that are represented by a flowchart-like structure. Decision Trees are unlike Support Vector Classifiers and neural networks, as they do not make statistical assumptions concerning the inputs or scale the data.

DT models have a structure similar to a tree, where data is broken down into smaller subsets at each branch. In the research, Decision Trees had a reasonably respectable success rate, falling just short of the accuracy attained by SVC [4].

- *Logistic Regression (LR)* – Logistic Regression is used to model the probability of a specific class or classes existing. In the literature, Logistic Regression has been shown to have impressive accuracy rates with training and testing images. This approach is only made more impressive because the program used reduced image sizes when making these comparisons to cut down on computational space and time [7].

B. Dataset

We use color facial images from the Facial Recognition Technology (FERET) database [8] in this study. The color FERET dataset contains images of 994 subjects. The size of each image is 512x768 pixels. Subjects in the FERET database do not have an equal number of images taken, some less and some more. Moreover, the database does not have an equal distribution of all races and genders, limiting our ability to choose a diverse group of subjects with sufficiently enough images per individual to train.

IV. METHODOLOGY

We pick 24 distinct individuals with at least 13 images from the FERET database for each experiment. Thus, we process 312 images in each one of our experiments. While 12 out of 13 images of a subject are used for training, 1 image of a subject is used for testing. We pick female and male subjects from a diverse group of Caucasians and African Americans. A subset of representative images that we used in the experimentation is shown in Fig. 1.

Fig. 1. Sample Images from the FERET Database

We prepare three different split of training datasets to evaluate the gender bias in facial recognition with various machine learning algorithms. Our training dataset splits are given in Table 1. One version of the training datasets is a gender-balanced set called GBAL with 12 female and 12 male subjects. We compiled two unbalanced training datasets called Female Dominant (FD) and Male Dominant (MD) sets. We use 1/3 over 2/3 split ratio in FD and MD datasets. While there are 16 female and 8 male subjects in the FD training dataset, we include 8 female and 16 male subjects in the MD training dataset. We intend to observe and evaluate the gender bias with the tested Machine Learning algorithms using these balanced and unbalanced training datasets.

TABLE I. TRAINING DATASETS

Label	Description	# Of Female vs. Male Subjects
MD	Male Dominant	8 Females and 16 Males
GBAL	Gender Balanced	12 Females and 12 Males
FD	Female Dominant	16 Females and 8 Males

We use our previously developed facial recognition system [9] for experimentation. The architecture of the used system is given in Fig. 2.

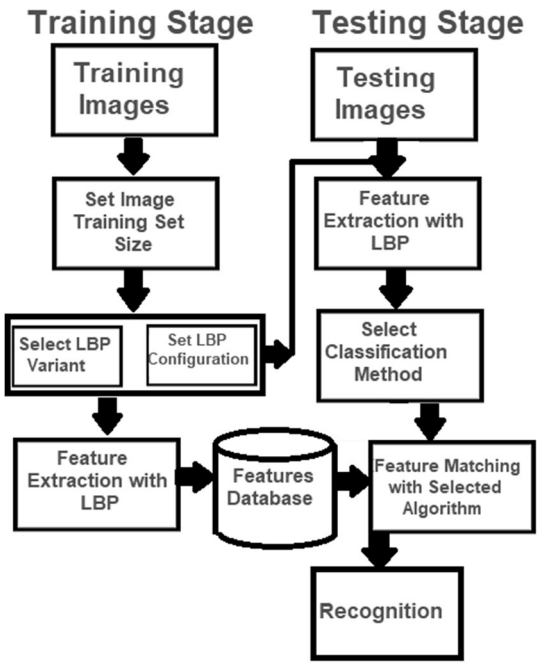

Fig. 2. Facial Recognition System Architecture [9]

This system utilizes the Local Binary Pattern (LBP) algorithm for the feature extraction process and enables users to configure the LBP parameters. LBP converts color images to gray-scale images before feature extraction. We use 12 images per subject for training. We set the radius to 1 and neighborhood size to 16 for the LBP operator. LBP has several variants, and

we use Uniform LBP for the experiments. We picked the optimum LBP configurations based on the observations in [9]. All the experiments are conducted with the 3 training datasets and 6 Machine Learning (ML) classification algorithms which are Support Vector Classifier (SVC), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Decision Trees (DT), and Logistic Regression (LR) and Naïve Bayes (NB). We do not include the NB in detailed and statistical analysis as it is the sole outlier with an underachieved prediction accuracy.

We utilize hit rates and miss rates to evaluate the performance of the tested algorithms with various training datasets. Our in-house facial recognition system can measure the performance of the predictions with four metrics: accuracy, precision, recall, and F1-score. Currently, the system only generates the visualized results using accuracy metric. Therefore, we present the experimental results based on accuracy metrics which are shown in the form of bar graphs in the following section of the paper. Miss rates metric is powerful in finding out potential bias in predictions. We define miss rate as the ratio of false-negatives to the total number of all the objects in a particular group. We observe miss rates for each gender group to evaluate the potential bias besides observing the accuracy. Miss rates are reported in the form of tables in the following section.

It is obvious that any machine learning model needs multiple images of a subject for training and identification. The distribution of races and the number of images per subject in the FERET database limits our ability to choose arbitrarily many subjects of diverse races and genders to test. The dataset includes lesser darker skin subjects than lighter skin subjects. As we aim to keep an equal distribution of darker and lighter skin subjects with the same number of images per subject to train, there is not too much room to grow the size of our experimental dataset. We research and shed light to the issue of gender bias in face recognition with our preliminary experiments using relatively small number of image sets retrieved from FERET database. We plan to work with image databases other than FERET which will give us opportunity to experiment with larger number of image sets such as FairFace [13] as a future work.

V. EXPERIMENTAL RESULTS

We used 3 training datasets and 5 Machine Learning Algorithms in our experiments. As introduced in Section IV, the 3 datasets are GBAL – Gender Balanced, FD – Female Dominant, and MD – Male Dominant. The 5 tested Machine Learning classifier algorithms are Support Vector Classifier (SVC), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Decision Trees (DT), and Logistic Regression (LR) which are introduced in Section III. We used the settings described in Section III for all the experiments. We followed the procedural steps given in Figure 3 to conduct our experimental study.

We recorded the miss rates for each gender after experimenting with every ML classifier algorithm. Our goal with measuring the miss rates includes the following:

i) to observe if the miss rates are balanced for male and female subjects;

ii) if not, to observe the direction of bias;

iii) to observe the gap between the male and female miss rates;

iv) to find out which ML algorithm best mitigates the bias.

1. Select the Training Dataset
2. Select the ML Classifier Algorithm
3. Configure LBP Parameters for Feature Extraction
4. Conduct Experiment and Record the Metrics
5. Repeat Steps 2-4 for each ML Classifier Algorithm
6. Repeat Steps 1-5 for each Training Dataset

Fig. 3. Experimentation Procedure

A. Experimentation with Male Dominant (MD) Training Set

The Male Dominant (MD) training set contains 16 male subjects and 8 female subjects, each with 12 images to train the system. Thus, we trained the system with a total of 288 images. We tested the system with 1 image per subject using 24 images. We processed a total of 312 distinct images in this experiment. We tested the ML classifier algorithms with the MD training set and recorded the performance with the accuracy metric. The graph in Fig. 4 shows the performance of the tested ML algorithms with the MD training set. While the average accuracy across ML algorithms is 88%, DT performed flawlessly with 100% prediction accuracy followed by SVC with 96% prediction accuracy.

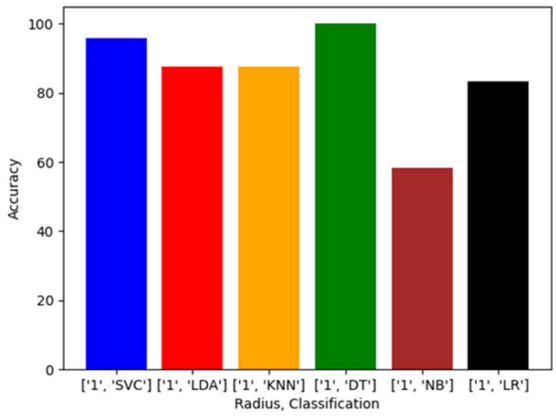

Fig. 4. Prediction Accuracies of ML Algorithms for MD Training Set

The breakdown of miss rates for female and male subjects in the experiment with the MD training set is given in Table 2.

Females' miss rates are either higher or equal in 4 of the tested ML algorithms, whereas the male miss rate is slightly higher in one ML algorithm, SVC. Females miss rates are 25% greater than that of male miss rates with the LDA and LR algorithms which is the most significant margin throughout all the experiments.

Average miss rates for female and male subjects in the experiment with the MD training set are shown in Table 3.

TABLE II. MISS RATES FOR MD TRAINING SET

Miss Rate Table	Dataset / MD			
ML Algorithm	Group	Misses	Out Of	Percent
SVC	F	0	8	0%
	M	1	16	6%
LDA	F	3	8	38%
	M	2	16	13%
KNN	F	1	8	13%
	M	2	16	13%
DT	F	0	8	0%
	M	0	16	0%
LR	F	3	8	38%
	M	2	16	13%

The average females miss rate is 18% which is twice the average male miss rate. The gap between female and male miss rates is 9%. Our findings with the MD training set show the evidence of bias towards male subjects in general.

TABLE III. OVERALL MISS RATES FOR MD TRAINING SET

Overall Female Miss Rate:	**18%**
Overall Male Miss Rate:	**9%**
The Difference of Overall Miss Rates:	**9%**

B. Experimentation with Balanced (GBAL) Training Set

The Gender Balanced (GBAL) training set contains 12 male subjects and 12 female subjects, each with 12 images to train the system. We tested the above ML classifier algorithms with the GBAL training set and recorded the performance with the accuracy metric. The graph in Fig. 5 shows the performance of tested ML algorithms with the GBAL training set. While the average accuracy across ML algorithms is 87%, DT and SVC algorithms performed flawlessly with 100% prediction accuracy.

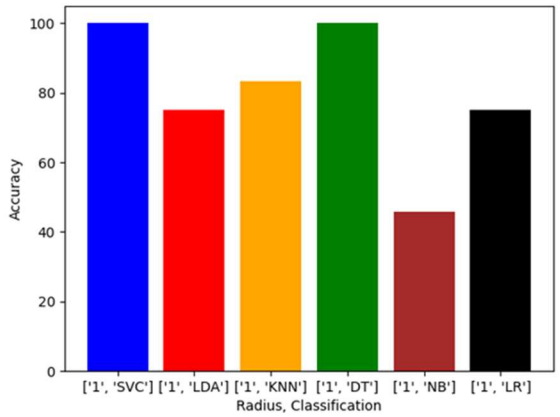

Fig. 5. Prediction Accuracies of ML Algorithms for GBAL Training Set

The breakdown of miss rates for female and male subjects in the experiment with the GBAL training set is given in Table 4. Female and male miss rates are equalized in 4 of the tested ML algorithms, while female miss rates are 16% higher than male miss rates only in the LR algorithm. A balanced dataset seems to mitigate the prediction bias except in the LR algorithm.

Overall miss rates for female and male subjects in the experiments with the GBAL training set are shown in Table 5. Overall female miss rate is 15%, while the overall male miss rate is 12%. The gap between female and male miss rates is reduced to 3% in GBAL experiments. We observed that a balanced dataset considerably reduces the miss rates gap between females and males.

TABLE IV. MISS RATES FOR GBAL TRAINING SET

Miss Rate Table	Dataset / GBAL			
ML Algorithm	Group	Misses	Out Of	Percent
SVC	F	0	12	0%
	M	0	12	0%
LDA	F	3	12	25%
	M	3	12	25%
KNN	F	2	12	17%
	M	2	12	17%
DT	F	0	12	0%
	M	0	12	0%
LR	F	4	12	33%
	M	2	12	17%

TABLE V. OVERALL MISS RATES FOR THE GLOBAL TRAINING SET

Overall Female Miss Rate:	**15%**
Overall Male Miss Rate:	**12%**
The Difference of Overall Miss Rates:	**3%**

C. Experimentation with Female Dominant (FD) Training Set

The Female Dominant (FD) training set contains 16 female subjects and 8 male subjects, each with 12 images to train the system. We tested the above ML classifier algorithms with the FD training set and recorded the performance with the accuracy metric. The graph in Fig. 6 shows the performance of tested ML algorithms with the FD training set. The average accuracy across ML algorithms is measured as 91%. DT algorithm again performed flawlessly at 100% prediction accuracy and followed by SVC with 96% accuracy.

The breakdown of miss rates for female and male subjects in the experiment with the FD training set is given in Table 6. Female miss rates are either higher or equal to the male miss rates in 4 of the tested ML algorithms, while male miss rates are slightly higher than female miss rates only in the SVC algorithm where one male subject is missed.

Overall miss rates for female and male subjects in the experiments with the FD training set are shown in Table 7. While the overall female miss rate is 10%, the overall male miss rate becomes 7.5%. The gap between female and male miss rates is reduced to 2.5% in FD experiments. We observed that the FD

training set with the highest number of female subjects minimizes the miss rates gap between females and males.

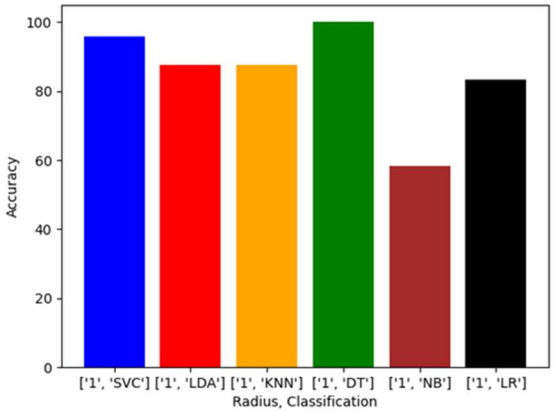

Fig. 6. Prediction Accuracies of ML Algorithms for FD Training Set

TABLE VI. MISS RATES FOR FD TRAINING SET

Miss Rate Table	Dataset / FD			
ML Algorithm	Group	Misses	Out Of	Percent
SVC	F	0	16	0%
	M	1	8	13%
LDA	F	3	16	19%
	M	0	8	0%
KNN	F	2	16	13%
	M	1	8	13%
DT	F	0	16	0%
	M	0	8	0%
LR	F	3	16	19%
	M	1	8	13%

TABLE VII. OVERALL MISS RATES FOR FD TRAINING SET

Overall Female Miss Rate:	**10%**
Overall Male Miss Rate:	**7.50%**
The Difference ofOverall Miss Rates:	**2.50%**

D. Comparison of Results

We put together the experimental results from imbalanced and balanced datasets to observe if there is a pattern in the measured accuracies and miss rates when we increase or decrease the number of female subjects in the training sets. We first analyze the average accuracies across all the training sets for the tested Machine Learning algorithms. The average accuracies across all the experiments are shown in Table 8.

We see that the highest average accuracy is 91% and obtained from the experiments with the Female Dominant (FD) dataset. The highest accuracy is achieved with the FD training set in four of the tested ML algorithms except for the SVC algorithm. The highest accuracy is achieved with the balanced dataset. In general, prediction accuracies either increased or stayed the same but not degraded when we switch from Male Dominant (MD) training set to Female Dominant (FD) training set by increasing the number of female subjects and decreasing the number of male subjects.

The comparison of overall miss rates for female and male subjects across all the training sets for the tested ML algorithms is given in Table 9. The overall female miss rate is highest with 18% in the experiments with Male Dominant (MD) training set, dropped to 15% in gender-balanced (GBAL) training set and minimized to 10% in Female Dominant (FD) training set.

TABLE VIII. AVERAGE ACCURACIES FOR ALL TRAINING SETS

Accuracy Table	Datasets / 24Sbj_12Img		
ML Algorithm	MD	GBAL	FD
SVC	96%	100%	96%
LDA	79%	75%	88%
KNN	88%	83%	88%
DT	100%	100%	100%
LR	79%	75%	83%
AVERAGES	88%	87%	**91%**

We see a clear pattern that the overall female miss rates decrease when we increase the number of female subjects from MD through GBAL and FD training sets. On the other hand, overall male miss rates do not show a peak when the number of male subjects is decreased from experiments with MD training set to FD training set. Moreover, male miss rates are observed at a minimum of 7.5% in the experiments with the FD training set. We observed that the gap between female and male miss rates reached the minimum of 2.5% in the experiments with FD training set where we have a split of 2/3 female and 1/3 male subjects for training.

TABLE IX. OVERALL MISS RATES FOR ALL TRAINING SETS

Miss Rates for Training Datasets	MD	GBAL	FD
Overall Female Miss Rate	18%	15%	10%
Overall Male Miss Rate	9%	12%	7.50%
The Difference of Overall Miss Rates	9%	3%	2.50%

VI. CONCLUSIONS

We observed that we need more female subjects to train traditional ML algorithms for better accuracy. In contrast, the lesser number of male subjects would not adversely affect the accuracies. When we have fewer female subjects in a dataset to train an ML algorithm, the female miss rates reached the maximum value and resulted in the highest gap between female and male miss rates. On the other hand, when we reduce the number of male subjects and increase the number of female

subjects in a dataset to train an ML algorithm, neither male miss rates nor the overall prediction accuracy degrades. Moreover, the highest average accuracies and lowest overall female and male miss rates are achieved in the experiments with the FD training set with more female subjects.

We found that some of the traditional Machine Learning classification algorithms are biased towards female subjects in facial recognition. We see that LR and LDA classification algorithms are vulnerable and biased towards female subjects and degrade when fewer female subjects are in the training sets. On the other hand, DT, SVC, and KNN algorithms mitigate gender bias. They are not adversely affected by the imbalanced MD and FD datasets and show similar prediction performances in either dataset.

Decision Tree (DT) algorithm performed flawlessly with our imbalanced datasets MD and FD as well as with the balanced dataset GBAL. DT models are typically effective and perform relatively better with small datasets and in high signal to noise ratio [20]. We believe that the perfect performance of DT algorithm is due to our small dataset size besides the high signal to noise ratio of the conducted image recognition experiments.

We plan to work with different datasets such as FairFace [13] and DemogPairs [14] and larger number of subjects to investigate gender bias in facial recognition. FairFace and DemogPairs datasets are aimed to mitigate gender and other demographic bias. We focused on evaluating gender bias in traditional Machine Learning algorithms in this study. We identify studying gender bias in facial recognition with advanced deep learning algorithms as future work. In addition, researching racial, ethnic, and age biases in facial recognition is also considered among potential future work.

REFERENCES

[1] 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, doi: 10.1109/TPAMI.2006.244.

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

[3] A. Krishnan, A. Almadan and A. Rattani, "Understanding fairness of gender classification algorithms across gender-race groups", *2020* 19th IEEE International Conference on Machine Learning and Applications (ICMLA), 2020, pp. 1028-1035, doi: 10.1109/ICMLA51294.2020.00167.

[4] T. J. Alhindi, S. Kalra, K. H. Ng, A. Afrin and H. R. Tizhoosh, "Comparing LBP, HOG and deep features for classification of histopathology images", 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1-7, doi: 10.1109/IJCNN.2018.8489329.

[5] S. Pang, S. Ozawa and N. Kasabov, "Incremental linear discriminant analysis for classification of data streams", in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 35, no. 5, pp. 905-914, Oct. 2005, doi: 10.1109/TSMCB.2005.847744.

[6] O. García-Olalla, E. Alegre, M. T. García-Ordás and L. Fernández-Robles, "Evaluation of LBP variants using several metrics and KNN classifiers", SISAP, 2013.

[7] Vanlalhruaia, Y. K. Singh and N. D. Singh, "Binary face image recognition using logistic regression and neural network", 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 3883-3888, doi: 10.1109/ICECDS.2017.8390191.

[8] P. J. Phillips, Hyeonjoon Moon, P. Rauss and S. A. Rizvi, "The FERET evaluation methodology for face-recognition algorithms", Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997, pp. 137-143, doi: 10.1109/CVPR.1997.609311.

[9] T. Gwyn, M. Atay, R. Kaushik and A. Esterline, "Evaluation of local ninary pattern algorithm for user authentication with face biometric", 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), 2020, pp. 1051-1058, doi: 10.1109/ICMLA51294.2020.00170.

[10] Martim Brandao, "Age and gender bias in pedestrian detection algorithms", ArXiv abs/1906.10490, 2019, n. pag.

[11] A. Khalil, S. G. Ahmed, A. M. Khattak and N. Al-Qirim, "Investigating bias in facial analysis systems: a systematic review", in IEEE Access, vol. 8, pp. 130751-130761, 2020, doi: 10.1109/ACCESS.2020.3006051.

[12] A. Das, A. Dantcheva and François Brémond. "Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach", ECCV Workshops, 2018.

[13] K. Kärkkäinen and J. Joo, "FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation", 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1547-1557, doi: 10.1109/WACV48630.2021.00159.

[14] I. Hupont and C. Fernández, "DemogPairs: quantifying the impact of demographic imbalance in deep face recognition", 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 2019, pp. 1-7, doi: 10.1109/FG.2019.8756625.

[15] P. J. Phillips, P. J. Grother, R. J. Micheals, D. Blackburn, E. Tabassi and J. M. Bone, "Face recognition vendor test 2002: evaluation report", National Institute of Standards and Technology (NISTIR), vol. 6965, pp. 1–54, 2003.

[16] P. J. Grother, G. W. Quinn, and P. J. Phillips, "MBE 2010: Report on the evaluation of 2D still-image face recognition algorithms," National Institute of Standards and Technology (NISTIR), vol. 7709, pp. 1–61, 2010.

[17] B. Friedman and H. Nissenbaum, "Bias in computer systems", ACM Transactions on Information Systems (TOIS), 14(3):330–347, 1996.

[18] D. Boyd, K. Levy, and A. Marwick, "The networked nature of algorithmic discrimination", Data and Discrimination: Collected Essays. Open Technology Institute, 2014.

[19] A. Howard and J. Borenstein, "The ugly truth about ourselves and our robot creations: the problem of bias and social inequity", Science and engineering ethics, 24(5):1521–1536, 2018.

[20] Z. Zhou and Y. Jiang, "NeC4.5: neural ensemble based C4.5", in *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 6, pp. 770-773, June 2004, doi: 10.1109/TKDE.2004.11.