computers

MDPI

Article

Evaluating Impact of Race in Facial Recognition across Machine Learning and Deep Learning Algorithms

James Coe * and Mustafa Atay *

Department of Computer Science, Winston-Salem State University, Winston-Salem, NC 27110, USA
* Correspondence: jcoe118@rams.wssu.edu (J.C.); ataymu@wssu.edu (M.A.)

Abstract: The research aims to evaluate the impact of race in facial recognition across two types of algorithms. We give a general insight into facial recognition and discuss four problems related to facial recognition. We review our system design, development, and architectures and give an in-depth evaluation plan for each type of algorithm, dataset, and a look into the software and its architecture. We thoroughly explain the results and findings of our experimentation and provide analysis for the machine learning algorithms and deep learning algorithms. Concluding the investigation, we compare the results of two kinds of algorithms and compare their accuracy, metrics, miss rates, and performances to observe which algorithms mitigate racial bias the most. We evaluate racial bias across five machine learning algorithms and three deep learning algorithms using racially imbalanced and balanced datasets. We evaluate and compare the accuracy and miss rates between all tested algorithms and report that SVC is the superior machine learning algorithm and VGG16 is the best deep learning algorithm based on our experimental study. Our findings conclude the algorithm that mitigates the bias the most is VGG16, and all our deep learning algorithms outperformed their machine learning counterparts.

Keywords: facial recognition; machine learning; deep learning; dataset; bias; race; ethnicity; fairness; diversity

Citation: Coe, J.; Atay, M. Evaluating Impact of Race in Facial Recognition across Machine Learning and Deep Learning Algorithms. *Computers* **2021**, *10*, 113. https://doi.org/10.3390/computers10090113

Academic Editors: Paolo Bellavista and Ana Filipa Sequeira

Received: 10 August 2021
Accepted: 6 September 2021
Published: 10 September 2021

1. Introduction

Many biometrics exist to provide authentication for users while in a public setting [1], such as personal identification numbers, passwords, cards, keys, and tokens [2]. However, those methods can become compromised, lost, duplicated, stolen, or challenging to recall [2]. The acquisition of face data [3] is utilized for verification, authentication, identification, and recognition [4], and has been a decades-old computer vision problem [5]. The ability to accurately interpret a face allows for recognition to confirm an identity, associate a name with a face [5] or interpret human feeling and expression [6]. Facial recognition for humans is an easy task [5], but becomes a complex task for a computer [4] to perform like human perception [5]. Although image analysis in real-time [7,8] is feasible for machines, and significant progress has been achieved recently [9]. Automatic facial recognition remains a difficult task that is challenging, tough, and demanding [2]. Many attempts to improve accuracy in data visualization [3] still reach the same conclusion that artificial intelligence is not equal to human recognition when remembering a small sample size of faces [10], and numerous questions and problems remain [9].

A system needs to collect an image of a face to use as input to compare against a stored or previously recognized image for successful recognition. This step involves many variables that severely impact the capabilities for successful face recognition [4]. Many users want authentication in a public setting and most likely, using a mobile device leads to unconstructed environments [6] and non-controlled changes [11]. These changes lead to limitations on nonlinear variations [11], making data acquisition difficult [1]. This problem has persisted for over fifty years [6] and often contributes to differing results that cannot

be replicated [10]. Further complications include the approach taken because different techniques can yield different results [10]. Some influences that contribute to these problems are the position, illumination, and expression [1–4,8,12,13]. Other influences include pose angles [1,8], camera distance, head size, rotation, angle, and direction [3,5,6]. The introduction of aging, scaling, accessories, occlusion [1,13], and hair [5] makes capturing these varying scales more difficult [14]. Limitations on the image quality such as noise, contrast [15], resolution, compression, and blur [12] also contribute to facial recognition inaccuracies.

Although image collection has some varying problems, machine and deep learning are relatively reliable [16] learning algorithms capable of handling large datasets available for research and development [17]. Plenty of facial recognition algorithm variants exist [18,19], and together these algorithms can improve human capabilities in security, medicine, social sciences [17], marketing, and human–machine interface [20]. These algorithms possess the ability to detect faces, sequence, gait, body, and gender determination [19], but still, trained algorithms can produce skewed results [16,17]. These uneven results often lead to a significant drop in performance [16] that can raise concerns about fairness [20]. With the reputation of companies that use facial recognition at stake, many ethical concerns have been raised because of the underrepresentation of other races in existing datasets [16]. Inconsistent model accuracy limits the applicability to non-white faces [17], and the dataset contributes to demographic bias [16]. Existing databases and algorithms are insufficient for training due to the low variability in race, ethnicity, and cultures [20]. Most datasets are not labeled for ethnicity [13], and unknown performance [20] and controversial biased models are a result of utilizing these datasets [13].

The use of limited models contributes to false matches, low adequacy, fairness, and reliability concerns [20]. Many experiments have observed these results and marked present mostly in faces with a higher melanin presence [16,21]. Convolutional neural networks have improved the capabilities of algorithms [13]. However, there is still much room for group fairness in datasets to mitigate the bias that has existed for decades leading to algorithms suffering from demographical performance bias that provides an imbalance to specific groups of individuals [22]. As racist stereotypes exist, there is a need to avoid mislabeled faces and achieve greater race classification accuracy [23]. Utmost importance should be placed on how human attributes are classified [17] to build inclusive models while considering the need for diversity in datasets [21]. Substantial importance should be emphasized that models need to be diverse due to being trained on large datasets with many variations [13]. The results are algorithms and techniques that mitigate bias [20].

The results include an imbalance for some races and demographical bias against specific ethnicities. Considering these inequalities, we investigate and evaluate racial discrimination in facial recognition across the various machine and deep learning models. The findings and results will allow us to discover existing bias with repeatable observations and if any algorithms outperform others while mitigating any bias.

Continuing previous research in [23], we continue to measure and evaluate the observable bias resulting from utilizing the five traditional machine learning models and techniques. Replicating the initial approaches and datasets used with conventional algorithms, we repeat the previous steps and conduct similar experiments with the three deep learning algorithms. Collecting the results from the deep learning models, we perform identical evaluation and bias measurements. Collecting results for each algorithm used allows us to compare performance, accuracy, and bias to evaluate the efficiency of algorithms. We present our findings in hopes of making a meaningful contribution to the decades-old computer vision problem and facial recognition fields. Our collected contributions are:

- Evaluation of racial bias across five machine learning algorithms using racially imbalanced and balanced datasets.
- Evaluation of racial bias across three deep learning algorithms using racially imbalanced and balanced datasets.
- Evaluation and comparison of accuracies and miss rates between tested algorithms.

- Report the algorithms that mitigate the bias the most.

2. Background

As humans, we can quickly identify a face and recognize an individual within a short time. The idea of facial recognition is done daily, and with minimal effort that we may consider this task easy. As time passes, we may see a face that seems familiar, but may not recall the name of the individual even though we recognize them. The introduction of computers to assist with this computer vision problem allows the capabilities to expand to remember substantially more faces. However, once it seemed easy for a human, it is a much more complicated task for a machine. The initial concept of facial recognition was a straightforward task with an obvious solution. The solution involved obtaining an image and precisely identifying or determining if the image matched against a database of stored images. However, four main obstacles present unique problems for facial recognition to be completed successfully. The first problem is the complexity of the task itself, where the next problem is shown in the influences on the images themselves. The final problematic area for facial recognition lies within the algorithms and datasets. Together these problems combined present a unique problem that is observable in technology that is used every day.

2.1. Complexity

For a human to visually recognize a face, we first must observe a face with our eyes, and this image is introduced to the ocular nerve. The idea is then sent to the brain, where the brain retrieves the name that we associate with that face if it is recallable. Many systems are working in tandem, and we often do not realize the complexity of a task that appears so simple. For a computer to first remember an image, it must convert that image into a data structure to compare against others [24]. In our case, we introduce a traditional photograph that we transform into an array of numbers to represent each image's pixel. After converting the pictures, they are then stored for comparison against additional photos. Calculating values for each pixel is a complex task that can involve thousands or millions of steps to complete depending on image size and quality, especially if you consider that each pixel comprises values representing the red, green, and blue values within itself.

2.2. Influences

Influences are one of the most varying items that severely impact facial recognition. The typical three items referenced in the field are pose, illumination, and expression (PIE). Most conventional images are collected by a photographer using a fixed camera with a seated individual in precisely controlled lighting. However, if we suddenly consider any image as usable, then a candidate for recognition may be standing, seated, or involved in an action. The lighting could be too bright or too dark, depending on the location of the image collection. Considering expression is another barrier because viewing traditional images may have a subject smiling. However, your system would still need to recognize your face if you were in a bad mood or had a different expression such as surprise, shock, or pain. Additional items that present problems are aging, angle, scaling, accessories, and occlusion. As we age, our face has subtle changes that make facial recognition more difficult because the look is slightly different. Using a mobile device to collect images can allow individuals to manage their appearance. Still, there is no guarantee that they will hold the camera perfectly level or at an exact distance away from their face. Accessories such as jewelry, glasses, facial hair, hairstyles, and hair color can support the occlusion of a look, making this complex task even more difficult. During our research, the topic of masks as a new item that covers faces and their impact on systems was discussed. During the pandemic a mask was introduced as a new accessory that covers a considerable amount of a face.

2.3. Algorithms

Once a facial recognition system is in use and the influences and complexity issues have been considered. The next item that impacts the miss rates, hit rates, and the system's accuracy is the algorithms themselves. Algorithms can contain several factors that may obtain results that are not precise. An algorithm is designed to do what it is programmed to do, and it will consistently execute as it is programmed. However, these algorithms are created by humans, and this underlying problem can surface within the intended results of an algorithm. Psychologists have studied this as something that humans instinctively do from childhood. Where humans interact with what they are familiar with or interact with more frequently. This phenomenon is observable in facial recognition algorithms designed in Eastern and Asian countries when compared against American counterparts. An algorithm designed in an Asian country will have some influence from its Asian designer and their daily interaction with other community members. This bias will likely result in a system that performs better for those types of individuals. Especially measuring results within that community when compared to other nations. Using inferior algorithms leads to skewed results, performance degradation, and applicability limitations. Biased algorithms can produce results that are not meaningful when considering other outside contributions. This bias demonstrates the need for careful thought when designing an algorithm to be inclusive for all intents and purposes.

2.4. Datasets

If the algorithm has carefully been planned during implementation, it is a particular assumption that it is only as good as the dataset it utilizes. Most datasets do not have labels for races, ethnicity, and demographics. Training of this type of dataset will yield an inaccurate model. This error would also lead to false matches and poor adequacy. Other obstacles in datasets include the fact that a nation may comprise many different races. Combining a similar race as one nationality complicates the accurate estimation of a specific demographic. For example, Malaysia is a multi-racial composition country with many races that all identify as Malaysian. Our dataset included more than 1000 test subjects, but we could only retrieve approximately twenty Black individuals with more than ten available images. There is an obvious imbalance and underrepresentation of specific races, ethnicities and, nationalities while surveying available datasets for research.

Many datasets exist and are available for research and development [10]. However, facial recognition requires many images of each subject, creating an immediate complication for this type of research. Our study utilizes the FERET [25,26] dataset, a publicly available dataset with access granted from the National Institute of Standards and Technology (NIST). Our selected dataset contained 1208 subjects. However, when examining the available images, the number of subjects with many images was predominantly White. This imbalance immediately impacted our research methods and processes, and the limitations for other races were noticeable. Only a few datasets contain many images of the same individual, predominantly comprised of celebrities [3]. Although many datasets are available, they lack the appropriate labeling and classifications that algorithms need to train and test against successfully. Properly documenting existing datasets would be a never-ending process, and still, the datasets would be insufficient for certain demographic and racial representations. An image revolution or campaign is needed to gather lots of images so that datasets become more diverse and inclusive with representations of all walks of life. Tuned datasets such as FairFace and DemogPairs successfully mitigated existing racial and demographic bias [3,7].

2.5. Machine Learning Algorithms

Machine learning algorithms make informed decisions by parsing data and learning from the collected data [27]. Typically, these techniques are used to examine large databases to extract new information as a form of data mining. Many large companies use machine learning to provide an improved user experience in human–computer interaction. These

improvements may include a recommendation based upon preferences or suggestions based on previous selections [27]. Machine learning algorithms require outside exchange if corrections or adjustments are necessary. A general analogy to envision a machine learning algorithm can be demonstrated by using a flashlight. If we teach the flashlight with a machine learning model, it will understand light and dark. If provided with the word dark, the flashlight will come on. The opposite will occur if the word light is used. However, if presented with a word or phrase it has not learned, such as dim, it will not function and require additional adjustments to accommodate the new expression [27]. Gwyn et al. [23] studied traditional machine learning techniques prompting further research in facial recognition, specifically regarding racial bias. We have chosen Support Vector Classifier (SVC), Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), Decision Trees (DT), and Logistic Regression (LR) algorithms from [23]. These algorithms are used for several applications and the approaches can vary such as regression, clustering, dimension reduction and image classification. For our study, we choose to apply these algorithms using classification methods and do not combine or use them in tandem to produce observable results for each algorithm.

Using SVC, image classification for a multi-class approach is achieved by constructing linear decision surfaces using high-dimensional features mapped by non-linear input [23]. LDA is used as a classifying approach by using a normally distributed dataset and projecting the input to maximize the separation of the classes [23]. Classification utilizing KNN is achieved by assigning classes to the output. The designation is concluded by the overall counts of its nearest cohorts, and the class assignment goes to the class with the most common neighbors [23]. The DT approach yields its classifications as a structure and treats each child or leaf as a subset without statistical reasoning [23]. The LR approach reaches the classification by considering the probability that a class exists where the outcome is that a training image and a test image belong to the same class [23].

2.6. Deep Learning Algorithms

Much like machine learning was a form of data mining, deep learning algorithms are a subset of machine learning that functions similarly, but with differing capabilities [27]. If a deep learning algorithm returns an incorrect prediction, the model itself will correct it internally. These algorithms may also use selections and preferences initially, but once deployed, they learn solely on the model and what it was trained on to predict and correct upon [27]. Vehicles that drive automatically or features on them that make lane corrections, braking, and accident avoidance autonomously are some examples of deep learning algorithms in use. We can continue the flashlight analogy to better assist with understanding deep learning algorithms. The light would not need a keyword to associate with action because it would compute from its model and learn to be on or off when required [27]. For our continuing research, we have chosen AlexNet, VGG16, and ResNet50 deep learning algorithms.

3. System and Experimental Design

The similarities in the construction of our machine learning and deep learning algorithms, as shown in Figure 1, denotes the similarities between the two. Both algorithms follow a similar route for execution and use inputs to test for successful recognition. Both algorithms store the results of their training for future comparison and testing. Still, the most notable differences are the number of iterations or epochs that deep learning uses to learn and perform better. Machine learning algorithms take input and apply the settings and execute a mathematically complex algorithm once.

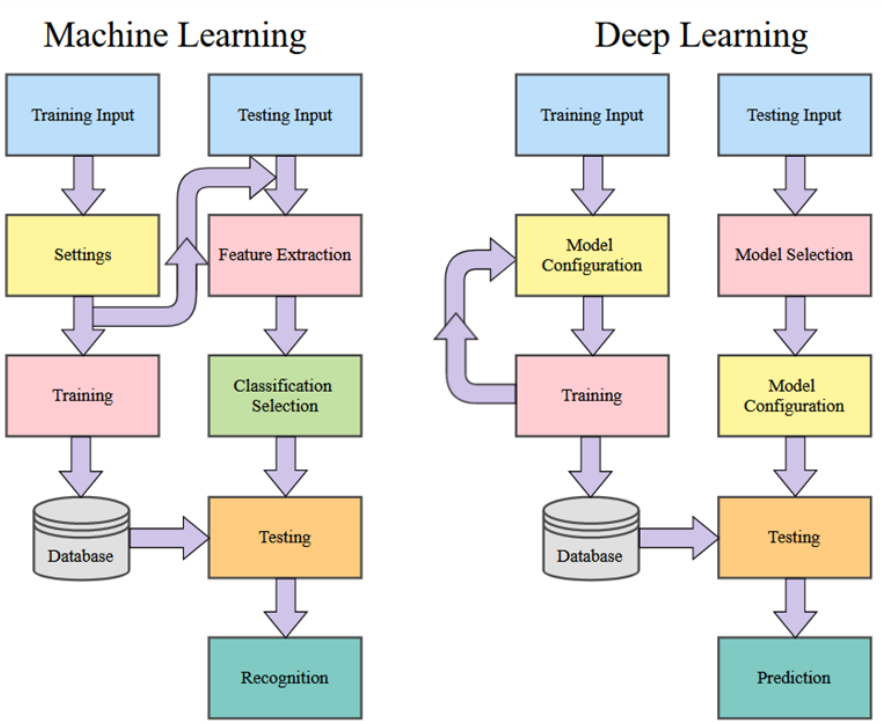

Figure 1. Overview of learning algorithms.

3.1. System Design and Architecture

Our machine learning algorithms take images and use them as input to train the system, as shown in Figure 1. After initialization, the graphical user interface (GUI) collects testing variables, so identical approaches are used. The following steps are for the system to train on inputs of balanced or imbalanced datasets and save them to a datastore. During the testing portion, the system takes the images as input and applies the same setting used for training. The system extracts the features and brings the classification selection for the desired algorithms. The testing phase compares the calculated image to the saved training images to recognize.

A similar approach is used for the deep learning algorithms where input is used for training. The model of the algorithm to be used is constructed. The training iterates or cycles for a selected number of epochs. During the revolutions of each period, the system can validate the training for accuracy and improve predictions on each iteration. Once the training is completed, the datastore is used to save the trained model, as shown in Figure 1.

3.2. System Software

Continuing prior research [23], Python® was the language used, which was a requirement for our continuing research and experimentation. Python is a powerful programming tool that granted us access to many libraries that assisted with implementing this project. Those imported libraries include Pandas, NumPy, OS, Scikit-Learn, Matplotlib, Keras, TensorFlow, CUDA, and cuDNN, as shown in Figure 2.

Pandas is a powerful utility to collect DataFrame information, and NumPy was successful in data structure manipulation. Tools included by importing Scikit-Learn include confusion matrices and built-in metrics reporting. Matplotlib was utilized for feedback and graphing during code execution. The subsequent imports were the game changers for this project. The Keras neural network library has an extensive catalog of tools built on the open-source library of TensorFlow. These high-level Application Programming Interfaces (APIs) assisted in layering, pooling, metric reporting, model construction, and optimization. The addition of the CUDA and cuDNN imports is where the significant drop in runtime occurred. After configuring them properly, they allow the GPU to be utilized

in tandem with the CPU. Overall runtimes significantly drop by more than 75%. Each of our model interpretations is paired with some of these imports and our FERET dataset compilation. The existing code implementations for our experimentation are conducted in Python 3.9.2 using PyCharm Community Edition 2020.3.

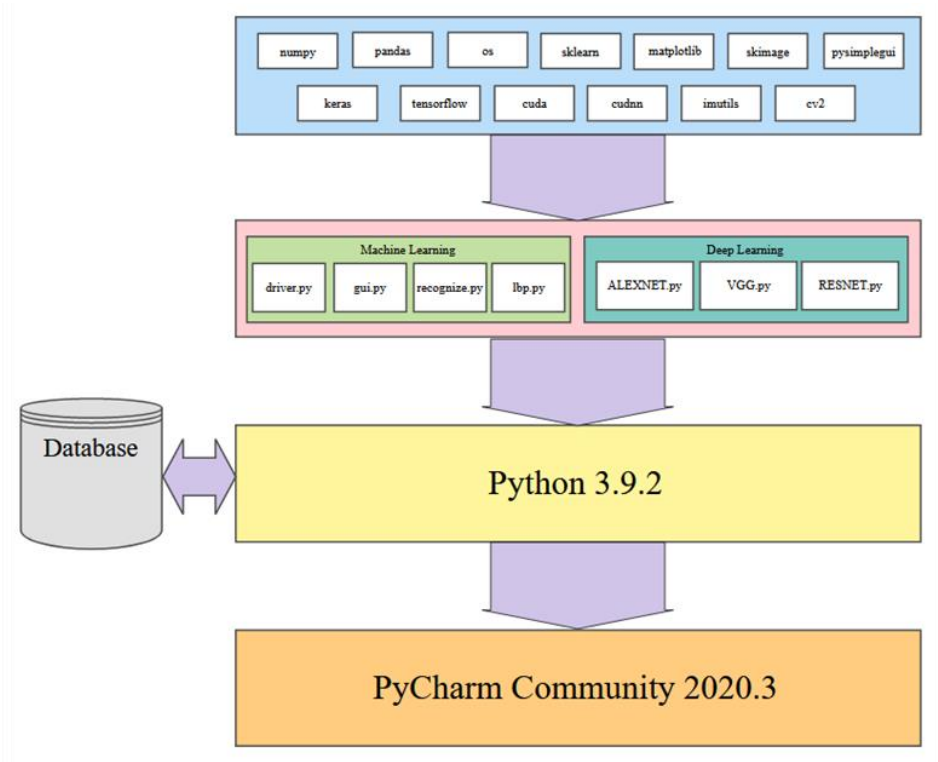

Figure 2. Software architecture.

3.3. Evaluation Plan

Exploring this topic requires three generated datasets and machine learning algorithms to compare performance and results against each. We conducted a plethora of research to collect images and place them into groups that will yield observable results that hopefully will have a meaningful contribution to the facial recognition field. Sampling is sourced from the FERET dataset emphasizing diverse, inclusive images with racial variations. Three datasets are distributed to yield a balance of equal amounts of image representations for each race. The other two sets are distributed to weigh heavier towards a particular race or ethnicity. The datasets are then analyzed with algorithms, and performance is rated on the accuracy of prediction modeling. Additional metric measurements include precision, recall, and F1 scoring. Algorithms and models utilized are Support Vector Classifier (SVC), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Decision Trees (DT), Logistic Regression (LR), AlexNet, VGG16, and ResNet50. In the initial testing on the sample of 24 subjects, we then alter the dataset to be biased towards a particular race. For dominant set 1, we choose 16 topics to represent the Black classification and 8 subjects to describe the White variety. We make a similar comparison for the next phase in experimentation, but this time we complement experiment 2. For dominant set 2, we choose 16 subjects to represent the White and 8 subjects to define the Black classification.

The weighting for each scenario is as follows:

- Balanced—12 Black subjects vs. 12 White subjects;
- Dominant Set 1—16 Black subjects vs. 8 White subjects;
- Dominant Set 2—8 Black subjects vs. 16 White subjects.

Everyone has an available 12 images because of the limitations of our dataset. Initially, we select 11 ideas for those available algorithms to train on and set the remaining image as the image to test against for accuracy. We then remove 3 photos to eliminate influences of

position, illumination, expression, or occlusion to explore accuracy. This setup allows us to test on the original image, but train on only 8 images. To mitigate the limitations of our dataset, we then revisited our initial training set that contained 11 images and augmented them. This method allows us to create additional pictures by inverting the original copies and giving us 22 photos to train.

The dataset configurations for each experimental scenario are:

- Eight images for training and 1 image for testing per subject;
- Eleven images for training and 1 image for testing per subject;
- Twenty-two Images for training and 1 image for testing per subject.

3.4. Evaluation Procedure

The evaluation involves careful consideration to continue previous research conducted in [23]. Initial research performed was explicitly applied to the field of facial recognition in general. Moving forward, we aim to use that research to observe existing bias and identify approaches to mitigate these inaccuracies. Metrics involved in these evaluations will include accuracy, precision, recall, F1 score, hit rates, and miss rates.

Our experimentation involves very precise steps that must be repeated, so the collected results are free from uncontrolled changes. The steps we have followed for each algorithm are:

1. Select the algorithm;
2. Select the dataset;
3. Measure the metrics;
4. Repeat steps 2 and 3 for each dataset with selected algorithm;
5. Repeat entire process for each algorithm.

3.5. Experimental Design

Details regarding machine specifics and methodologies are provided in this section to give an insight into the technology used. These details will serve as a gauge for others to compare to when considering turn-around and runtimes. Our experimentation was performed on a Dell® Alienware® Aurora® Ryzen® machine running a Windows 10® operating system. The Central Processing Unit (CPU) is an AMD® Ryzen® 9 3900XT 12-Core Processor 3.79 GHz. The available memory includes a 512 Gigabytes (GB) Solid State Drive (SSD) and 32 Gigabytes Random Access Memory (RAM). It consists of an NVIDIA® GeForce® RTX 2070 8GB GD DR6 for graphics that significantly contributed to expediting experimentation after installing the required libraries. The Integrated Development Environment (IDE) used was PyCharm Community Edition 2020.3®, a product of JetBrains®. Code implementation for our project was designed using Python 3.9.2 using our IDE. Importing TensorFlow, CUDA and cuDNN allows the machine to use the stream executor to create dynamic libraries that utilize the graphics card for CPU support. The addition of these dynamic libraries significantly reduced the overall runtimes of all algorithms. Before optimizing the configurations, runtimes for the machine learning algorithms took approximately 180 s to complete depending on the number of images utilized. With the optimization configuration, the runtimes are reduced to about 32 s for each selected model. The deep learning algorithms use a sizeable learning epoch, so overall runtimes depend on that setting. Our generalized settings of 50 periods yield runtimes of approximately 18 min, where our optimized settings reduce that time to around 4 min. Taking averages of our three deep learning algorithms produces an average of 7 s per epoch.

4. Experimental Results

The collected results yield a comprehensive overview to visualize performance across the racially imbalanced datasets, much like a sliding scale. Collectively, we document the results for each algorithm in its category and then compare their results between the two types of algorithms.

4.1. Racial Bias across ML Algorithms

Our results began with the execution of our algorithms using the first dataset with eight training images. This approach gives us eight ideas to train on and one image to test for accuracy. We performed the experiments as directed on our balanced dataset to serve as a baseline for comparison. Support Vector Classifier was the superior algorithm of the five machine learning choices. However, the accuracy barely surpassed the 80% mark, as shown in Figure 3, which is well below the industry standard and user expectations.

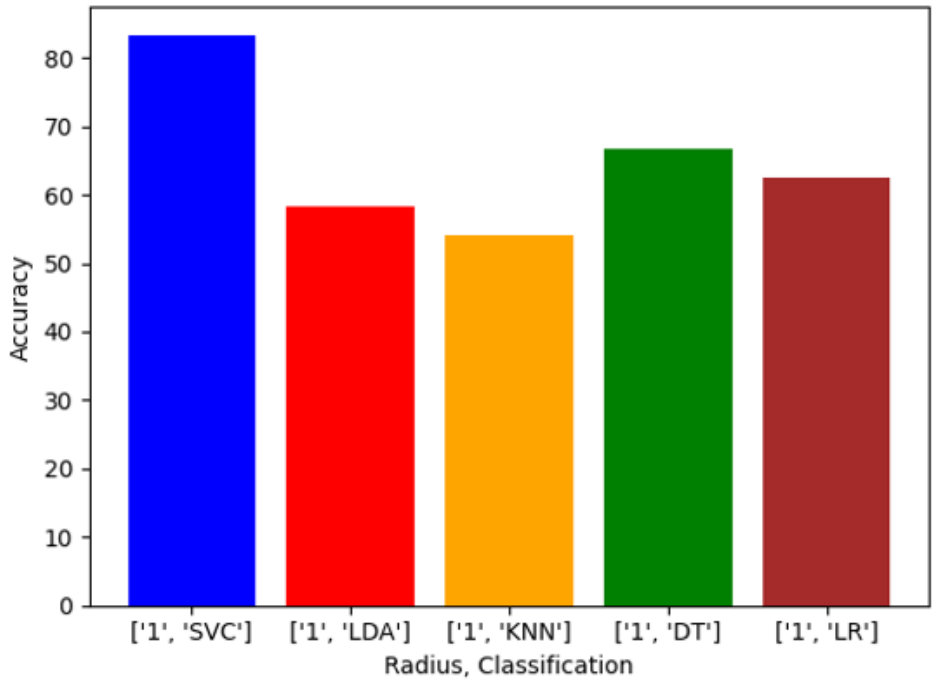

Figure 3. ML accuracies with balanced dataset and 8 training images.

Another notable result is the low level of performance for the other algorithms, especially the K-Nearest Neighbors algorithm barely surpassed the 50th percentile. Applying this experiment to the different datasets demonstrates the importance of the dataset. We repeat the steps for each dataset to measure the differences while exhibiting bias towards a specific race. These types of experiments differ slightly from our baseline results. You can see that the dominant dataset towards Blacks has declining accuracies for Decision Trees and Logistic Regression. The two datasets in Figure 4 yield similar results compared to the White dataset yielding slightly better results when applied to our machine learning algorithms.

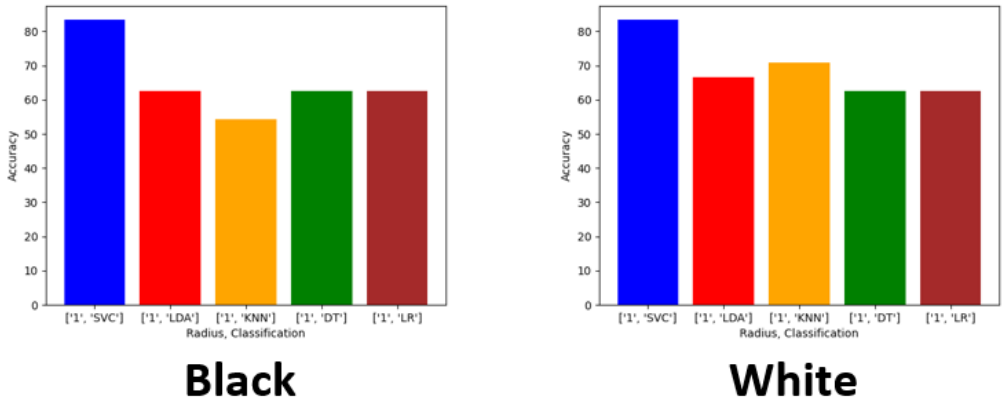

Figure 4. ML accuracies with imbalanced datasets and 8 training images.

The results for this first experiment are not very different, but considering the results already demonstrates a potential problem. We compile the accuracies for each algorithm, and the results show that deviating from a fair and balanced approach can yield unwanted side-effects, as shown in Table 1. The results show that using a dataset that is dominant towards Blacks underperforms compared to its cohort. Specific algorithms perform better than others, but averages demonstrate an unbiased insight that a problem exists.

Table 1. Accuracies of machine learning algorithms using 8 training images.

Accuracy Table	Datasets		
ML Algorithm	**WD**	**RBAL**	**BD**
SVC	83%	83%	83%
LDA	67%	58%	63%
KNN	71%	54%	54%
DT	63%	67%	63%
LR	63%	63%	63%
AVERAGES	**69%**	**65%**	**65%**

Considering the number of miss rates for each race while simultaneously monitoring the dataset type provides a different metric that reinforces the notion that bias in a dataset contributes to uneven results, as shown in Table 2.

Table 2. Average ML miss rates for datasets using 8 training images.

ML Algorithms Miss Rates for Datasets	WD	RBAL	BD
Average Black Miss Rate	32.50%	38.33%	32.50%
Average White Miss Rate	30.00%	31.67%	40.00%
The Difference of Average Miss Rates	2.50%	6.66%	7.50%

Finally, we evaluate the other available metrics of precision, recall, and F1 scoring. A complete scoring metrics table is in [28], and for analysis purposes, we provide the averages in Table 3.

Table 3. Additional ML metrics precision, recall and F1.

ML Metrics Averages	WD	RBAL	BD
Precision	69%	66%	65%
Recall	100%	100%	100%
F1	82%	79%	78%

Across the datasets, average metrics increase and decrease when changing from our balanced baseline dataset. When deviating towards a dataset that is White dominant (WD), the metric averages increase for precision and F1 scores, and the opposite occurs when a Black prevailing dataset is used. The collected results from the experimentation using the first dataset are compiled for further analysis and comparison.

The next portion of our experiments will repeat the steps previously completed for the eight training images. Still, instead, this time, we replace the eight training images with our dataset with 11 pictures to train. The thought for this approach was that the addition of more training images would allow the algorithms to perform better and achieve higher accuracies and performance metrics. We first establish our baseline by using the dataset that has one representation for each racial profile. We utilize the same five machine learning algorithms, and the results are shown in Figure 5.

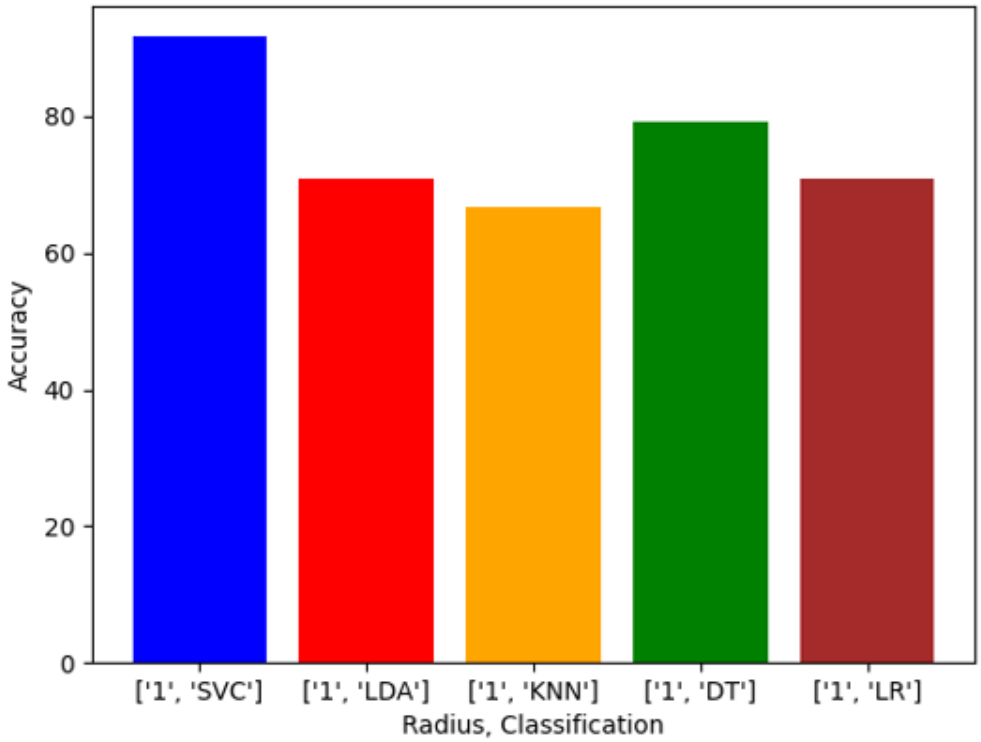

Figure 5. ML accuracies with balanced dataset and 11 training images.

When comparing the results to our previous experiments, the accuracies have improved for all algorithms. Support Vector Classifier is still the top achiever with 92% accuracy, and K-Nearest Neighbors is still the lowest performer at almost 70%. With our baseline established, we can compare the results of each imbalanced dataset and see immediate changes, as shown in Figure 6. The most significant difference shown here is that SVC performs at 96% when paired with the White dataset. All the other algorithms barely achieve approximately 60% when using the Black dataset.

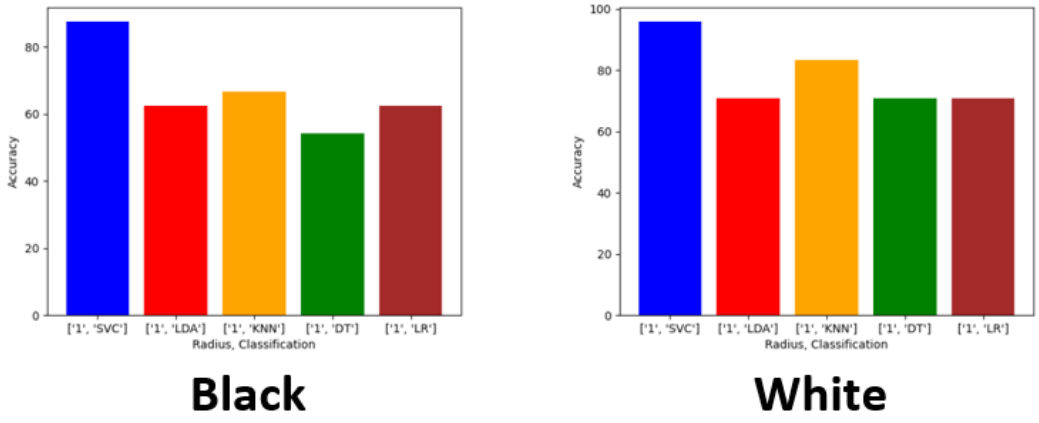

Figure 6. ML accuracies with imbalanced datasets and 11 training images.

In this round of experimentation, we find that the averages for each dataset have increased. However, when using the Black dataset, the standards remain in the 60-percentile range. The balanced dataset and White dominant (WD) dataset improved accuracies to exceed the mid-70-percentile, as shown in Table 4.

Table 4. Accuracies of machine learning algorithms using 11 training images.

Accuracy Table	Datasets		
ML Algorithm	**WD**	**RBAL**	**BD**
SVC	96%	92%	88%
LDA	71%	71%	63%
KNN	83%	67%	67%
DT	71%	79%	54%
LR	71%	71%	63%
AVERAGES	**78%**	**76%**	**67%**

Our miss rate findings indicate an overall decline and are likely due to better training because of additional training images. The miss rate percentages are still well beyond what a user may expect or desire. The gap between them statistically shows bias towards the White dataset, as shown in Table 5.

Table 5. Average ML miss rates for datasets using 11 training images.

ML Algorithms Miss Rates for Datasets	WD	RBAL	BD
Average Black Miss Rate	25.00%	25.00%	33.75%
Average White Miss Rate	20.00%	23.33%	32.50%
The Difference of Average Miss Rates	5.00%	1.67%	−1.25%

Again, we consider additional metrics to gain an insight that can support our initial findings, and again we see that using the averages in Table 6. Using a weighted dataset towards Blacks yields lower performance metrics for accuracy, precision, and F1 scores. Ref. [28] includes the complete scoring metrics table since we use the averages for generalized analysis.

Table 6. Additional ML metrics precision, recall and F1.

ML Metrics Averages	WD	RBAL	BD
Precision	78%	76%	67%
Recall	100%	100%	100%
F1	87%	86%	79%

We attempted to circumvent the limitations presented to us regarding our dataset for our next round of experimentation. We took the initial 11 images for training and augmented them to give us 22 representations to train the algorithms with more images. In this instance, we used the inverse representation of each available image and created a new dataset for training. Again, the expectation is that more training images will yield better performance and lower miss rates. We use the same approach and initially use the balanced dataset to establish a baseline for comparison purposes. We now see that a balanced dataset has achieved 96% accuracy for this iteration, as shown in Figure 7. These baseline results show improvement, but are still lacking accuracies that should be expected in facial recognition.

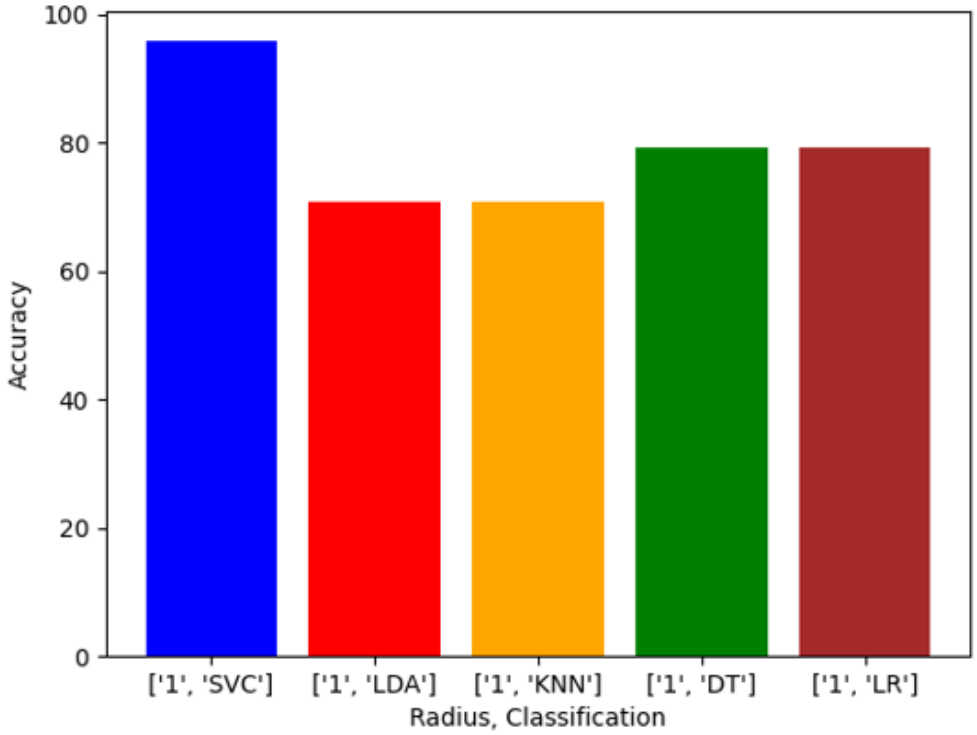

Figure 7. ML accuracies with a balanced dataset and 22 training images.

With our baseline results established, we then move on to compare the results of each imbalanced dataset. Each dataset used was weighted to favor Blacks or Whites, and the expectations are that more training images would yield higher accuracy. This time, we see that using the White dominant (WD) dataset yields better accuracies when compared to the previous experiments. Still, we also note that the Black counterpart has some declines in performance, as shown in Figure 8. Although the graphs appear similar, the y-axis for the White portion has a max value of 100, where the max value for the Black graph is 90.

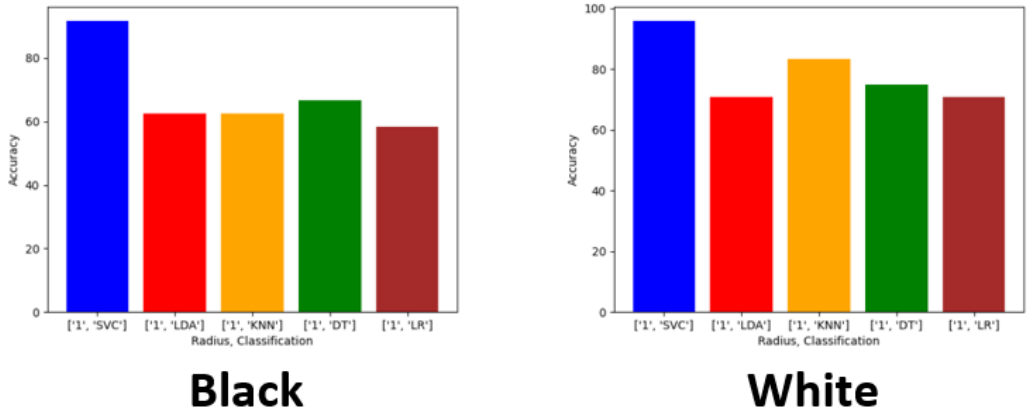

Figure 8. ML accuracies with imbalanced datasets and 22 training images.

We investigate the averages of all the algorithms while considering race, and a noticeable performance gap is shown in Table 7.

This difference is between the Black dominant (BD) dataset and the others, with the other two outperforming by 10%. Observations for this round of experimentation include the accuracy improvements for each dataset, but a worsening divide in performance. This problem is dually observable in the miss rate calculations. Overall, our miss rates have declined, but the margin has increased when considering the differences, as shown in Table 8.

Table 7. Accuracies of machine learning algorithms using 22 training images.

Accuracy Table	Datasets		
ML Algorithm	**WD**	**RBAL**	**BD**
SVC	96%	96%	92%
LDA	71%	71%	63%
KNN	83%	71%	63%
DT	75%	79%	67%
LR	71%	79%	58%
AVERAGES	**79%**	**79%**	**69%**

Table 8. Average ML miss rates for datasets using 22 training images.

ML Algorithms Miss Rates for Datasets	WD	RBAL	BD
Average Black Miss Rate	17.50%	20.00%	28.75%
Average White Miss Rate	22.50%	21.67%	37.50%
The Difference of Average Miss Rates	−5.00%	1.67%	8.75%

The same conclusion can be made when considering the additional metrics. The introduction of other images has allowed the performances to improve, but comparing the versions, shows an existing difference based on racial weighting in the datasets. Table 9 provides the averages for the metrics, and complete findings and scoring metrics are in [28].

Table 9. Additional ML metrics precision, recall and F1.

ML Metrics Averages	WD	RBAL	BD
Precision	79%	79%	69%
Recall	100%	100%	100%
F1	88%	88%	81%

4.2. Racial Bias across DL Algorithms

We were careful to use the exact dataset representations and configurations for the experimentation for deep learning algorithms. Our initial testing followed the same approaches and methods, and we first applied the AlexNet deep learning algorithm using the eight training images datasets. Matplotlib provides an insight into the model performance for each epoch to visually represent the accuracy and loss of the model. For these tests, we have also utilized the included confusion matrix support for prediction and accuracy visualization. During our training, loss and accuracy metrics are gathered to generate a visual representation for each epoch, as shown in Figure 9.

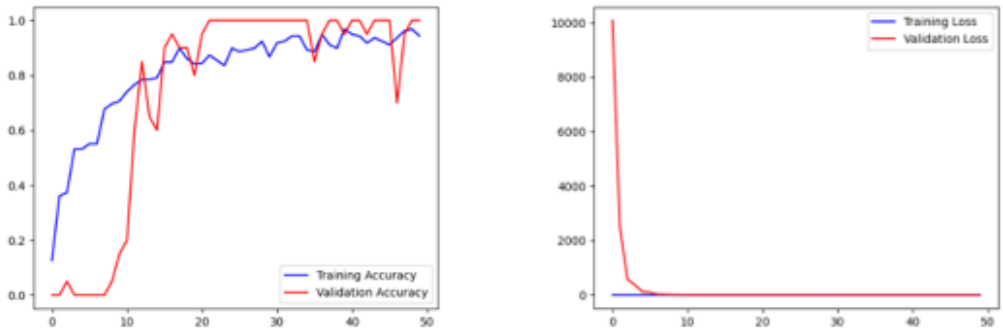

Figure 9. Validation and loss graph examples.

Ideally, the graphs should show decreasing loss and increasing accuracy for each cycle. These graphs are general reference material and have minimal differences if the algorithm performs as desired, and input is applied appropriately, but quickly indicates when a problem arises. The complete accuracy and loss graphs can be found in [28].

The confusion matrix is formatted along the y-axis to include the labels that are known to be true. The x-axis is formatted with tags that represent the prediction of the given test image based on the training. Ideally, the projections create a diagonal formation if all predictions are accurate. A deviation from this diagonal formation indicates a miss. Using our balanced training dataset with eight available images, we generate the results and immediately see that VGG16 only has one miss, and the others have two misses, as shown in Figure 10. The complete confusion matrices are in [28].

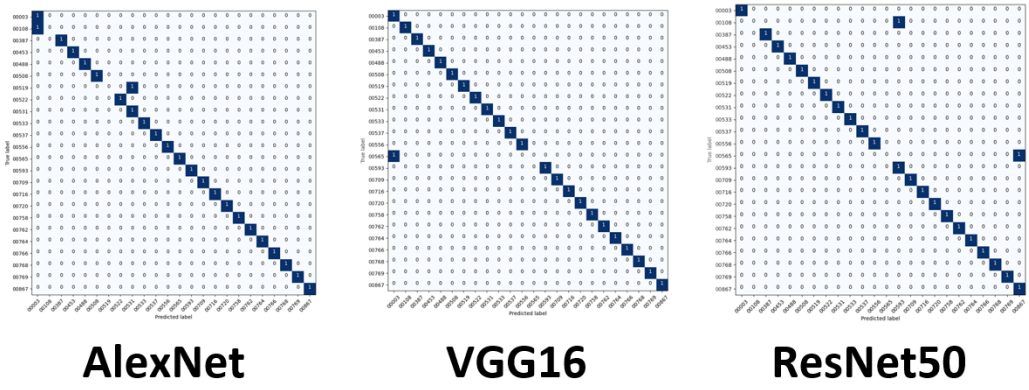

Figure 10. DL confusion matrices using balanced datasets and 8 training images.

The results are excellent if you recall the previous percentages recorded for traditional machine learning approaches and models. These balanced matrices serve as our baseline, and imbalance matrices produced similar results for each model. However, the most notable takeaway from these results is consistency. One model may outperform the other, but the models themselves are more consistent across the datasets, as shown in Figure 11. It appears these types of algorithms do not display a bias for a particular dataset.

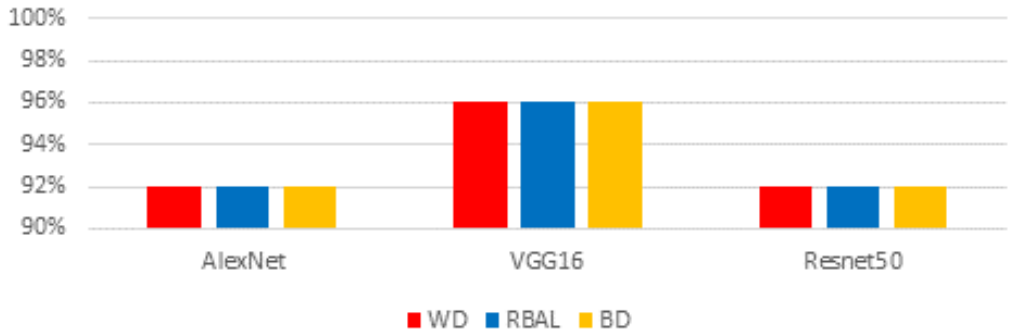

Figure 11. DL accuracies on all datasets using 8 training images.

These consistent results are also observable in the overall averages of the deep learning algorithms. VGG16 did outperform AlexNet and ResNet50 by 4%, but the general standards between the datasets are equal, as shown in Table 10.

Measuring the miss rates does show a sizeable difference statistically, but with such high accuracies, this difference can be misleading. Table 11 shows that it should be expected that missing more of the race is biased because there are more opportunities to miss that race.

Table 10. Accuracies of deep learning algorithms using 8 training images.

Accuracy Table	Datasets		
DL Algorithm	**WD**	**RBAL**	**BD**
AlexNet	92%	92%	92%
VGG16	96%	96%	96%
Resnet50	92%	92%	92%
AVERAGES	**93%**	**93%**	**93%**

Table 11. Average DL miss rates for datasets using 8 training images.

DL Algorithms Miss Rates for Datasets	WD	RBAL	BD
Average Black Miss Rate	8.33%	5.56%	4.17%
Average White Miss Rate	6.25%	8.33%	12.50%
The Difference of Average Miss Rates	2.08%	2.78%	8.33%

We continue our observations of the eight-training image experiment with the deep learning algorithms by measuring the additional metrics. Keras allows us to measure many metrics with many variations such as weighted, micro, and macro. To keep our comparisons in similar fashions, we will quote the averages using the micro scores because they are comparable to the machine learning results. Ref. [28] includes the complete additional scoring metrics. Although some metrics are better than the others, if we look at the averages in Table 12, the results are consistent, like the previously reported accuracies.

Table 12. Additional DL metrics precision, recall and F1 using 8 training images.

DL Metrics Averages	WD	RBAL	BD
Precision	97%	97%	97%
Recall	93%	93%	93%
F1	95%	95%	95%

As we continue the same methods, we followed our machine learning approaches, and we repeat the previous steps using 11 images for training. This round of training produced similar loss and accuracy graphs like the first experiment, and the complete selection of charts is in [28]. When investigating accuracy, we first compare the generated confusion matrices, in Figure 12, for our balanced datasets with 11 training images.

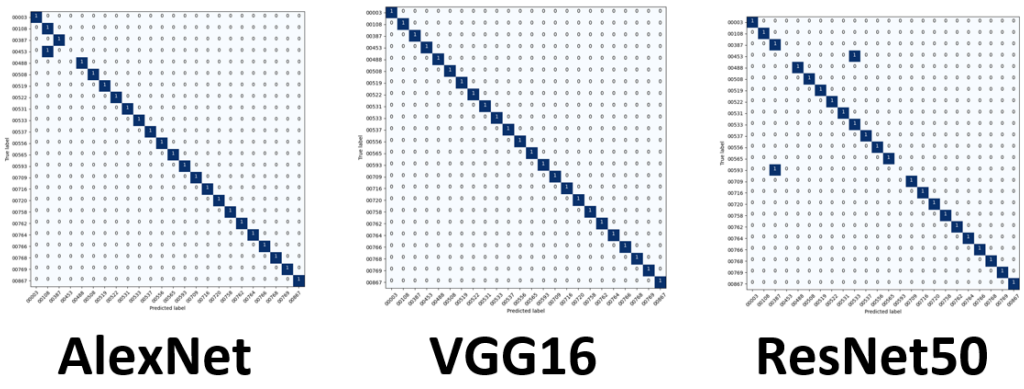

Figure 12. DL confusion matrices using balanced datasets and 11 training images.

The matrices represent our baseline results, and the other matrices are similar, which indicates minimal differences. The complete library of confusion matrices is in [28]. When

experimenting with the imbalanced datasets. We find again that the results are consistent, as shown in Figure 13.

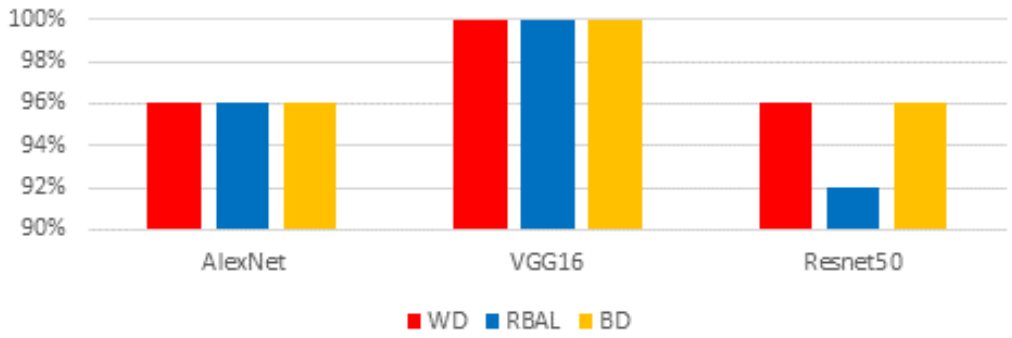

Figure 13. DL accuracies on all datasets using 11 training images.

VGG16 was the better performing model by successfully recognizing all faces. AlexNet and ResNet50 produced similar results, mostly only missing one look, but a key detail in ResNet50 results is the change from our balanced dataset. Deviating from our baseline approach yields better outcomes for both racial groups. Considering the averages of accuracies of all three algorithms in Table 13 supports the findings that a racially weighted dataset does not hinder performance.

Table 13. Accuracies of deep learning algorithms using 11 training images.

Accuracy Table	**Datasets**		
DL Algorithm	**WD**	**RBAL**	**BD**
AlexNet	96%	96%	96%
VGG16	100%	100%	100%
Resnet50	96%	92%	96%
AVERAGES	**97%**	**96%**	**97%**

Investigating the miss rates for these deep learning algorithms arrives at the same averages as our eight training image experiments, but differently. In Table 14, our higher Black dataset miss rate results from the percentage of missed Whites. The miss rate is relative to the fact that there are fewer Whites, giving the impression of a bad average. The averages are minimal when considering the accuracies at which the algorithms performed, and this small number of misses could fall within the margin of error.

Table 14. Average DL miss rates for datasets using 11 training images.

DL Algorithms Miss Rates for Datasets	**WD**	**RBAL**	**BD**
Average Black Miss Rate	4.17%	5.56%	0.00%
Average White Miss Rate	2.08%	2.78%	8.33%
The Difference of Average Miss Rates	2.08%	2.78%	8.33%

Additional metric findings indicate a similar result as the previous experiment. The complete metric analysis is in [28]. The results indicate consistency again. We provide the averages in Table 15. In this instance, we find that utilizing a dominant dataset performed slightly better than our baseline results.

Table 15. Additional DL metrics precision, recall and F1 using 11 training images.

DL Metrics Averages	WD	RBAL	BD
Precision	97%	99%	99%
Recall	97%	96%	97%
F1	97%	97%	98%

The final round of experimentation with our deep learning algorithms used a dataset that we created using 22 images in hopes of improving accuracy because our literature reviews indicated that these algorithms are capable of better performance. The accuracy and loss graphs in [28] were like all the others, but our confusion matrices slightly improved for this experiment. Our balanced or baseline test results were AlexNet and VGG16 performing flawlessly and ResNet50 missing only one face, as shown in Figure 14.

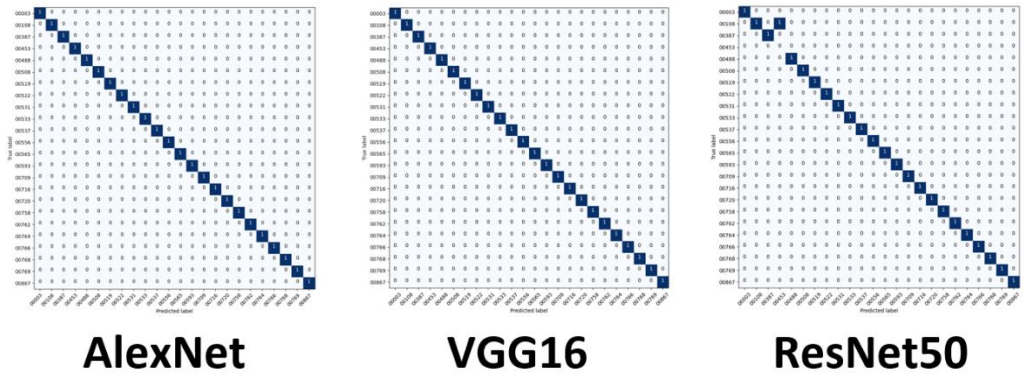

Figure 14. DL confusion matrices using balanced datasets and 22 training images.

The other matrices indicate similar performance with ResNet50 missing one subject when using the Black dominant (BD) dataset, but missing none when using the White prevailing dataset, as shown in Figure 15.

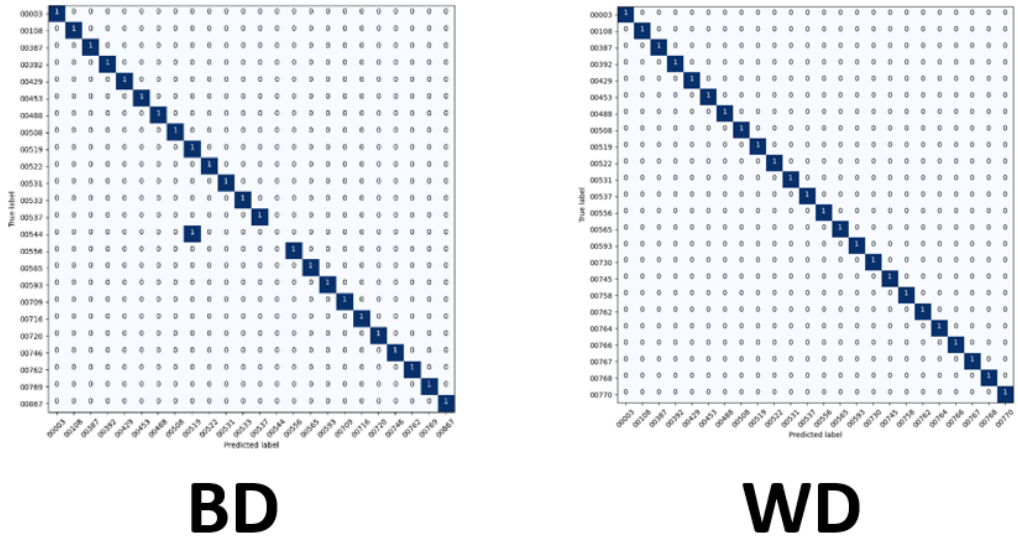

Figure 15. ResNet50 imbalanced confusion matrices using 22 training images.

Like the other comparisons, consistency is a critical contributing factor to these successful passes. Performing with such accuracy and missing only one test subject is near-perfect results. Figure 16 shows how AlexNet and VGG16 share similar consistency, and ResNet50 nearly matching them.

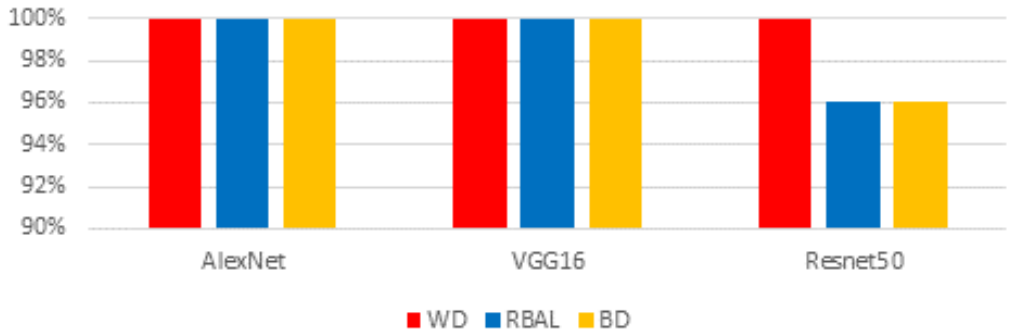

Figure 16. DL accuracies on all datasets using 22 training images.

The averages for these three deep learning algorithms demonstrate excellent results. As mentioned, AlexNet and VGG16 missed zero faces. ResNet50 missed one face using the balanced dataset and the Black dominant dataset, as shown in Table 16.

Table 16. Accuracies of deep learning algorithms using 22 training images.

Accuracy Table	Datasets		
DL Algorithm	**WD**	**RBAL**	**BD**
AlexNet	100%	100%	100%
VGG16	100%	100%	100%
Resnet50	100%	96%	96%
AVERAGES	**100%**	**99%**	**99%**

The miss rates indicate similar results with near-perfect execution. It is expected to see a total percentage. Again, the miss rates for this experiment result from missing a White test subject when using a Black Dominant (BD) dataset, which should be expected. Table 17 shows the minimal differences of the miss rate percentages when using the different datasets. There is a slight underperformance, but the results have improved by adding more images, and this could be considered within the margin of error.

Table 17. Average DL miss rates for datasets using 22 training images.

DL Algorithms Miss Rates for Datasets	WD	RBAL	BD
Average Black Miss Rate	0.00%	2.78%	0.00%
Average White Miss Rate	0.00%	0.00%	4.17%
The Difference of Average Miss Rates	0.00%	2.78%	4.17%

The additional precision, recall, and F1 scoring metrics corroborate these findings and again show a slight difference that could be considered within the margin of error. It is to be expected again that a total percentage would result in near-perfect execution, as shown in Table 18.

Table 18. Additional DL metrics precision, recall and F1 using 22 images.

DL Metrics Averages	WD	RBAL	BD
Precision	100%	100%	100%
Recall	100%	99%	99%
F1	100%	99%	99%

4.3. Racial Bias between ML and DL Algorithms

We have thoroughly documented our findings up to this point, and this section serves as a global perspective and completes analysis for machine learning and deep learning algorithm experimentation. Testing results using the traditional machine learning algorithms resulted in skewed results. These results ultimately displayed that not only does bias exist, but it can also alter results depending on the approach or algorithm. The bias that we observe from our testing indicates that the machine learning algorithms we tested with were biased towards White Subjects. In each instance of trying, the result was underperformances when using a Black dominant (BD) dataset and the opposite when using a White weighted dataset, regardless of algorithms used. We do find that when testing with our machine learning algorithms, Support Vector Classifier (SVC) was the best for mitigating this existing bias, and K-Nearest Neighbors (KNN) was the worst option for bias mitigation, as shown in Figure 17.

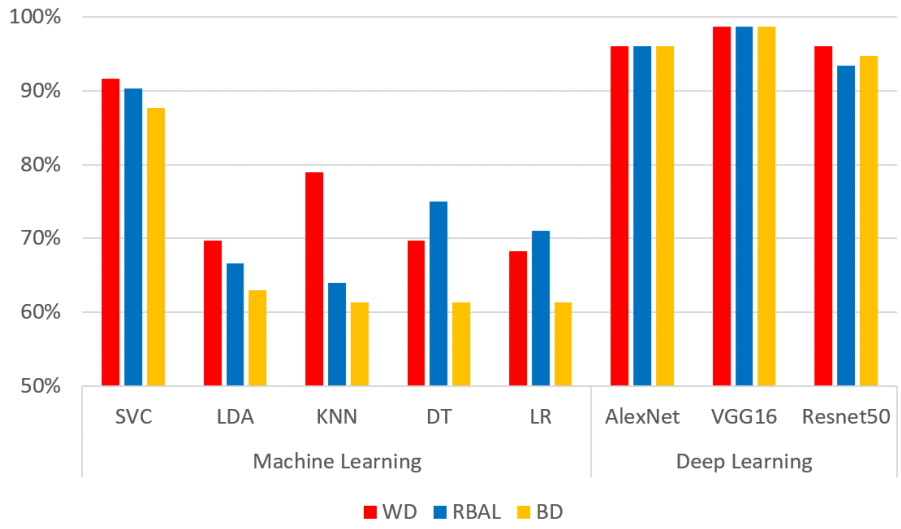

Figure 17. Complete algorithms summary with all datasets.

We also determine that when using deep learning algorithms, the level of complexity dramatically increases. However, better performance is a want, and higher accuracy is achieved when using these algorithms. We notice that we could not witness any bias using deep learning algorithms, as these algorithms seem to mitigate the previously detected bias. Accuracy, performance, and all other metrics indicate equality. We do find that VGG16 was the best at mitigating the bias and overall had quicker execution times. We also conclude that AlexNet and ResNet50 were essentially the same. Considering the complexity within the layering that ResNet required, we would give AlexNet a slight advantage. Using deep learning algorithms significantly increased performance while mitigating bias and producing consistent results, as demonstrated in the average totals in Table 19.

An additional consideration to verify the superiority of these deep learning algorithms can be seen by surveying the miss rates. The initial miss rates were significantly higher for the machine learning algorithms. The results were high in numbers, and the difference in the miss rates across racially weighted datasets was vast. The use of deep learning algorithms drastically reduced the miss rate, but it did so for both ethnicities, as shown in Figure 18.

We find that machine learning algorithms allow bias to exist, while deep learning algorithms do better at mitigating this undesirable consequence. These findings were observed by analyzing many metrics in an organized fashion to produce verifiable conclusory results. We also note the importance of the dataset and the consideration given when choosing an algorithm, model, approach, or technique.

Table 19. Complete accuracies for all algorithms.

Traditional Machine Learning Algorithms Global Totals				
Accuracy Table	**Datasets**			**AVERAGES**
ML Algorithm	**WD**	**RBAL**	**BD**	
SVC	92%	90%	88%	90%
LDA	70%	67%	63%	66%
KNN	79%	64%	61%	68%
DT	70%	75%	61%	69%
LR	68%	71%	61%	67%
AVERAGES	**76%**	**73%**	**67%**	72%
Deep Learning Algorithms Global Totals				
Accuracy Table	**Datasets**			**AVERAGES**
DL Algorithm	**WD**	**RBAL**	**BD**	
AlexNet	96%	96%	96%	96%
VGG16	99%	99%	99%	99%
Resnet50	96%	93%	95%	95%
AVERAGES	**97%**	**96%**	**96%**	96%

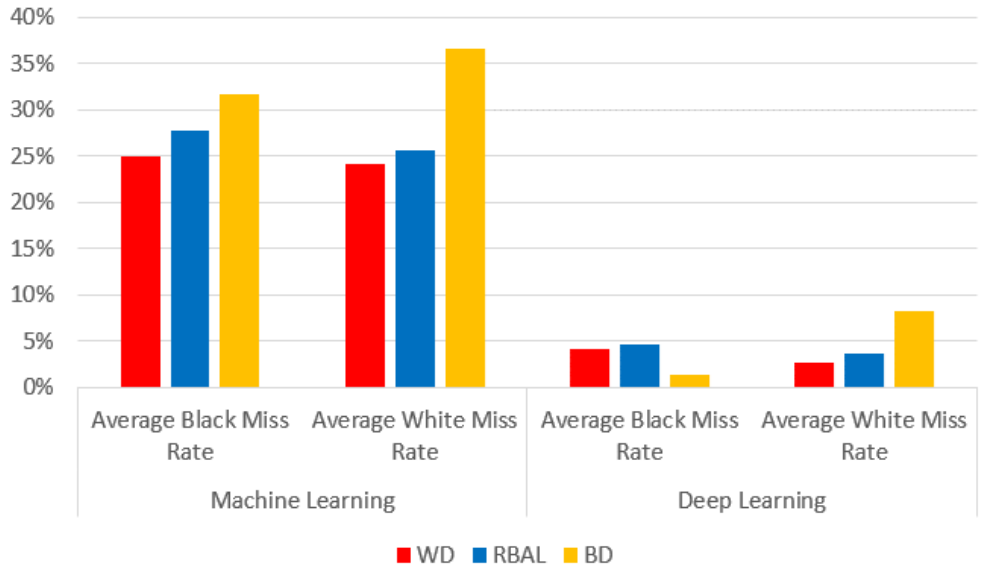

Figure 18. Miss rates for all datasets.

5. Conclusions and Future Work

We have conducted thorough coverage throughout our research covered in this document. We first gave a little insight into facial recognition with our introduction, preliminaries, and problem statement. We provided four problems related to facial recognition during our problem statement, and we laid out our contributions to the scientific community. Additional issues related to race, ethnicity, algorithms, and datasets followed to provide an encompassing generalized view of our research. Next, we reviewed our system design, development, and architectures. We gave an in-depth evaluation plan for each type of algorithm, dataset, and a look into the software and its architecture. Concluding those informative items, we then thoroughly explained the result and findings of our experimentation. We provided analysis for the machine learning algorithms and compared them across the differing dataset configurations. We then replicated those experiments using deep learning algorithms and explored their performance across the different dataset

types. Concluding the analysis, we compared the results of two kinds of algorithms and compared their accuracy, metrics, miss rates, and performances.

We evaluated racial bias across five machine learning algorithms using racially imbalanced and balanced datasets. The five machine learning algorithms explored were Support Vector Classifier (SVC), Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), Decision Trees (DT), and Logistic Regression (LR). We evaluated racial bias across three deep learning algorithms using racially imbalanced and balanced datasets. The three deep learning algorithms explored were AlexNet, VGG16, and ResNet50. We evaluated and compared the accuracy and miss rates between all tested algorithms and reported that SVC is the superior machine learning algorithm and VGG16 is the best deep learning algorithm based on our experimental study. Our findings conclude the algorithm that mitigates the bias the most is VGG16, and all our deep learning algorithms outperformed their machine learning counterparts.

The deep learning algorithms utilized for this study all have layering as a commonality. This layering is what assists these algorithms in performing significantly better than their machine learning counterparts. While comparing these models the significant difference between them are the numbers of layers, their connections, strides, pooling, and density. The number of layers for each algorithm yields the trainable parameters [29]. AlexNet uses 8 layers resulting in 60 million trainable parameters where VGG16 has 16 layers resulting in 138 million parameters [29]. ResNet50 uses 50 layers that yield 23 million trainable parameters but using significantly more layers requires additional time to train and geometric fallout likely led to our findings.

Our research has barely scratched the surface as it relates to facial recognition. Our future work can be expanded in many directions towards the approaches we used here. Our team is currently investigating similar research about the impacts of genders and ages across algorithms. There are plenty of other algorithms that can be explored, including variants of the ones we have used. VGG16 and ResNet50 have variants that could expand on our research. Different algorithms such as Xception and Inception also have variants that could develop this research even further. Expansion for this research does not end with algorithm approaches as other datasets may also be helpful. We considered experimenting with more datasets such as FairFace or DemogPairs as potential future work. These datasets are focused on mitigating bias and would have served as an excellent addition to this research. A final consideration for research expansion should include many racial representations, as this research focuses on two ethnicities.

Author Contributions: Conceptualization, J.C. and M.A.; Methodology, J.C. and M.A.; Software, J.C.; Validation, J.C.; Formal Analysis, J.C. and M.A.; Investigation, J.C.; Resources, J.C.; Data Curation, M.A.; Writing—Original Draft Preparation, J.C.; Writing—Review & Editing, J.C. and M.A.; Visualization, J.C.; Supervision, M.A.; Project Administration, M.A.; Funding Acquisition, M.A. All authors have read and agreed to the published version of the manuscript.

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

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Supporting experimental results available at https://9920d21e-b718-4d9f-91f5-be2bdcf3d554.filesusr.com/ugd/ee66e4_0a830271770e4872b81fc6abf6e83b89.pdf.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Khalil, A.; Ahmed, S.G.; Khattak, A.M.; Al-Qirim, N. Investigating Bias in Facial Analysis Systems: A Systematic Review. *IEEE Access* **2020**, *8*, 130751–130761. [CrossRef]
2. Klare, B.F.; Burge, M.J.; Klontz, J.C.; Bruegge, R.W.V.; Jain, A.K. Face Recognition Performance: Role of Demographic Information. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 1789–1801. [CrossRef]
3. Hupont, I.; Fernandez, C. DemogPairs: Quantifying the Impact of Demographic Imbalance in Deep Face Recognition. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019; pp. 1–7.
4. Alhindi, T.J.; Kalra, S.; Ng, K.H.; Afrin, A.; Tizhoosh, H.R. Comparing LBP, HOG and Deep Features for Classification of Histopathology Images. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–7.
5. Loo, E.K.; Lim, T.S.; Ong, L.Y.; Lim, C.H. The Influence of Ethnicity in Facial Gender Estimation. In Proceedings of the 2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA), Penang, Malaysia, 9–10 March 2018; pp. 187–192.
6. Masi, I.; Wu, Y.; Hassner, T.; Natarajan, P. Deep Face Recognition: A Survey. In Proceedings of the 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Parana, Brazil, 29 October–1 November 2018; pp. 471–478.
7. Brownlee, J. A Gentle Introduction to Deep Learning for Face Recognition. 30 May 2019. Available online: https://machinelearningmastery.com/introduction-to-deep-learning-for-face-recognition/ (accessed on 4 January 2021).
8. Das, A.; Dantcheva, A.; Bremond, F. Mitigating Bias in Gender, Age and Ethnicity Classification: A Multi-Task Convolution Neural Network Approach. In *Lecture Notes in Computer Science*; Springer International Publishing: Cham, Switzerland, 2019; pp. 573–585.
9. Faudzi, S.A.A.M.; Yahya, N. Evaluation of LBP-Based Face Recognition Techniques. In Proceedings of the 2014 5th International Conference on Intelligent and Advanced Systems (ICIAS), Kuala Lumpur, Malaysia, 3–5 June 2014; pp. 1–6.
10. Karkkainen, K.; Joo, J. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv* **2019**, arXiv:1908.04913. Available online: http://arxiv.org/abs/1908.04913v1 (accessed on 22 January 2021).
11. Yashoda, M.; Sharma, R.S. Face Recognition: Novel Comparison of Various Feature Extraction Techniques. In *Harmony Search and Nature Inspired Optimization Algorithms*; Springer: Singapore, 2019; pp. 1189–1198.
12. Pali, V.; Goswami, S.; Bhaiya, L.P. An Extensive Survey on Feature Extraction Techniques for Facial Image Processing. In Proceedings of the 2014 International Conference on Computational Intelligence and Communication Networks, Bhopal, India, 14–16 November 2014; pp. 142–148.
13. Ryu, H.J.; Adam, H.; Mitchell, M. Inclusive Face Net: Improving Face Attribute Detection with Race and Gender Diversity. *arXiv* **2018**, arXiv:1712.001939. Available online: http://arxiv.org/abs/1712.00193v3 (accessed on 29 January 2021).
14. Li, L.; Fieguth, P.; Guo, Y.; Wang, X.; Pietikäinen, M. Local Binary Features for Texture Classification: Taxonomy and Experimental Study. *Pattern Recognit.* **2017**, *62*, 135–160. [CrossRef]
15. Dass, R.K.; Petersen, N.; Visser, U.; Omori, M. It's not Just Black and White: Classifying Defendant Mugshots Based on the Multidimensionality of Race and Ethnicity. In Proceedings of the 2020 17th Conference on Computer and Robot Vision (CRV), Ottawa, ON, Canada, 13–15 May 2020; 2020; pp. 238–245.
16. Patil, S.A.; Deore, P.J. Face Recognition: A Survey. *Inform. Eng. Int. J.* **2013**, *1*, 31–41.
17. Pietikãinen, M. Local Binary Patterns. *Sch. J.* **2010**, *5*, 9775.
18. do Prado, K.S. Face Recognition: Understanding LBPH Algorithm—Towards Data Science. Towards Data Science. 10 November 2017. Available online: https://towardsdatascience.com/face-recognition-how-lbph-works-90ec258c3d6b (accessed on 9 December 2020).
19. Rosebrock, A. Local Binary Patterns with Python & OpenCV. 7 December 2015. Available online: https://www.pyimagesearch.com/2015/12/07/local-binary-patterns-with-python-opencv/ (accessed on 4 January 2021).
20. Siju, V. A Survey on Machine Learning Algorithms for Face Recognition. *Int. Res. J. Eng. Technol.* **2008**, *7*, 1072–1075.
21. Wang, M.; Deng, W. Deep Face Recognition: A Survey. *Neurocomputing* **2021**, *429*, 215–244. [CrossRef]
22. Yucer, S.; Samet, A.; Noura, A.-M.; Toby, P.B. Exploring Racial Bias within Face Recognition via Per-Subject Adversarially-Enabled Data Augmentation. *arXiv* **2020**, arXiv:2004.08945. Available online: http://arxiv.org/abs/2004.08945 (accessed on 3 February 2021).
23. Gwyn, T.; Atay, M.; Kaushik, R.; Esterline, A. Evaluation of Local Binary Pattern Algorithm for User Authentication with Face Biometric. In Proceedings of the 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 14–17 December 2020; pp. 1051–1058.
24. Ahonen, T.; Hadid, A.; Pietikäinen, M. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 2037–2041. [CrossRef] [PubMed]
25. Jonathon, P.P.; Wechsler, H.; Huang, J.; Rauss, P.J. The FERET Database and Evaluation Procedure for Face-Recognition Algorithms. *Image Vis. Comput.* **1998**, *16*, 295–306.
26. Phillips, P.J.; Moon, H.; Rizvi, S.A.; Rauss, P.J. The FERET Evaluation Methodology for Face-Recognition Algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1090–1104. [CrossRef]
27. Artificial Intelligence vs Machine Learning vs Deep Learning—GeeksforGeeks. 2019. Available online: https://www.geeksforgeeks.org/artificial-intelligence-vs-machine-learning-vs-deep-learning/ (accessed on 23 January 2019).

28. Coe, J.; Gipson, H. Appendix A: MACHINE LEARNING ADDITIONAL METRICS. Available online: https://9920d21e-b718-4d9f-91f5-be2bdcf3d554.filesusr.com/ugd/ee66e4_0a830271770e4872b81fc6abf6e83b89.pdf (accessed on 5 August 2021).
29. Özgenel, Ç.F.; Sorguç, A.G. Performance Comparison of Pretrained Convolutional Neural Networks on Crack Detection in Buildings. In Proceedings of the 35th International Symposium on Automation and Robotics in Construction (ISARC), International Association for Automation and Robotics in Construction (IAARC), Berlin, Germany, 20–25 July 2018.