Machine Learning Methods to Improve Fairness and Prediction Accuracy on Large Socially Relevant Datasets.

by

BHANU JAIN

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

August 2021

To my daughters, with whom I learnt what it takes to succeed

and

to my grandparents, parents, sister, nephews, niece, uncle, aunts, friend

and

to all who make our world a better place with their

hard work, selfless service, intellect, honesty, ethics, compassion, presence, and love.

ACKNOWLEDGEMENTS

I would like to thank my supervising professors Dr. Ramez Elmasri and co-advisor Dr. Huber for guiding me, encouraging me, motivating me, for sharing their knowledge and wisdom with me, and for the generous gift of their time throughout my pursuit of PhD. I am thankful to Dr. Leonidas Fegaras and Dr. Vassilis Athitsos for serving in my committee, for guiding me, and for sharing their knowledge, time, and valuable teaching resources with me.

I would like to thank my grandparents, parents, sister, uncle and aunts for conveying through their life and actions the importance of discipline, hard work, education, service, ethics, compassion, and love. I want to thank my sister, daughters, nephews, niece, and dear friends for encouraging me, motivating me, loving me, and for being great role models.

I am thankful to my friends Chitra, Fatma, Israel, Madhu, Rodrigo, Samiksha, Chris, Mary, Theodora, M. Hani, Shalini, Daniel, Elizabeth, Tish, Sally, Satish ji, Aryaman, and Sanjana for their friendship, support and encouragement. I am thankful to M. Rezaei, M. Shaito, Aishwarya, Bhanu, Hasitha, Rachana, Bader, Tariq, Mousa, Remesh, Fariba, Sanika, Jasmine, Maria, Akilesh, Jubin, Fadiah, Marnim, Aashara, Zaman, Yash, Spardha, Donna, Ariella, Giun, Ashish, and Huadi for making it fun to work at UTA.

I am thankful to Dr. Jiang, Dr. Li, Dr. Khalili, Dr. Peterson, Dr. Ghidini, Dr. Gonzalez, Mr. Levine, Dr. Csallner, Dr. Conly, Dr. Stefan, Dr. Tiernan, Dr. Makedon, Dr. Kamangar, Dr. Eary, Dr. Guizani, Ms. Gotcher, Ms. Dickens, Ms.

McBride, Mr. Coates, Mr. Irie, Mr. Harris, Mr. Orcutt, and my students for making life at UTA a rich experience.

I am grateful to all who gave me an opportunity to contribute to their lives and to walk a part of their journey with them just as I am grateful to all who have contributed towards my life and walked my journey with me.

July 30, 2021

ABSTRACT

Machine Learning Methods to Improve Fairness and Prediction Accuracy on Large Socially Relevant Datasets.

Bhanu Jain, Ph.D.

The University of Texas at Arlington, 2021

Supervising Professor: Ramez A. Elmasri, Co-supervisor: Dr. Manfred Huber

Machine learning-based decision support systems bring relief to the decision-makers in many domains such as loan application acceptance, dating, hiring, granting parole, insurance coverage, and medical diagnoses. These support systems facilitate processing tremendous amounts of data to decipher the embedded patterns. However, these decisions can also absorb and amplify bias embedded in the data.

An increasing number of applications of machine learning-based decision support systems in a growing number of domains has directed the attention of stakeholders to the accuracy, transparency, interpretability, cost effectiveness, and fairness encompassed in the ensuing decisions. In this dissertation, we have focused on fairness and accuracy embodied in such predictions. When making machine learning based forecasts, there are a series of sub-problems within the overarching problem of addressing bias and accuracy in decisions that we address in this work: 1) detecting bias in the predictions, 2) increasing accuracy in predictions, 3) increasing prediction accuracy without tampering with the class labels and while excluding sensitive at-

tributes that trigger bias, 4) quantifying bias in a model, and finally 5) reducing a model's bias during the training phase.

In this dissertation we develop machine learning methods to address the aforementioned problems to improve fairness and prediction accuracy while using three large socially relevant datasets in two different domains. One of the two Department of Justice recidivism datasets as well as the Census-based adult income-based datasets hold significant demographic information. The second recidivism dataset is more feature rich and holds information pertaining to criminal history, substance-abuse, and treatments taken during incarceration and thus provides a rich contrast to the largely demographic datasets when comparing fairness in predicted results.

Our approach is focused on data preparation, feature enrichment in activity and personal history-based datasets, model design, and inclusion of loss function regularization alongside the traditional binary cross entropy loss to increase both fairness and accuracy. We achieve this without tampering with the class labels and without balancing the datasets. To stay squarely focused on fairness, we do not include the sensitive attributes in our input features while training the models.

In the experiments we show that we can increase accuracy and fairness in the predictions based on the three dataset beyond what has been achieved in the published literature. The results demonstrate that our fairness improvement approach via loss functions is applicable in different domains with different sensitive attributes and can be applied without manipulating class labels or balancing skewed datasets.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

In this chapter, we first introduce the area of focus of this dissertation, which is to develop machine learning methods to improve fairness and prediction accuracy on large socially relevant datasets. In Section 1.1 we focus on defining the problem that we wish to tackle in this work. Then, in Section 1.2, we state the motivation for our work. This is followed by an outline of our research contributions in Section 1.3 and an outline of the dissertation in Section 1.4.

1.1 Problem Statement

Machine learning-based decision support systems bring relief and support to the decision-maker in many domains such as loan application acceptance, dating, hiring, granting parole, insurance coverage, and medical diagnoses. These support systems facilitate processing tremendous amounts of data to decipher the patterns embedded in them. However, these decisions can also absorb and amplify bias embedded in the data and render these predictions inequitable and prejudiced for sub populations. Therefore, in this dissertation, we work to develop machine learning techniques to improve both fairness and accuracy of the predictions. The input data is often unbalanced and this mandates that newly developed techniques can handle unbalanced data. Finally, it is important to quantifiably measure bias in results.

1.2 Motivation to Improve Fairness and Accuracy in Machine Learning Based Predictions

Machine learning-based decision support systems developed and deployed in numerous domains directly impact many aspects of human lives. These systems help in bringing support to the decision-maker in domains such as medical diagnosis [1], loan application acceptance [2], dating [3], hiring [4], granting parole [5], and predicting insurance reserve [6], to name just a few. These support systems help process incredible amounts of data but can also heighten the bias embedded in these datasets. For example, biased results can be observed in the risk-assessment software used in criminal justice [6], or in travel, where fare aggregators can direct Mac users to more extravagant hotels [7], or in the hiring domain where females may see fewer high paying job ads [8]. In healthcare, there is evidence that healthcare professionals exhibit bias based on race, gender, wealth, weight, etc. [9]. This bias permeates into the data and is reflected in the results of diagnostics, prediction, and treatment [10]. Due to the widespread impact of AI-based decisions, both the accuracy and fairness of these recommendations has become a popular and pertinent topic of research [11]. Hence we focus on both accuracy and bias in this dissertation.

Special interest groups negatively affected by decision support systems can often be categorized by sensitive attributes such as race, gender, affluence level, weight, and age, to name a few. While machine learning-based decision support systems often do not consider these attributes explicitly, biases in the data sets, coupled with the used performance measures can nevertheless lead to significant discrepancies in the system's decisions. For example, many minorities have traditionally not participated in many domains such as loans, education, employment in high paying jobs, receipt of health care services, etc. This can lead to unbalanced datasets as the minority-based data may be combined with the majority sensitive attribute-based data. Similarly,

some domains like homeland security, refugee status determination, incarceration, parole, loan repayment, etc., may be already riddled with bias against certain subpopulations, even in the absence of AI based decision support system. Such human bias seeps into the datasets used for AI based prediction systems which, in turn, amplify it further.

Thus, as we begin to use Artificial Intelligence (AI) based decision support system, it becomes important to ensure fairness for all who are affected by these decisions. Therefore, our approach focuses on increasing accuracy and fairness on any sensitive attributes. We use three datasets from two different domains and two different sensitive attributes to illustrate the efficacy and adaptability of our work to diverse domains. To demonstrate a quantifiable decrease of bias in the models via our approach, we introduce a measure called Bias Parity Score (BPS) to represent bias in a single measure. Furthermore, our BPS-inspired loss functions are capable of mitigating bias in unbalanced dataset and are thereby capable of addressing the traditional lack of participation of minorities in pertinent datasets.

1.3 Dissertation Contributions

In this dissertation, we introduce efficient and adaptable methods to spot, quantifiably measure, and reduce bias via Bias Parity Score BPS and BPS inspired loss functions while also increasing accuracy in predictions via feature enrichment. Our contributions illustrated through a sequence of five sets of experiments can be divided into three parts:

- In the first part described in Chapters 4, 5, and 6, we work with two socially relevant datasets in the recidivism domain to spot bias in predictions and analyze the effects of types of datasets on accuracy and fairness in predictions. We utilize very similar neural network architectures to learn recidivism predic-

tions for the two datasets and demonstrate that demographic datasets produce less accurate and fair predictions as compared to those produced with a feature rich personal activity and history-based dataset. We develop a framework in Chapter 6 to further enrich a dataset to ameliorate the prediction accuracy.

- In the second part covered in experiments described in Chapter 7, we propose a new fairness measure called Bias Parity Score (BPS) to measure bias quantifiably in the prediction models. The BPS score leverages an existing intuition of bias awareness and summarizes it in a single measure. We demonstrate how BPS score can be used to quantify bias for a variety of statistical quantities and how to associate disparate impact with this measure. We demonstrate how to use BPS to select the least biased model without sacrificing accuracy.
- In the third part covered in experiments described in Chapter 8, we formulate and introduce loss functions that are inspired by BPS. We apply the loss functions in the context of various fairness measures and study their performance, potential drawbacks, and deployment considerations in achieving fairness in terms of accuracy, False Positive Rate, False Negative Rate, True Positive Rate, True Negative Rate, or a combination of the statistical measures. By using these loss functions, we can measurably reduce bias in the results for the two race-based cohorts. Additionally, we investigate potential divergence and stability issues that can arise when using these fairness loss functions, when shifting significant weight from accuracy to fairness. Furthermore, to evaluate the applicability of our approach of using loss functions and different regularization weights in conjunction with the traditional binary cross entropy to increase fairness across sensitive attributes-based groups, we carry out the same experiments with a census-based income dataset from a different socially relevant domain and utilizing a different sensitive attribute, namely gender.

We show that we can improve fairness in the three socially relevant dataset results in the two distinct domains beyond what has been achieved in the published literature. The results demonstrate that with feature enhancement and a good choice of fairness loss function, we can reduce the trained model's bias without deteriorating accuracy even in unbalanced datasets.

1.4 Dissertation Organization

In Chapter 1 we introduce the topic of our research. In Chapter 2, in addition to the related work we list and define the bias metrics that we have used throughout the dissertation. In Chapter 3, we describe three datasets used in this dissertation. In Chapter 4, we describe our first set of experiments with a demographical dataset to make predictions and analyze the results. This is followed by Chapter 5, that describes our second set of experiments with a feature rich dataset. Here we compare the results with those of a demographic information based dataset using a similar neural network architecture.

In Chapter 6, we lay out an approach to reduce sensitive atribute-based bias and increase prediction accuracy by enriching a dataset followed by selecting a model based on False Positive Rate Parity. This enables us to choose a prediction model with least bias and high accuracy.

In Chapter 7, we lay out techniques to derive a feature-rich representation of a dataset from the temporal information in the data, which improves fairness and accuracy of predictions. We introduce a new fairness measure called Bias Parity Score (BPS) that summarizes bias in a single measure.

In Chapter 8, we introduce a family of BPS inspired loss functions and use them as a regularization component during the model training process in neural networks.

We demonstrate that with a good choice of fairness loss function we can reduce any specific kind of bias in a model even in unbalanced datasets.

CHAPTER 2

Bias Metrics and Literature Review

In this chapter we introduce the Bias Metrics used in our research and in previous work and discuss the relevant Related Work in the area.

2.1 Bias Metrics

For various experiments covered in Chapters 4, 5, and 6, we computed several metrics for our experiments: Accuracy, FPR (False Positive Rate), FNR (False Negative Rate), TPR (True Positive Rate), and TNR(True Negative Rate). We evaluated bias by comparing FPR and FNR for the two major subpopulations. To compute False Positive Rate Parity, we computed the ratio of False Positive Rates (FPR2/FPR1) for the two subpopulations and marked it for various models as shown in the second graphs in Figure 6.3 and Figure 6.4. The following definitions explain the metrics for our experiments in recidivism prediction.

True Positive(TP): future recidivist predicted to recidivate.

True Negative(TN): future non-recidivist forecasted to not recidivate.

False Positives(FP): future non-recidivist incorrectly forecasted to recidivate.

False Negatives(FN): future recidivist incorrectly predicted to not recidivate.

Accuracy: truthfulness of the predicted label - recidivist or non-recidivist.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

False Positive Rate(FPR): total number of false positive predictions (FP) divided by the total number of all the non-recidivists (Negatives).

$$FPR = \frac{FP}{FP + TN} \tag{2.2}$$

False Negative Rate(FNR): total number of incorrect negative predictions divided by the total number of all the recidivists (Positives).

$$FNR = \frac{FN}{FN + TP} \tag{2.3}$$

True Positive Rate or Sensitivity(TPR): total number of true positive (TP) predictions divided by the total number of all positives (recidivists).

$$TPR = \frac{TP}{FN + TP} \tag{2.4}$$

True Negative Rate or Specificity(TNR): total number of true negative (TN) predictions divided by the total number of all negatives (non-recidivists).

$$TNR = \frac{TN}{FP + TN} \tag{2.5}$$

A model becomes accurate as both FPR and FNR approach a value of zero. A high FPR signals that many non-recidivists are being predicted to be recidivists while a high FNR indicates that many recidivists are being falsely predicted to be non-recidivists. A high FPR and high FNR both individually or simultaneously lead to lower accuracy. A high TPR, (also called Sensitivity or Recall)is desirable as it indicates a higher proportion of actual recidivists that are correctly identified. Similarly, a high TNR (also called Specificity or Selectivity) is desirable as it indicates the proportion of non-recidivists that are correctly identified

Even though high accuracy, high TPR, and high TNR are always coveted in machine learning models for recidivism predictions, race-based bias is here kept under

control by having similar FPR and FNR for different subpopulations. Similar FPR and FNR for different subpopulations is often achieved at the cost of accuracy.

2.2 Related Work

Several studies in recidivism prediction such as those by [12], [13], [14], and [15], have used datasets described in Sections 3.1 and 3.2 that we have also used in the current work. In the following we discuss the most relevant related work both in recidivism prediction as well as in the broader area of fairness and bias in prediction systems.

2.2.1 Recidivism Prediction

Impact of incarceration: One work by Bhati [12] used this data to consider criminal history of offenders to delve into a theoretical approach to explore crimes averted by offenders' incarceration. Yet another study [13] used the same dataset to study the degree of impact that incarceration had on an offender's criminal trajectory and found that while 4% of offenders displayed a criminogenic effect, 40% of the offenders were released to a lower criminogenic trajectory than without the incarceration.

Effect of sentencing models work: The work in [14] used one of the same datasets as the current work to study the effect of different sentencing models on offender recidivism. The data was used to compare release programs in six states and the study found that the results varied by state, where mandatory parole release had an impact in some, parole board discretionary release was more successful in other states, while there was no noteworthy impact on recidivism in some states.

Bureau of Justice Statistics Report (3 year follow-up): A report by the Bureau of Justice Statistics on the second dataset is provided in [16]. The dataset

includes four ways to measure recidivism: re-arrest, reconviction, resentence to prison, and return to prison (with or without a new sentence). The report uses a three-year follow-up period after offender's release to study recidivism. In our experiments, we used reconviction as a measure of recidivism. This work included records of prisoners who were alive for the 3 year follow-up period. Since we used all records, and not merely the ones within the three years follow-up period since their 1994 release, we used a feature vector in our input variables that specifies whether a prisoner was dead or incarcerated for life.

Bureau of Justice Statistics Report (9 year follow-up): This is a follow-up study [17] on prisoners released in 2005 across 30 states that found 83% of state prisoners were re-arrested within 9 years of release. As per this report, 44% of prisoners followed were re-arrested in the first year, 34% during their third year, and about 24% during the ninth year. This report clearly indicated that prisoners continue to recidivate long after the three-year follow-up period used in other works such as those of Zeng et al. [15] and Ozkan [18]. It set us in the direction of splitting all arrest records in our dataset to include all arrest cycles that were available.

Re-arrest and Adjudication as Measures of Recidivism Work: As indicated above, the report by the Bureau of Justice Statistics on the dataset in [16] used 4 different variables to measure recidivism. Work by Zeng et at. [15] and Ozkan [18] on this data used rearrest and reconviction respectively as a measure of recisivism. In contrast, we have chosen to only use reconviction as a measure of recidivism in all of our experiments. We believe that arrest without a conviction is not a sufficient premise for prediction. Hence, our outcome variables laterCNV, laterDRUG etc record binary values of whether an offender was convicted for any, or for a specific crime any time after the release.

Input Variables and Past Arrest Cycles Related Work: Zeng et at. [15] did not use some of the personal activity-based features such as vocational or educational courses attended, and substance abuse treatments during the 1994 arrest cycle as very few prisoner records from the 1994 release had valid entries for these. We did use these variables and circumvented the "too few records" problem by considering each arrest cycle as a standalone record when the parole decision needs to be made. Hence, these variables could be included in the records from 1994 and later arrest cycles. We used -1 for missing values in pre-1994 cycles and added a binary feature vector to indicate that the arrest cycle record was from 1994 or afterwards to help neural networks decipher the temporal pertinence of these features in related experiments (presented in Chapter 6).

The work in [15] used only records from 1994 for prediction purposes. In contrast, we used all the arrest records embedded with the 1994 release data for accruing previous criminal activity information and the 1994 and post 1994 records for prediction purposes to use substance-abuse related variables efficiently. We used all the arrest records from 1994 onwards for prediction purposes. All records - before, during and after the 1994 release were used to calculate crimes committed in each of the broad crime categories. Like [15], we used drug, general violence, sexual violence, and fatal violence crime categories to align our experiments (presented in Chapter 5). However, our crime categories were based on adjudication rather than on arrest for these crimes. Furthermore, we categorized for property, public and other crimes as well, since these were required to cover all types of crimes included in the adjudication variables.

Sentence Length Correlation with Lower Recidivism work: As per the work by Tiedt et al. [19], longer prison sentences correlate with lower recidivism. Instead of limiting themselves to a single state, this study used criminal history data

from 30 state repositories. They found that it is difficult to extricate effects related to individual factors like age and previous criminal history from that of state practices and policies on recidivism. They found a greater disparity in sentence length amongst 45 years old or older prisoners and less so in those released before they turned 45.

Conviction as a Measure of Recidivism Work: Another work by Ozkan [18] used reconviction within three years after release as the outcome variable. Even though, the work did consider 15 cycles of prior arrests in the feature selection process for the 1994 arrest cycle, the study took binary variables for 15 prior arrests, while we took a rolling sum of priors for each of the 26 individual crime categories before we normalized the input vectors for our experiments (described in Chapter 6).

Interpreting Black Box Models Work: As we use artificial neural networks to increase prediction accuracy for recidivism, we are aware of the frequent criticism of artificial neural networks being an opaque algorithm [20]. To investigate this, the work in [21] presents an overview of literature covering the methods to explain systems based on opaque and obscure machine-learning models. They propose classification of methods to account for the specific explanation problem, the type of explanation adopted, the black box model opened, and the type of data used in the black box models.

Recidivism Prediction Based on Asymmetric Costs Work: Even though many works of research have focused on the accuracy of recidivism prediction, the work by Won et al. [22] brings to attention the disparate cost of errors in misrecognizing those who do not cause recidivism versus those who do cause recidivism. They weighed in that False Positive Errors cause additional monitoring while False Negative Errors raise social and economic costs. This study found XGBoost based models to have higher predictive accuracy and lower misclassification cost than those powered by logistic regression, decision trees, artificial neural networks, and support vector

machines. Hence, they aligned XGBoost based recidivism prediction with asymmetric error cost. This work stirred us to include XGBoost to conduct experiments with our dataset. However, as per our consideration, a rising ~~F~~false positive rate for one race is a cause of concern due to its implications in race-based bias. Therefore, we have chosen to monitor both false positive and false negative rate.

Race Related Works: Several works [23] [24] [16] have shown different rates of recidivism amongst Caucasians and African Americans. The Bureau of Justice Statistics report [23] specified that African Americans recidivate at a higher rate than Caucasians. Furthermore a report by Durose et al. [23] stated that different types of crimes have different recidivism rates with the highest rates in public order crimes and relatively lower for drugs, property, and least for violent crimes. Furthermore, we believe that people who commit similar types of crimes, share other similarities - such as similarity of prior criminal and substance abuse-based activities. Therefore, we used several activity and substance abuse based input variables listed and explained in Chapters 5 and 6, and aligned our experiments with crime categories.

As we pursue the goal of reducing race-based bias in recidivism predictions, we are cognizant of the ethical concerns around the issue. Some studies left the sensitive attribute of race [25] [10] out of the set of input variables while others [26] indicate that removing contentious attributes like gender, ethnicity, race etc. means excluding critical information. The work in [27] emphasized that eliminating sensitive attributes such as race and gender may help some and harm others.

A work by Calders et al. [28] proposed training a model individually for every value of the sensitive attribute of gender. Additionally, the work in [29] used the approach of exposing the classifier to data related to one race at a time without ever using the race feature vector explicitly. This allows using the race related information without pitting one race against another. In our current work we have used different

approaches, including this technique to study their impact on increased prediction accuracy while monitoring the particular choice's effect on race-based bias in the decisions.

Another study [30] sought to compare different fairness-enhancing classifiers with one another under a variety of fairness measures using different datasets, and to figure out the basis of the differences. They found that many of these measures were strongly correlated with each another. Furthermore, they found that the fairness-preserving algorithms they worked with, were sensitive to variations in the composition of the dataset, thus suggesting that fairness interventions were more "brittle" than believed earlier.

Disparity and Bias Related Works: Many anti-discrimination laws have come into existence to prevent unfairness meted out to individuals based on sensitive attributes such as gender, race, age, etc. Many studies [31, 11, 32] have shed light on this unfairness in machine learning-based prediction results. The presence of bias in prediction results could steer a practitioner away from machine learning-based support systems, particularly given that Dressel et al. [10] found that laypeople were as accurate as algorithms in predicting recidivism. However, other work by Jung et al. [33] in the same domain found that algorithms performed better than humans in predicting recidivism using three datasets. Jung et. al found this performance gap to be even more prominent when humans received no immediate feedback on the accuracy of their responses and in the presence of higher numbers of input features. Thus, given the need for using machine learning-based decision support systems, it is essential to find ways to increase accuracy and decrease bias in predictions.

Miron et al. [32] studied the causes of disparity between cohorts and found that static demographic features have a higher correlation with the protected features than the dynamic features such as substance abuse, peer rejection, and hostile

behavior. They found that static features cause disparity between groups ~~on~~in terms of group fairness metrics. In the current work, we also found that using several features like substance abuse variables, treatments taken and courses attended increased the prediction accuracy and decreased bias in the prediction.

2.2.2 Fairness and Bias in Prediciton

Bias and Fairness: Bias has been classified into three categories by Zafar et al. [34]: *disparate treatment, disparate mistreatment, and disparate impact. Disparate treatment* represents different outputs for different subgroups with the same (or similar) values of non-sensitive features. *Disparate mistreatment* indicates different misclassification rates for different subgroups; For example, when different subgroups have different False Negative rates (FNR) and False Positive rates (FPR). *Disparate impact* suggests decisions that benefit or hurt a subgroup more often than other groups.

Various studies such as [10, 25, 11], including our own work [29, 35] have demonstrated this unfairness in the results of machine learning-based predictions in many applications such as diagnosing diseases, predicting recidivism, making hiring decisions, etc, to name just a few. The presence of bias in machine learning-based prediction results leads to disparities in impact and treatment of some cohorts. This disparity has steered many researchers to conduct several studies such as [36, 37, 38, 34], to decrease unfairness in the results.

Biswas et al. [39] trained two statistical models: one on balanced data and the other on unbalanced data. They found that balanced data led to fairer prediction models than the one made with unbalanced data. In the current work, we worked with both balanced and unbalanced data sets. In addition, some of the feature enhancement work presented in this dissertation showed its benefits by changing an

unbalanced recidivism data set into an enriched dataset that is a balanced dataset for Caucasians and African Americans. In another work, Chouldechova et al. [40] identify three distortions that trigger unfair predictions by machine learning-based systems: human bias embedded in the data, the phenomenon that reducing average error fits towards majority populations, and the need to explore. Human bias embedded in the data is a major contributor to bias resulting from the dataset itself. For example, recidivism data typically records rearrests but needs to predict reoffense. Therefore, in our work, we captured adjudication in each arrest cycle and used that in the history considered and not prior arrests, assuming that re-arrests could have human bias embedded but adjudication/reconviction may have relatively less human bias. In contrast, a study by Zeng et al. [15] with whom we compare our results, uses arrests for predicting recidivism for the same dataset. Reducing Average Error Fits Majority Populations refers to the phenomenon that machine learning algorithms trained with accuracy or a similar error measure tend to favor correctly learning larger groups over smaller groups. Fitting larger subgroups reduces overall error faster than fitting smaller groups and hence minorities are disadvantaged. This effect can be measured and somewhat reduced by additional data [41]. To this end, we split the original dataset records as defined in Section 8.4.1 and used point of release in each arrest-release cycle as a point of time to make the parole decision. This results in gaining many points of time of release in related arrest-release cycles (historic cycles) for predictions and hence increases accuracy and other statistical measures for both Caucasian and African American cohorts. Furthermore, we added past criminal activities by including different numbers of past arrest-release cycles in different experiments as described in Section 8.4.1 and compared in the results listed in Tables 7.1–7.6. *The Need to Explore* refers to the observation that identical settings tend to not translate to different problems, requiring the experimental inves-

tigation of different options to determine the best system for the problem. In general, the training data depends on past algorithmic actions; for example, one can observe recidivism only if a suboptimal decision was taken to release a recidivistic offender. To this end, we considered all arrest cycles of a given offender and not just the 1994 release cycle and labeled our data using subsequent adjudications (instead of future arrests, which could be more biased). Using each arrest-release cycle as a standalone record gave us access to many suboptimal decisions where people did re-offend. In contrast, two studies that we compare our results with [15, 18] used only 1994 data for training and testing purposes, even though they used historical arrests features in the records. Since the dataset contains data only up to approximately 1997, this could have been a limitation if we included only 1994 records for training and testing.

Bias and Fairness Metrics Work: Many fairness measures are used to measure unfairness in AI-based systems [28, 42, 43, 11]. Bias in machine learning-based systems has been observed in many works such as [10, 29, 35, 25, 11]. Very often, the values of pertinent metrics like FPR, FNR, etc are observed in sets for different sub-populations. The difference in their values indicates the existence of bias, for example in terms of higher FPR value and lower FNR value for a disadvantaged group and vice versa for an advantaged group in recidivism. Similarly, bias shows up as lower FPR value and higher FNR value for a disadvantaged group and vice versa for an advantaged group in hiring decisions and loan applications. However, a general, quantitative measure of bias for a predictive model that can be used for any statistical measure has been missing so far. BP score, a metric we introduce in this work, fills this gap. A BP score of 100 for FNR means that the FNR score for the two cohorts being considered is identical and hence indicates perfect FNR fairness. Accordingly, a BP score of 90 is much better than one of 80. BP score offers a decision maker a way to quantify bias in a model to compare models and to accept or to reject it

for decision making by observing one quantifiable fairness measure for each of the pertinent statistical measures.

In this dissertation, we abstract an existing intuition and summarize bias in a lone measure, namely bias parity (BP) score of any statistical measure, which is the ratio of that statistical measure between two subsets of the population. We use the lower statistical measure value in the numerator and the greater one in the denominator to achieve a symmetric measure for a pair of sub-populations. By multiplying this fraction by 100, we express it as a percentage. This is maximized (i.e., reaches 100%) when there exists no bias (i.e., full parity) with respect to the employed measure. The BP score has the virtue of being a measure between 0 and 100 expressing full bias (0) versus no bias (100). By using BPS, we are able to move from a binary fairness notion of (fair/unfair) to a fairness score for any statistical measure deemed important for fairness in a given situation, particularly as we take into consideration the results demonstrated by Chouldechova that, as recidivism prevalence differs across subpopulations, all fairness criteria cannot be simultaneously satisfied [42]. The work in [42] illustrates how disparate impact can occur when error rate balance fails for a recidivism prediction instrument.

Work by Krasanski et al. [44] states the p% rule [45], an empirical rule that proscribes sensitive group identification from being less than a percentage of the favored group identification. For this rule, there is a legal context and the Uniform Guidelines on Employee Selection Procedures [46] mandate adherence to the 80% or more rule. BP score gives the practitioner the prerogative to decide the ideal threshold for BPS of a requisite metric, e.g., FPR and FNR BPS in recidivism, particularly as the ability to collect pertinent data and improve algorithms further progresses. Many machine learning researchers continue to use the 80% threshold. For example, a work by Feldman et al. [36] states that the Supreme Court has resisted a "rigid

mathematical formula" to define disparate impact. Feldman et al. adopt the 80% rule recommended by the US Equal Employment Opportunity Commission (EEOC) [46] to specify whether a dataset has disparate impact. Feldman et al. link the measure of disparate impact on the balanced error rate (BER) and show that a decision exhibiting disparate impact can be predicted with low BER. Given a dataset $d = \{(X, A, Y)\}$, with X representing the non-sensitive data attributes, A the sensitive ones, and Y the data element's correct binary label, and a classification function $f(X)$, BER can be defined as:

$$BER(f(X), Y) = \frac{P[f(X) = 0|Y = 1] + P[f(X) = 1|Y = 0]}{2}.$$

A data set d = {(X, A, Y)} is ϵ-fair if for some classification algorithm, f : X $\rightarrow$ Y , BER(f(X), Y) $\leq \epsilon$.

Feldman et al. [36] define Disparate Impact ("80% rule") by stating that for a given data set d, d ={(X, A, Y)}, with protected attribute A (e.g., race, sex, religion, etc.), remaining attributes X, and binary class to be predicted Y (e.g., "will hire"), d has disparate impact if

$$\frac{P(Y = 1|A = 0)}{P(Y = 1|A = 1)} \leq \tau = 0.8$$

Similarly, for a prediction C of Y, the classification has disparate impact if

$$\frac{P(C = 1|A = 0)}{P(C = 1|A = 1)} \leq \tau = 0.8$$

Krasanakis et al. [44] employ D_{FPR} and D_{FNR} as the differences in FPR and FNR, respectively, of the protected and unprotected group while computing the overall disparate mistreatment. They combine those two metrics into $|D_{FPR}| + |D_{FNR}|$.

Given the predicted classification output is $C = f(X,\ A)$, the differences are represented as

$$D_{FPR} = P(C \neq Y \mid Y = 0, A = 1) - P(C \neq Y \mid Y = 0, A = 0)$$

$$D_{FNR} = P(C \neq Y \mid Y = 1, A = 1) - P(C \neq Y \mid Y = 1, A = 0)$$

The related work recognizes that the presence of bias can be represented as a ratio or the difference of a statistical measure for the two pertinent cohorts. When laid out side by side for several statistical measures, as shown in Tables 7.1–7.6, BP scores can tell us which model has least bias and hence should be selected from a set of possible models. As a generic measure, BPS can be used for all statistical measures and thus offers a unifying technique for representing bias.

Bias Mitigation Works Some recent works [47, 44] summarize bias mitigation techniques into three categories: i) preprocessing [48, 36, 49] input data approaches, ii) in-processing approaches or training under fairness constraints that focuses on the algorithm and, iii) postprocessing approaches that seeks to improve the model.

The first technique involving preprocessing input data is built on the premise that disparate impact in data results in disparate impact in the classifier trained on such data. Therefore, these techniques are comprised of massaging data labels and reweighting tuples of the dataset. Massaging is altering the class labels that are deemed to be mislabeled because of bias, while reweighting involves increasing weights of some of the tuples over the others in the dataset. Calders et. al. [48] assert that these massaging and reweighting techniques yield a classifier that is less biased than without such a process. They also note that while the massaging labels method is intrusive in nature and can have legal implications [31], the reweighting method

on the dataset to make the labels not dependent on the sensitive attribute does not have these drawbacks.

The second technique of fairness under constraints [34] chooses one or more of the disparate impact metrics [50, 34, 51]. This is followed by modifying the imposed constraints during the classifier training or by additional linear program constraints that steer the model towards the optimization goals [34, 51, 50, 52]. For example, in a recent work Iosifidis et. al. [50] change the training data distribution and monitor the discriminatory behavior of the learner. When this discriminatory behavior exceeds a threshold, they facilitate different fairness notions by adjusting the decision boundary to prevent discriminatory learning.

The third technique to improve the results involves a postprocessing approach to comply with the fairness constraints. The approach can entail selecting a criterion for unfairness relative to a sensitive attribute while predicting some target. In the presence of the target and the sensitive attribute, Hardt et. al. [53], show how to adjust the predictor to eliminate discrimination as per their definition.

Definitions of Fairness. Current literature on fairness recommends several formal concepts of fairness which require that one or more demographic or statistical properties are held across multiple subpopulations in the corpus. Demographic parity, also referred to as statistical parity, mandates that the decision rates are independent of the values of a sensitive attribute that represents membership of different subgroups [54, 48, 55, 56]. For binary classification problems this is often mathematically represented as $P(C = 1|A = 0) = P(C = 1|A = 1)$, where $C \in \{0, 1\}$ is the decision made by the system. This, criterion, however, makes an underlying equality assumption between the subpopulations which might not hold for all problems. To address this, several recent works [53] focus on error rate balance where fairness requires subpopulations to have equal false positive rates (FPR) or equal false negative

rates (FNR) or both. Another commonly used parity condition is equality of odds which commands equal true positive rate (TPR) and equal true negative rate (TNR). While perfect parity as a constraint would be desirable, it often is not achievable and thus quantitative measures representing the degree of parity have to be used [57]. Refer to [43, 40, 11] for a more complete recent survey of computational fairness metrics.

CHAPTER 3

Datasets

The work in this dissertation is focused around three different datasets, two of them being in the area of recidivism, and one being a standard Census Bureau dataset focused on income characteristics. Use of these datasets permits the study of different aspects of increasing accuracy while reducing bias in the context of different sensitive attributes. Moreover, they allow effective comparisons with past results.

3.1 Dataset 1: "Criminal Recidivism in a Large Cohort of Offenders Released from Prison in Florida, 2004-2008"

Dataset 1 [58] is used for experiments described in Chapters 4, 5, and 8. It is assembled using the information and resources acquired from the Florida Department of Corrections (FDOC) and the Florida Department of Law Enforcement (FDLE) [59].

The Offender-Based Information System (OBIS) stores and maintains offender related information, such as sentence, termination date, offender location, new violations, DNA profile, status, risk classification etc., while the FDLE's Computerized Criminal History (CCH) database, holds criminal history information pertaining to all arrests made in Florida. The dataset is comprised of information acquired from OBIS and CCH pertaining to offenders released between January 1996 and December 2004. The offenders without criminal history were excluded during the dataset creation, as were the offenders released outside Florida. The FDOC data files provided the current offense information for this dataset.

3.1.1 Dataset 1 Description

The dataset categorizes charges into six groups: violent charges (murder, manslaughter, sexual offenses, and other violent offenses); robbery; burglary; other property charges (including theft, fraud, and damage); drug-related charges; and other charges (including weapons and other public order offenses). The dataset also categorizes crimes ~~by~~into four main categories (MAINCAT): violent, property related, drugs, and all others. The dataset uses FDLE criminal history records for offender rearrests information within 3 years of release while using arrest date for recidivism event purposes and to calculate various time related features. Additionally, FDOC court docket information is used to capture reconviction information. The dataset captures several pieces of information in many ways. We used the attributes listed here: MAINCAT (crime main category), ADMAGE (admission age of offender), RELAGE (release age of offender), TIME SRV (time served in jail by the offender), CHIST (number of crimes committed by the offender prior to being arrested), RACE, MARITAL (offender's marital status), EMPLOY (offender's employment status prior to being arrested), SEX, EDUCLAM (education claimed) and SUPER (whether parolees were supervised or not after release). Many features like MAINCAT, MARITAL, EDUCLAM etc. in this dataset are represented as categorical data, so we converted each attribute value as a column indicating the presence or absence of that attribute value. This removed the bias that can result from encoding the data values using arbitrary categorical values during the normalization process.

The dataset is comprised of 156,702 valid cases, all of which are included in all race models. The ratio of recidivists to non-recidivists in the entire dataset is 41:59. There are proportionately more non-recidivists amongst the Caucasian race than in the African American race in every category. In the entire dataset this ratio for Caucasians was 34:66, while that for the African American sub populace was 46:54.

3.2 Dataset 2: "Recidivism of Prisoners Released in 1994"

The second dataset is from the "Recidivism of Prisoners Released in 1994" study [60]. This dataset is comprised of 38,624 records - one for each convict released in 1994 from one of 15 states in the USA. As these states account for two-thirds of prisoners released that year, the dataset is a good representative of the offenders released in 1994. The individual criminal history information in the dataset is acquired from the State and FBI automated RAP sheets).

3.2.1 Dataset 2 Description

Each record in the dataset is comprised of a convict's 1994 release cycle and past arrest records before the 1994 release and those ensuing in the subsequent three or more years. Each record consists of variables pertaining to a maximum of 99 arrest cycles before and after the 1994 release cycles, each record containing information about arrests, adjudications, and sentences for each of these arrest records. These records maintain 91 fields related to the 1994 release, 64 fields for each of the prior or ensuing arrest cycles, and up to 99 multiples of the 64 fields based on the total number of arrests associated with the offender. Thus, each convict record is comprised of up to 6,427 fields.

There are more Caucasians in this dataset than African Americans and the difference is almost exclusively in records with fewer than 10 arrest cycles. There are more Caucasians with fewer than 10 arrest records than African Americans. As shown in Figure 3.1, after 10 arrest records, the proportion of the two races becomes similar. There are very few offenders with more than 40 arrest records each.

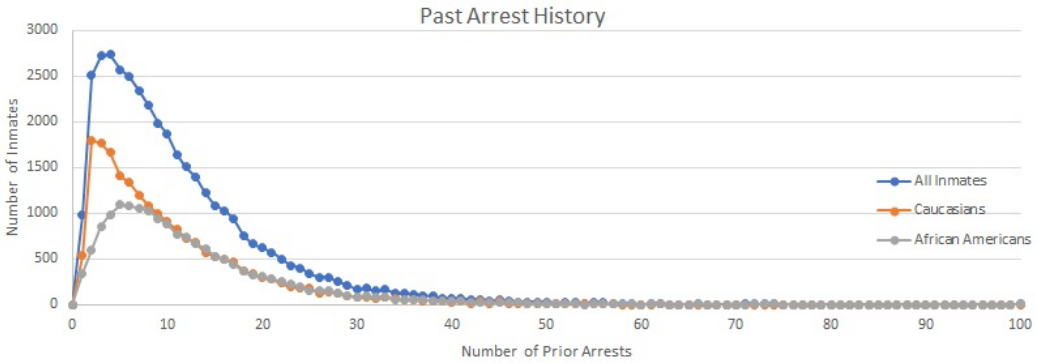


Figure 3.1: The Number of Prior Arrests for Different Sub-populations.

3.3 Dataset 3: "Adult Income Data Set"

Our third datset "Adult Income Data Set" [61] from the UCI repository [62] was extracted from 1994 Census database. It is also known as "Census Income" dataset and is available via the University of California Irvine's (UCI) Machine Learning Repository. The dataset has 48,842 records and as per the UCI website, each record represents an individual who was older than 16, had an adjusted gross income greater than $100, worked for at least an hour a week and represented more than 1 person in the general population during the 1994 census. The dataset holds socioeconomic status, education, and job information pertaining to the individual. Therefore, it is a demographical dataset just like Dataset 1 but in a different domain than Datasets 1 and 2. Unlike Datasets 1 and 2, where race is the sensitive attribute, gender is the sensitive attribute in Dataset 3. For our work, similar to most previous work involving this data set, we deem a salary higher than $50 K a "positive" outcome and less than $50 K to be a "negative" outcome. It is used for predicting whether an individual's income exceeds $50K/yr or not and for related statistical measures.

3.3.1 Dataset 3 Description

The dataset is comprised of 48,842 records that are a mix of 6 continuous and 8 discrete attributes, and one class label. 3,620 records had some attributes missing. However, we imputed these missing values as the mean value of the attribute. The 6 continuous attributes age, fnlwgt, education-num, capital-gain, capital-loss, and age were normalized and the 7 discrete variables were treated using one-hot encoding to convert each attribute value as a binary vector indicating the presence or absence of that attribute value. We used income as a binary target variable and all remaining attributes as input variables.

The ratio of records with class label '>50K' : '<=50K' is approximately 24:76. The ratio of males to female in the sensitive attribute gender is approximately 2:1. There are proportionately more "positive" outcomes (income above 50K) amongst males than amongst females. In the entire dataset the ratio of positive (above 50K) to negative (below 50K) outcomes for males was approximately 9:4, while that for the female sub population was approximately 8:1.

We split the dataset in a 80-20 ratio for training and testing purposes. The training data was further split in a 90-10 ratio for training and validation purposes. The dataset is used in experiments described in Chapter 8.

CHAPTER 4

Singular Race Models: Addressing Accuracy and Bias in Predicting Recidivism

4.1 Singular Race Models (SRM)

Our goal for these experiments was to study the potential of training separate, Singular Race models versus a single, all-race model to increase the accuracy of predicting recidivism while removing the potential for any race from being discriminated against. We conducted five sets of experiments: one for the all-crime category and four for each of the main crime categories. In each set, to start with, we trained a model with data from all races and tested it with three different sets of data encompassing all races, only Caucasian, and only African American. All-race models in each set used records attributed to all races without using a feature vector that stated an offender's race information. By contrast, the Caucasian and African American Singular Race models used data solely from the subpopulation for training, again without including the race feature vector. The all-race models in each crime category constituted our baseline to compare the results of the Singular Race Models. In the same vein, the crime related models include violent, property, drugs, or other crime related subsets only, without the race feature being included in the dataset. The various crime related base models were trained on that particular crime-based data pertaining to all races and tested on data pertaining to Caucasian and African American races without including the race feature in the training or test data while having the Singular Race Model dataset restricted to one race at a time.

We wanted to explore if models trained for a sensitive attribute value (Caucasian or African American race) would prove to have higher predictive accuracy for the

favored sensitive value over a model trained for all possible values of a sensitive attribute. Our thought process was that comparing test subjects with others who were similar to them in the training data, would lead to higher predictive accuracy. Additionally, we believe that by comparing individuals to others just like them or of the same race and committed crime category, we might reduce or eliminate bias resulting from the unbalanced nature of the complete data set dramatically.

We used only Caucasian and African American race models because they represent over 98% of the dataset - where 42.6%, 55.6%, 1.6% of the dataset is attributed to Caucasians, African Americans, and Hispanic races, respectively, while all other races combined constitute 0.2% of the dataset.

Another recent study by Ozkan [18] investigated numerous statistical models to improve predictive accuracy of recidivism prediction. It compared a conventional logistic regression model with other machine learning models like random forests, support vector machines, XGBoost, neural networks, and search algorithms and found XGBoost and neural networks to outperform all other models. This guided us to include artificial neural networks for our experiments with Singular Race Models. We selected neural networks, the classifier that gave the best accuracy with our dataset.

We used a three-layered deep learning neural network model to develop, train, and explore these new models. We used the Keras wrapper for low-level libraries like TensorFlow [63] to program and develop our models. Adam [64], an optimization algorithm for stochastic gradient descent, was utilized for training our deep learning models.

4.2 Datasets

Dataset 1 described in Section 3.1 was used for the experiments included in this chapter. It is the raw data from the studies "Recidivism of Prisoners Released

in 1994" [60]. It was assembled using the information and resources acquired from the Florida Department of Corrections (FDOC) and the Florida Department of Law Enforcement (FDLE) [59]. A detailed description of the dataset can be found in Chapter 3.

4.3 Experiments

Our goal for these experiments was to increase the accuracy of predicting recidivism while removing the potential for one race from being discriminated against. We conducted five sets of experiments: one using the all-crime category and additional ones to predict the four main crime categories. In each set, to start with, we trained a model with all races and tested it with three different sets of data, namely ones comprised of all races, only Caucasians, and only African Americans. All-race models in each set used records attributed to all races without using a feature vector that stated offender's race information. Similarly, the Caucasian and African American models used data solely from the subpopulation without including the race feature vector explicitly. The all-race models in each crime category constituted our baseline to compare the results of Singular Race Models. In the same vein, the crime related models were exposed to violent, property, drugs, or other crime related subsets only, without the feature of race being included in the dataset. The various crime related base models were trained on that particular crime-based data pertaining to all races and tested on data pertaining to all races, Caucasian and African American races without including the race feature in the training or test data while having the Singular Race Model dataset restricted to one race at a time.

We wanted to explore if models trained for a sensitive attribute value (Caucasian or African American race) would prove to have higher predictive accuracy for the favored sensitive value over a model trained for all possible values of a sensitive

attribute. Our thought process was that comparing test subjects with others who were similar to them in the training data, would lead to higher predictive accuracy. Additionally, we thought that by comparing individuals to others just like them or of the same race and committed crime category, we will reduce or eliminate bias dramatically.

We used only Caucasian and African American race models because they represent over 98% of the dataset - where 42.6%, 55.6%, 1.6% of the dataset is attributed to Caucasians, African Americans, and Hispanic races respectively, while all other races combined constitute 0.2% of the dataset.

A study [18] investigated numerous statistical models like logistic regression model with other machine learning classification models like random forests, support vector machines, XGBoost, neural networks, and Search algorithm to improve predictive accuracy of recidivism prediction. It found XGBoost and neural networks to outperform all other models [18]. This guided us to experiment with several models - artificial neural network, K-nearest neighbors (k=10 and 5), Random Forests, AdaBoost, Decision Tree, and Support Vector Machines for comparison to see which of these provided the highest accuracy and could be used for further experimentation with Singular Race Models. We conducted ablation analysis to understand the effectiveness of each of the features in the dataset. This study showed that MAINCAT, SUPER, MARITAL, EMPLOY, RELAGE, ADMAGE, TIME_SRV, CHIST, EDUCLAM features brought out the best prediction accuracy. Even though adding RACE and ETHNICTY increased the predictive accuracy of artificial neural networks, we excluded these and sliced the dataset by Caucasian and African American races for various crime related models.

Table 4.1: Experiment Results for Various Classification Models

All Crimes All Races	**Artificial Neural Network**	KNN(10)	Random Forest	AdaBoost	Decision Tree	Support Vector
Accuracy	**0.652**	0.625	0.649	0.649	0.640	.646
FPR	0.235	0.196	0.176	0.209	0.242	.184
FNR	0.513	0.637	0.607	0.557	0.533	.603

4.3.1 Selecting the Best Classifier

In order to select a classifier for Singular Race Models, we compared the prediction accuracy of several conventional machine learning models (Table 4.1) like K-nearest neighbors (k=10 and 5), random forests, AdaBoost, Decision Tree, support vector machines, and artificial neural network. For finding the accuracy of each of these models, we used the entire dataset comprised of 156,702 valid records with 80% used for training and the rest for testing the classification model. We left out RACE, ETHNICITY, and GENDER from the dataset. Only MAINCAT, SUPER, MARITAL, EMPLOY, RELAGE, ADMAGE, TIME_SRV, CHIST, EDUCLAM features were used for the best classification model selection. Since artificial neural networks had the highest predictive accuracy of 0.652, it was selected for further experiments in Singular Race Models. At each step, we wanted to note not merely the accuracy but also recidivism rates in the dataset, as well as the False Positive Rate and False Negative Rate in the predictions as these represent measures of the bias in the results.

The Artificial Neural Network that we selected for constructing Singular Race Models accepts inputs, and processes it in hidden layers. The neural networks use and adjust weights during the training phase to detect patterns in the input data to predict a classification decision [65]. These weights are then used to predict classification results for the test data. For this, we used Keras, an open-source python library.

The K-nearest Neighbors (KNN) algorithm is a supervised machine learning algorithm that uses the K closest training instances to predict the class of the test record. We experimented with both K= 10 and 5, but chose K=10 due to its higher accuracy. We used Scikit-learn's KNeighborsClassifier model for our experiments.

Random Forests is an ensemble learning method [66]. The algorithm constructs and fits many decision trees in the training phase and uses the aggregate of the classification decisions to label the test record. Aggregation increases the accuracy and the algorithm is strong against overfitting. We used Scikit-learn's RandomForestClassifier model with default settings for our experiments.

AdaBoost is a machine-learning algorithm that combines the results of weak-learner classifiers [67] using weighted sum to create a strong learning algorithm. The weak-learning classifiers are tweaked in a way to boost the occasions where they accurately label the previously mislabeled instances. We used Scikit-learn's AdaBoost-Classifier model for our experiments.

Decision Tree, a machine learning modeling technique, accepts various input variables to predict an outcome. It graphically represents potential solutions based on various conditions. Each leaf of the tree represents a target class for the classification problem. We used Scikit learn's DecisionTreeClassifier model with max_depth equal to 6 for our experiments.

Support Vector Machines, a machine learning algorithm, accepts input to fit it and find a hyperplane that best classifies or rather divides the data into different classes. The features of the test records can then be used to locate its predicted class based on its location relative to the hyperplane. We used Scikit learn's LinearSVC model for our experiments.

4.3.2 Singular Race Models

We built Singular Race Models, a novel way of splitting the dataset one race at a time, to train and test single race-based models and to increase prediction accuracy without setting one race against another. In our experiments with several classification models, artificial neural networks had the highest accuracy so we used them to generate Singular Race Models for four different crime categories and to place their results side by side and compare these with base models created using all crimes and all races.

Our dataset did not contain offender related personal information like academic courses attended, vocational training or months of internship completed during incarceration or after release on parole. This propelled us to design Singular Race Models that could use the demographic information one race at a time and thereby increase the accuracy without putting any one race at a disadvantage.

4.3.3 Artificial Neural Networks

Neural networks learn patterns in the input data as the human brain does and use multiple interconnected neural units to do so, thereby mutually affecting each other's activation state. During the process of neural network model building, we experimented with different values for various hyperparameters like number of neurons, number of layers, batch size, activation, optimization etc. to find the best results. For example, we altered batch size and epochs values to several thousands. However, we found that increasing batch size beyond 250 and epochs beyond 25 did not increase the prediction accuracy. Therefore, we settled for a batch size and epochs value of 250 and 25, respectively. The results of experimenting with different neural networks parameter values led us to a three-layered topology neural structure with 32 neurons in the input and in the hidden layer, and finally 1 neuron in the output

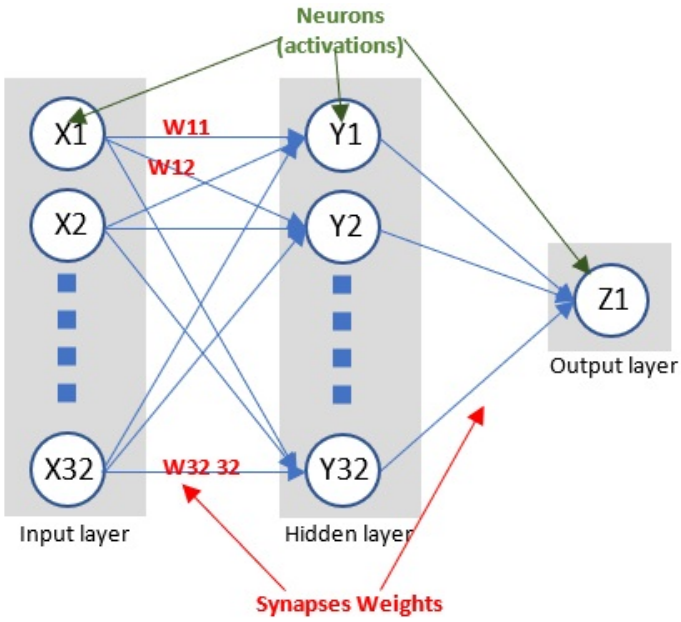


Figure 4.1: The Number of Prior Arrests for different sub-populations.

layer as shown in Figure 4.1. We used the dropout regularization technique to prevent overfitting so that randomly selected nodes were dropped out with a 10% probability during each cycle as the neural network updated the weights associated with each of the neural units while the network learnt using backpropagation. We used the Keras wrapper for low-level libraries like TensorFlow to program and develop our models. Adam, an optimization algorithm for stochastic gradient descent, was utilized for training our deep learning model. This being a binary classification problem (recidivist or not), we chose the logarithmic loss function binary_crossentropy for training the model. We used ReLU activation function in the first two layers and Sigmoid in the output layer.

4.3.4 Bias Metrics for Experiment Results

Besides investigating the effect of using Singular Race Models on accuracy of predictions, we also wanted to study their effect on bias in the predictions. Bias

metrics used in this chapter are TP, TN, FP, FN, Accuracy, Null Accuracy, FPR, and TPR. These are explained in detail in Chapter 2.

4.3.5 Experiment 1: All Crimes

We wanted to increase the accuracy of prediction by using race information and yet remove the results' dependence on race. By entirely removing the sensitive attribute of race, we could be fair, we thought, but also lose accuracy by losing pertinent information. So, we trained our models in a way that the information regarding race could not be used against any one. In order to do so, we trained three different models: All_crimes, All_crimes_caucasian, and All_crimes_africanAmerican.For training the All_crimes model, we used a subset of the data without distinguishing by race. Three test results from this model are comprised of subsets of all races, Caucasian only, and African American only subpopulations.

All_crimes_caucasian, and All_crimes_africanAmerican models are trained and tested using members of only the Caucasian and African American subpopulations, respectively. In each of these three models the training data pertained to all crime categories.

The All_crimes model was trained on eighty percent of the dataset containing the following information: MAINCAT, SUPER, MARITAL, EMPLOY, RELAGE, ADMAGE, TIME_SRV, CHIST, EDUCLAM from the original dataset. We did not use directly race related features vector in any of the training or test datasets. For the Singular Race Models, we segmented the dataset by race and provided a homogenous dataset devoid of any race related features in the dataset to the model. We split the dataset 80% for training and 20% for testing purposes and the experimental results are depicted in Table 4.2. The All_crimes_caucasian dataset was comprised of all features in All_crimes data set but limited to the data records where the race was

Table 4.2: Experiment Results for All Crimes

All Crimes	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.406	0.341	**0.457**	0.341	**0.457**
Non-recidivists	0.594	**0.659**	0.543	**0.659**	0.543
Null Accuracy	0.594	**0.659**	0.543	**0.659**	0.543
Accuracy	0.652	0.675	0.633	**0.685**	0.638
FPR	0.235	0.180	0.290	0.121	**0.345**
FNR	0.513	0.603	0.458	**0.692**	0.383

Caucasian. Similarly, the African American dataset was comprised of all features in All_crimes data set but limited to the records where the race was African American.

4.3.6 Results of Experiment 1: All crimes

In Experiment 1 for All_crimes (Table 4.2), the accuracy for the model with all the races was 0.652. This model predicted recidivism with an accuracy of 0.675 and 0.633 for the Caucasian and African American subpopulations, respectively. The All_crime_caucasian model has a higher prediction accuracy as compared to the baseline at 0.685 but shows less improvement when one compares the models' null accuracies with their accuracies. The All_crimes_caucasian model has a higher FNR as compared to the baseline. A high FNR represents recidivists that are erroneously released on parole and thus represents more preventable crimes. The All_crimes_africanAmerican model has an accuracy of 0.638, but it improved the most amongst the three models - particularly as one compares the null accuracy with the accuracy: The All_crimes model improved from 0.594 to 0.652, the Caucasian model from 0.659 to 0.685, and the African model from 0.543 to 0.638.

Table 4.3: Experiment Results for Violent Crimes

Violent Crimes	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.314	0.232	**0.381**	0.232	**0.381**
Non-recidivists	0.686	0.768	0.619	**0.768**	0.619
Null Accuracy	0.686	0.768	0.619	**0.768**	0.619
Accuracy	0.702	0.758	0.656	**0.769**	0.654
FPR	0.113	0.073	0.154	0.031	**0.167**
FNR	0.700	0.802	0.652	**0.894**	0.637

The increase in FPR in the All_crimes_ africanAmerican model as compared to the baseline represented that the African American race model would keep more non-recidivist African American offenders incarcerated as compared to the baseline model and the All_crimes_caucasian model.

When comparing Singular Race Models, FNR for the African American race model is lower than that of the Caucasian race model. A relatively lower FNR of the African American race model means that proportionately more potential future crimes perpetrated by African Americans will be prevented as compared to those committed by Caucasians.

One should also notice that the base model predicts recidivism with higher FPR and lower FNR for African American race than for Caucasian race. The magnitude of this bias increases in Singular Race Models even though our original hypothesis was that if the neural network learns from people from the same group, the prediction accuracy will be higher, and bias will be lower. The results showed that the accuracy did improve but the bias also increased in the direction established in the base model. Since the base model already represented the bias in the absence of race information, it means that several dataset features are strongly correlated with race.

4.3.7 Experiment 2: Violent Crimes

We repeated the same experiments as in Experiment 1 but restricted the records to those pertaining to violent crimes.

4.3.8 Results of Experiment 2:

In Experiment 2 for violent crimes (Table 4.3), the accuracy for the model with all the races was 0.702. The Singular Race Model for the African American race in this category is the only one in all experiments that failed to improve the accuracy with respect to the base models. Violent_crimes_caucasian model has higher accuracy than the base line but just like Experiment 1, improved less when null accuracy is compared with the accuracy. Additionally, the FNR of the Caucasian model increased from 0.802 to 0.894. A high FNR represents violent crimes that society would like to be spared off so prediction using Violent_crimes_caucasian model for Caucasians means 89% of the Caucasian recidivists will be declared non-recidivists and released on parole. Violent_crimes_africanAmerican model has a lower accuracy, but it has learnt more than the other two models in the category (comparing accuracy with null accuracy). Additionally, the FPR for Violent_crimes_african American model is higher (worse) than for the other two models in the category and FNR decreased (improved) as compared to the base line model. This combination of FPR and FNR results essentially means that prison population of African American violent crime offenders will increase while that of Caucasians will decrease.

The base model indicated bias even in the absence of race information, suggesting that several dataset features are strongly correlated with race.

4.3.9 Experiment 3: Property Crimes

We repeated the same experiment as in Case 1, for property crimes.

Table 4.4: Experiment Results for Property Crimes

Property Crimes	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.472	0.420	**0.538**	0.420	**0.538**
Non-recidivists	0.528	**0.580**	0.462	**0.580**	0.462
Null Accuracy	0.528	**0.580**	0.538	**0.580**	0.538
Accuracy	0.612	0.613	0.612	**0.618**	**0.618**
FPR	0.345	0.311	0.399	0.273	**0.518**
FNR	0.436	0.493	0.379	**0.534**	0.266

4.3.10 Results of Experiment 3:

In Case 3 for property crimes (Table 4.4), the accuracy for Singular Race Models increased as compared to the Property_crimes_all_race model. All three models represent very different rates of improvement from the data (comparing null accuracy and accuracy). The Property_crimes_caucasian model learnt the least amongst the three models. Additionally, the FNR of Property_crimes_caucasian model increased to 0.534. The FNR for the base model was already higher for the Caucasian subpopulation (0.493) than the African American subpopulation (0.379) and it increases further for the Singular Race Models for Caucasians to 0.534, while it decreased to 0.266 for African Americans.

A high FNR here represents preventable property crimes. The Property_crimes_africanAmerican model's FPR is higher (worse) than that of the other two models which means that when used, this model will keep more African American offenders incarcerated even when they will not re-offend. A higher FNR of the Caucasian model represents that Caucasians likely to reoffend after Property related crimes will be released in over 53% of the cases while 27% of the African property crime related reoffenders will be released.

Table 4.5: Experiment Results for Drug Crimes

Drug Crimes	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.441	0.327	**0.482**	0.327	**0.482**
Non-recidivists	0.559	**0.673**	0.518	**0.673**	0.518
Null Accuracy	0.559	**0.673**	0.518	**0.673**	0.518
Accuracy	0.639	0.690	0.621	**0.700**	0.626
FPR	0.244	0.119	0.301	0.100	**0.341**
FNR	0.508	0.702	0.463	**0.712**	0.409

Thus, we observe that the Singular Race Models for both Caucasian and African American race models increased the accuracy of recidivism prediction. However, the bias measured by FPR and FNR observed in the base line model in the two subpopulations increased further in the Singular Race Models.

The property crime related base model represented bias even without race information indicating dataset feature correlation with the missing race feature in this category, too.

4.3.11 Experiment 4: Drug Crimes

We repeated the same experiment as in Case 1, for drug crimes.

4.3.12 Results of Experiment 4:

In Experiment 4 for drug crimes (Table 4.5), the recidivism prediction accuracy of the Singular Race models increased as compared to the base model and its treatment of the subpopulations.

The Caucasian model had 1% better accuracy than the base model, 1% lower FPR (good) and 1% higher FNR (not so good). The accuracy of the African American related Drug crimes model also improved over the base model. FNR decreased to 0.409

Table 4.6: Experiment Results for Other Crimes

Other Crimes	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.415	0.348	**0.478**	0.348	**0.478**
Non-recidivists	0.585	**0.652**	0.522	**0.652**	0.522
Null Accuracy	0.585	**0.652**	0.522	**0.652**	0.522
Accuracy	0.631	**0.654**	0.610	**0.654**	0.613
FPR	0.263	0.189	0.350	0.157	**0.409**
FNR	0.518	0.640	0.425	**0.714**	0.363

and this was an improvement but the FPR increased (worsened) to 0.341. Moreover, in drug related crimes, too, one can expect the singular race model for Caucasians to release 70% of the re-offenders to re-offend.

In this crime category, too, the presence of bias in the base model represents the dataset features' correlation with race.

4.3.13 Experiment 5: Other Crimes

We repeated the same experiment as in Case 1, but with other crimes.

4.3.14 Results of Experiment 5: Other Crimes

In Experiment 5 for other crimes (Table 4.6), the results are similar to the crime models covered in Experiments 1 through 4.

The African American model learnt more than the other models despite having a relatively lower accuracy. Just like in previous cases, here too, FPR worsened (increased) and FNR improved (decreased) as compared to the subpopulation treatment in the base model. The base model showed bias even in the absence of race information.

4.4 Discussion

In this study we segmented the dataset by race to create Singular Race Models - models that trained and tested using a single race at a time with the goal being to increase accuracy and reduce bias by comparing people with those who were similar to them and even committed similar crimes. The base models were developed using records of people from all races. All Singular Race Models except for the African American model in the violent crime category had higher accuracies as compared to the base model of their crime category. However, the magnitude of bias in all Singular Race models increased with respect to that established by the base models.

The fact that bias was present in the two race-based cohorts even in the absence of race information and even increased in magnitude in the preexisting direction in the presence of a more homogenized race-based population, primarily means that many features present in the dataset are correlated to race. Secondly, it means that there are several pieces of information that influence the outcome but are not captured in the dataset. These additional factors must be correlated in a unique way to one sub group and differently to the other to influence the results so strongly and hence the system can figure out the memberships of the subjects even in the absence of race information.

The Caucasian race-based models showed the highest accuracies but the least improvement in accuracy with respect to the null accuracy in all sets of experiments. This means that the system could learn significantly about the Caucasian offenders even from an all-race group of offenders. However, the disadvantaged group, the African American race, which had higher FPR and lower FNR in the base model, had the highest increase in accuracy as compared to the null accuracy. This means that the model had most to learn from the Singular Race Model for this cohort.

However, the bias in the African American Singular Race Model also increased in the direction established by the base model.

One essential characteristic of the dataset is that it is based on demographic information only and is completely devoid of the personal information regarding what the offender might have done before, during or after incarceration like courses attended, events attended, new skills acquired etc. - and this makes it hard for the artificial neural network to learn from anything other than the demographic information and hence the bias in the results.

The dataset that we used had a more balanced recidivism to non-recidivism ratio amongst African Americans than amongst the Caucasians and this had an important role to play in the bias in terms of FNR and FPR in the result. Since the African American subpopulation data showed more recidivism than the Caucasian related data, the test data for the former group showed higher FPR and lower FNR and vice-versa in the latter subpopulation in the base multi-race model. This bias was further increased in the Singular Race Models and the test population showed a further increase in FPR and decrease in FNR in the African American models while demonstrating a decrease in FPR and a decrease in FNR in the Caucasian models.

In simple terms, it means that the higher recidivism amongst African American offender data also leads to higher attribution of recidivism to members of the group. Lower recidivism in the Caucasian data leads to lowering the ascribing of recidivistic behavior to members of this subgroup. We believe that one way to mitigate bias would be to go beyond the demographics in the dataset and include more features representing personal information regarding the offenders.

4.5 Conclusion

In this study we focused on creating Singular Race Models - separate race-based models that use a single race to train and test recidivism prediction. The homogenous nature of training and test data was utilized with the intention to generate models that would provide us with higher predictive accuracy and minimize the predictive bias as they compared and learnt from people similar to each other and even commit similar kinds of crimes. We found that all Singular Race Models increased the predictive accuracy as compared to the base model. The only exception to this rule was the African American based violent crime model that did not improve the already elevated accuracy. The data for this crime category was the most unbalanced amongst all crime categories and highly skewed with fewer recidivists than non-recidivists in each category. The side effect of higher accuracy via Singular Race Models was that the magnitude of bias of the prediction demonstrated in the base models was increased further.

The ratio of recidivists to non-recidivists in the all-crime dataset, the Caucasian dataset, and the African American dataset, is approximately 41:59, 34:66, and 46:54, respectively. This ratio is skewed more amongst the subpopulations for various crime related groups. Therefore, in a future work, it would make sense to use Singular Race Models with more balanced datasets to verify its influence on the bias in the results. It will also make sense to use other machine learning algorithms to test Singular Race Models. Additionally, the dataset used here has limited features. Having a richer dataset with more offender activities before, during, and after incarceration - such as vocational trainings completed, may also turn out to be a major force in producing better Singular Race Models capable of predicting recidivism with a higher accuracy.

In this component of the dissertation, we looked at the dataset with the Machine Learning lens and found that the Singular Race Models increased the accuracy of

predicting recidivism. We also found that grouping individuals by races and the crime category increased the accuracy for each subpopulation in most cases. The presence of bias even in the absence of race leads us to ask which features are strongly correlated with the race and most importantly, why are they correlated? We believe that investigating this correlation and finding the factors that lead to this correlation will lead us to contribute to effective prison reforms and hence we would like to explore these in our future work.

CHAPTER 5

Including Activity Information to Reduce Bias and Increase Prediction Accuracy

5.1 Introduction

We seek to increase accuracy and reduce the bias by increasing offender activity-based information in the input feature set. To achieve this, we aggregated the previous convictions for each of the seven broad crime categories committed before each arrest cycle and used this for predicting recidivism at the end of each arrest cycle. Additionally, we organized our experiments by output variables for eight crime categories and by two races. We used Singular Race Models (SRM) to increase accuracy and to measure and analyze bias in results.

The four main contributions of the work included in this chapter are:

1. We studied the effect of offender's personal activity information on the accuracy of predicting recidivism and use Singular Race Models to study the effect on race-based bias in the predicted results.
2. We examined the accuracy and bias in the results of prisoner recidivism predictions via Singular Race Models for eight crime categories: convictions (for any crime), fatal, sexual, general, property, drug, public, and other crimes.
3. We compared the results of the current work with a work based on a demographic information-based dataset(Dataset 1) to observe and display the change in race based bias in the presence of personal criminal history, substance-abuse, and related treatment data when used in conjunction with Singular Race Models.

4. We analyzed the results and discussed the significance of the type of data in increasing the overall accuracy and fairness in prediction results of each individual crime category and all the crime categories together.

5.2 The Dataset

In this section we describe Dataset 2 [60] that we use for our current work along with Dataset 1 [58], with which we often compared our current work. This is followed by a sub section on dataset preparation, a sub section on input feature selection, and another one on target variables. This section ends with a description of the dataset limitations.

5.2.1 Data Availability

The dataset used in this study are the raw data from the studies "Recidivism of Prisoners Released in 1994" [60] and "Criminal Recidivism in a Large Cohort of Offenders Released from Prison in Florida, 2004-2008" [58]. Due to the sensitive nature of the data, ICPSR prevents us from posting the data online. The datasets analysed during the current study are available in the ICPSR repository. [1] [2]

5.2.2 Basic Dataset Description

The dataset used in this research came from Dataset 2 [60], a study based on "Recidivism of Prisoners Released in 1994". This report contains information pertaining to 38,624 randomly sampled prisoners out of a total of 302,309 prisoners released on parole in 1994 from U.S. prison systems in 15 States. Since the 302,309 prisoners released represent two-thirds of the prisoners released nationwide in 1994,

[1] https://www.icpsr.umich.edu/icpsrweb/NACJD/studies/27781

[2] https://www.icpsr.umich.edu/icpsrweb/NACJD/studies/3355

the dataset represents the released prisoner distribution in 1994 relatively well. The criminal history related data covered in the dataset is assembled from State and FBI automated RAP sheets. This dataset includes information concerning arrests, adjudications, and sentences pertaining to each crime cycle associated with each prisoner released in 1994.

This dataset maintains information regarding the prisoners released in 1994 from 15 U.S. states (Arizona, California, Delaware, Florida, Illinois, Maryland, Michigan, Minnesota, New Jersey, New York, North Carolina, Ohio, Oregon, Texas, and Virginia). It logs their criminal history for 3 or more years following their release in 1994. The 1994 offender release records include arrest cycles prior to and following the 1994 release cycle. This dataset is comprised of 38,624 rows - one per offender - and has 6,427 columns. Each row holds 91 fields pertaining to 1994 related features. Each row also holds 64 fields for other arrest cycles and up to 99 multiples thereof. Each record can have a max of 99 arrest cycles related to an offender.

Whenever possible, we have compared our current results from the activity based Dataset 2 [60] with those described in the study in Chapter 4 obtained from a different largely demographic information based Dataset 1 [58] by crime category as described in the previous chapter. This previous study does not have experiments for Sexual, General, and Public crime categories due to limitations of underlying Dataset 1. Hence, we were unable to compare the results of these crime categories using Dataset 2 [60]. As discussed in Chapter 4, the previous study also uses artificial neural networks to generate base models and SRMs with similar bias metrics. Dataset 1 [58] was compiled using Florida Department of Corrections (FDOC) and the Florida Department of Law Enforcement (FDLE) resources [59] and is comprised of demographic information pertaining to offenders released between January 1996 and December 2004 in Florida. In the comparison experiments, MAINCAT, SUPER,

MARITAL, EMPLOY, RELAGE, ADMAGE, TIME_SRV, CHIST, EDUCLAM features from Dataset 1 were used. These represented the main crime category, whether supervised or not, marital status, employment status, offender's release and admission age, time served in jail, number of crimes committed before arrest, and education level respectively, thus largely representing demographic information with very limited personal attributes for each offender.

5.2.3 Dataset Preparation

Figure 5.1: Data Preparation by Splitting Dataset 2 Records.

To work with the more extensive Dataset 2 and to increase the amount of personal information available for the prediction system, we split our dataset records to create one record for each arrest cycle as shown in Figure 5.1. Thus, if a prisoner was released in 1994 and had had no other arrests, we created one record out of that. Additionally, we created as many standalone records as the number of arrest cycles. After the split, we rearranged each offender's arrest in the chronological order as in the original [60] dataset. The 1994 record was always placed at the beginning of the

record, followed by all other 0-99 arrest cycles. We used all the arrest cycles prior to any given one to summarize each type of offense in the eight broad categories (convictions (for any crime), fatal, sexual, general, property, drug, public, and other category) and added these to each record that corresponded to one arrest cycle. In other words, as per our encoding, each of these newly minted arrest cycle records held information about the number of times the offender had committed different types of crimes in any of the prior cycles. We treated each arrest cycle as a point of time when a decision to release the offender needs to be taken. We used each of these records to calculate personal criminal activity-based input features for the next arrest cycle of a given offender. Thus, each of these derived criminal-activity based features could now contribute towards an increase in the prediction accuracy. Our original Dataset 2 [60] contained at most the three most serious judgement charges in each cycle. Each of these three crimes belonged to one of the seven broad crime categories. So, for each record (arrest cycle originally), we added seven input variables to hold the total number of times an offender had been convicted for each of the seven crime categories. An eighth variable counted the number of times an offender had been convicted before. We incremented this eighth variable by one, each time an offender was convicted in a previous arrest cycle - irrespective of whether it was for one or more of the seven broad offense categories. We used these eight variables to serve as additional input features in each arrest cycle. In our experiments, we used all the records to calculate the criminal activity-based information for the subsequent arrest records but used only the records associated with the 1994 release and beyond for creating our training and testing datasets.

5.2.4 Input Feature Selection

For our experiments, we took several input feature vectors available to us and derived some others to increase the accuracy of predicting recidivism. Specifically, the features used are vocational courses (VOCAT), Educational courses (EDUCAT), HIV positive or not, substance abuse-related variables (DRUGAB, DRUGTRT, ALCABUS, ALCTRT), SEXTRT, number of prisoners represented by this samples case (WEIGHT), dead or undergoing life sentence (DeadOrLifeCnf), Admission Age for each cycle (AgeC), Admission Age for the first arrest cycle (AdAgeC1), crimes adjudicated for (or not) in an arrest cycle (convictions, fatal, sexual, general, property, drug, public, and other category), sum of crimes adjudicated for in previous arrest cycles (CUMcnv , CUMfatal, CUMsexual, CUMgeneral, CUMproperty, CUMdrug, CUMpublic, and CUMother category), number of times in previous arrest cycles, confined(CUMJ001CNF), involved in domestic violence (CUMJ001DMV), convicted (CUMJ001CNV), confined (cumJ001CNF), involved in Fire Arms (CUMJ001FIR) and whether the record was from the 1994 arrest cycle or a later cycle ('after94R). We encoded the values of the categorical attributes using a one-hot vector. This allowed us to eliminate the bias that can result from using specific integer values for each of the features.

5.2.5 Target Variables

We formulated eight recidivism prediction problems by generating eight variables: laterCNV, laterFATAL, laterSEXUAL, laterGENERAL, laterPROPERTY, laterDRUG, laterPUBLIC, and laterOTHER. These were binary variables representing whether an offender was reconvicted for any crime, adjudicated for fatal violence, sexual violence, general violence, property offenses, drug related offenses, public order offenses, and other offenses in any arrest cycle after the given arrest cycle for which

the prediction was being made. Since the 1994 arrest cycle and the associated before and after arrest cycle variables are written differently in the dataset, we read the crime committed from SMPOFF26, for the 1994 records, while for the crime cycles before and after 1994, these were read from J001OFF1, J001OFF2, J001OFF3. In order to calculate the eight target variables, all three crimes in each of the following crime cycles were considered. For example, if an offender had been adjudicated for both fatal violence and property related crime in any of the subsequent cycles, both laterFATAL and laterPROPERTY variable recorded one for that cycle. However, when an offender was adjudicated for a crime category - say fatal crime - but never recommitted a fatal crime again after release, the laterFatal for that arrest cycle was set to 0. Moreover, if an offender was adjudicated for any crime whatsoever in any of the cycles after a certain cycle, laterCNV was set to 1. For the 142,573 records, we recorded the percentage of records that each of laterCNV, laterFATAL, laterSEXUAL, laterGENERAL, laterPROPERTY, laterDRUG, laterPUBLIC, and laterOTHER variables recorded 1 (convicted) or not (0). Figure 5.2 shows the percentage of records that have later crimes for each of the categories. Additionally, for each record we also tracked the maximum number of broad crime categories the offender committed after release. These crimes could be committed in several arrest cycles/records after the one that we were considering. Figure 5.3 shows the distribution of the number of crimes considered in each of the records.

Even though each of the crime categories could be subdivided further, we did not narrow our search any further lest we reduce the size of our dataset too much based on the crime category. Besides, limiting the crime categories also allowed us to align our work with [16] and [15] with similar major crime categories. Unlike [15], who chose arrest for any crime as opposed to conviction, we chose to use adjudication to establish whether someone recommitted a certain kind of crime. We based this

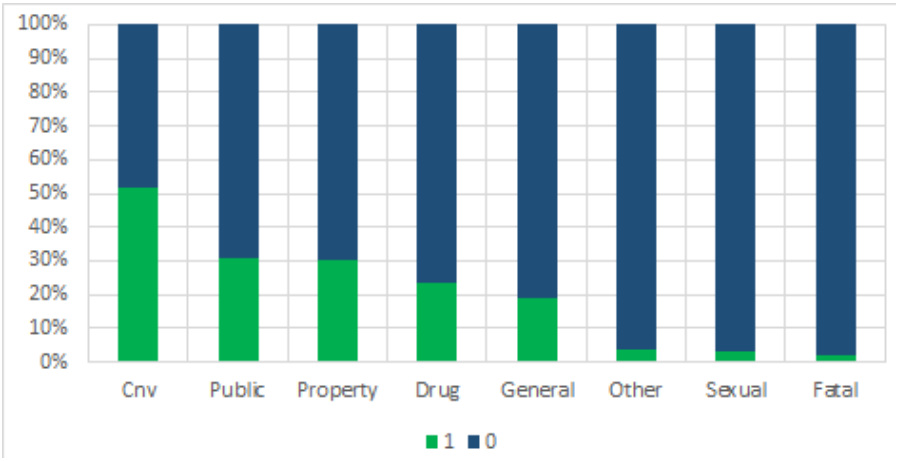


Figure 5.2: Percentage of Convicted(1) Not Convicted(0) Records for Different Crime Categories

decision on the reasoning that unless a person is adjudicated for a crime, an arrest is not sufficient to characterize guilt. An arrest without conviction could forebode the bias in the system and would need further investigation.

5.2.6 Dataset Limitations

Some of the variables included in this dataset and used in this study do not provide very detailed information. For example, EDUCAT and VOCAT do not encode the kind of courses they represent. Inmates' prior education level, marital status, employment history, family history etc. are not shared in the dataset. Additionally, the substance abuse characterization and corresponding treatment information is available for a small fraction of the population. Hence, it was impossible to compare the results using merely demographic information-based data and more personal activity-based information for the same dataset. Therefore, we resorted to comparing the bias in this study with the similar study from the previous chapter (see also [29]) that was based on the demographic information-based dataset to predict recidivism

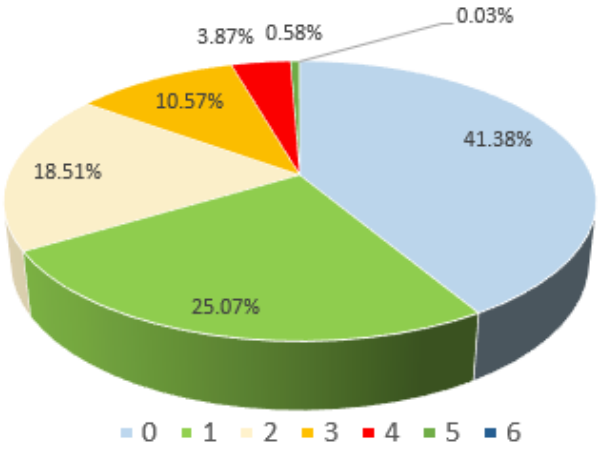


Figure 5.3: Percentage of Records with # of Crimes

while using reconviction as an output variable and set up similar experiments as in this study.

5.3 Methodology for Our Experiments

In this section we describe how we have structured our experiments and follow it by how we selected artificial neural networks for our base model and SRMs. We then define Singular Race Models and our neural network architecture along with the various hyper parameters. We follow this up with a subsection on how to assess bias in the results and another one on how race-based bias can be identified in the results.

5.3.1 Structuring of Our Experiments

In our study, we want to increase the accuracy of predicting recidivism and decrease race-based bias by grouping offenders based on their similarities. We want to pursue these offender similarities in terms of the types and frequency of crimes committed, the addictions faced, and the courses participated in. We believe that similar prior activities make offenders pursue similar crime trajectories after release on parole. Therefore, by providing input in terms of the crimes committed and

treatment courses participated in and completed until the time of the parole decision, we believe we can achieve higher accuracy in predicting recidivism and decrease bias in predictions. Therefore, we align our experiments by crime categories and employ the neural network to look for factors that align themselves for various crime categories and use Singular Race Models [29] to monitor race-based bias in each of the set of experiments.

For each of our experiments, we use several available and derived feature vectors: Vocational courses, Educational courses, HIV positive or not, substance abuse-related variables like drug abuser or not, took drug treatment, alcohol abuser or not, took alcohol treatment, sexual treatment, weight or number of state prisoners represented by each case, dead or undergoing life sentence, admission age for each cycle, admission age for the first arrest cycle, crimes adjudicated for in an arrest cycle (convictions, fatal, sexual, general, property, drug, public, and other category), sum of crimes adjudicated for in eight broad crime categories in previous arrest cycles, number of times confined/convicted/involved in domestic violence or fire arms in previous arrest cycles, and whether the record belonged to the 1994 arrest cycle or a later one (after94R). Furthermore, in order to reduce the bias arising from utilizing specific integer values for categorical data, we encode the categorical attribute values employing a one-hot vector.

We conduct eight sets of experiments, namely all-crime category and seven broad crime categories: all crimes, fatal, sexual, general, property, drug, public and other. The derived variables laterCNV, laterFatal, laterSexual, laterGeneral, laterProperty, laterDrug, laterPublic, and laterOther are the output variables for each of these sets of experiments. The output variables specify whether an offender will be convicted for any crime, fatal crime, sexual crime, general crime, property crime, drug crime, public crime, or other crime respectively in any of the subsequent

arrest cycles after the release in a given arrest cycle. In each of the eight sets of experiments, we trained three models and conducted five experiments. In each set, we trained models with data from all races, only Caucasians, and only African Americans. These models used all race or one race related data without ever using a race feature explicitly. In other words, the three models were not aware of the race of the offenders in the dataset. This gave us the leeway to prevent explicit race-based discrimination while still monitor the presence of race-based bias in the prediction results.

In each set of experiment, we tested the all-race model with the test data from all races, the African American race and the Caucasian race. We tested the SRMs trained on African American and Caucasian data respectively with the African American and Caucasian offender test dataset only. This enabled us to compare the prediction accuracy for all three models while also keeping track of the race-based bias in each result set.

We employed Caucasian and African American race models for SRM because they constituted 49% and 47% of our current Dataset 2 [60] respectively. Additionally, we compared our SRM models with those generated using Dataset 1 [58]. Caucasian and African American races constituted 43% and 56% of the latter dataset. The offenders from other races did not have sufficient records in both datasets to train and test different models.

Whenever possible, we compared the results from our study with those from the previous study in Chapter 4 that used neural networks to create SRM and used a dataset with primarily demographic information. As indicated previously, this previous study used almost exclusively demographic information-based features, including crime category (MAINCAT), post-release supervision (SUPER), marital status (MARITAL), pre-arrest employment status (EMPLOY), age of release (RE-

LAGE), age of admission (ADMAGE), time served in prison (TIME_SRV), number of convictions before current record (CHIST), education level before incarceration (EDUCLAM) features that are demographic in nature. In the work presented in this chapter, the input feature vector includes similar information as captured by MAIN-CAT, ADMAGE, TIME_SRV in the previous study. However, while CHIST in the previous study had access to only the number of convictions in previous cycles, in our study in this chapter we could calculate a much higher level of granularity in terms of history of different kinds of convictions in previous crime cycles from the raw dataset for the current study. This, together with the information regarding courses and treatments taken, gives this study significantly more access to individual, personal information for each of the offenders.

5.3.2 Selection of Artificial Neural Network to Generate SRMs

Our objective was to find a way to have higher accuracy and lower race-based bias in predicting recidivism in Dataset 2. Since FPR and FNR in conjunction can expose race-based bias, we tabulated accuracy, FPR, and FNR for several classifiers. We sought the highest possible accuracy with the lowest possible FPR and FNR values. In pursuit of this goal, we compared several machine learning models (see Table 5.1). In particular we evaluated Neural Networks, XGBoost, Linear SVC, AdaBoost, Nearest Centroid, and Decision Tree and chose neural networks based on its prediction accuracy for our derived input feature vectors from Dataset 2.

We divided our dataset of 142,573 records into training and testing subsets in using an 80-20 ratio. For neural networks we took a 10% validation split from the 80% training subset and used the validation to identify the best model. The training and test data had a recidivist to non-recidivist ratio of approximately 52 to 48. We used the same training and test sets for all 6 classifiers. We used the default parameters

Table 5.1: Experiment Results for Several Classification Models (Dataset 2)

All Crimes All Races	Artificial Neural Network	XGB	Linear SVC	AdaBoost	Nearest Centroid	Decision Tree
Accuracy	**0.676**	0.661	0.519	0.645	0.590	0.666
FPR	**0.291**	0.364	0.999	0.395	0.334	0.349
FNR	0.354	0.317	0.000	0.318	0.481	0.321

for these classifiers. For each of the Machine Learning algorithms, we used the same feature vectors with all crimes and all races, albeit the race feature vector was not explicitly included in the input vectors. We observed that the artificial neural network had the highest predictive accuracy amongst the six algorithms. Additionally, we found that FNR was reasonable with lowest FPR in the neural networks' result. Even though FNR in Linear SVC was minimal, the FPR was very high and the accuracy was very low. This result is similar to the results in Chapter 4 where we also found neural networks to perform better than several other classifiers when working with Dataset 1, a demographic information-based recidivism dataset.

As indicated previously, our initial choice of classifiers to test was informed by a recent study [18] based on recidivism which compared the results of logistic regression, random forests, support vector machines, XGBoost, neural networks, and Search algorithm for predicting recidivism. Even though, XGBoost and ANN had outdone all other classifiers used in that study, in our case we found that artificial neural networks had a higher accuracy amongst the two and a lower FPR between the results of the classifiers. Hence, we again chose artificial neural networks for our Singular Race Models.

5.3.3 Singular Race Models

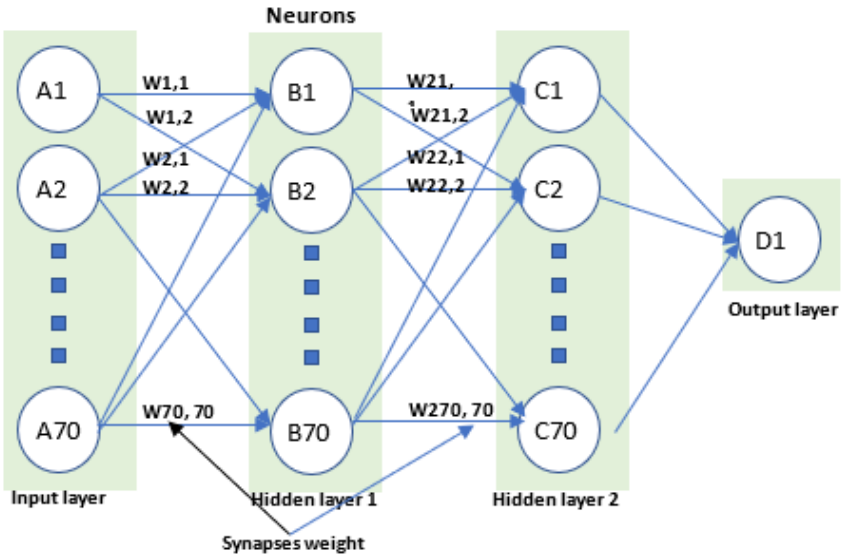
Singular Race Models (SRM) enable exposing a classifier to data of only one race at a time without ever using the race feature as part of the feature vector. SRMs allow models to gain information from data that pertains to only one race at a time. This empowers the models to gain accuracy from the racial homogeneity of the dataset without putting one race at a disadvantage as compared to another. The work in Chapter 4 found Singular Race Models to increase accuracy in most models generated with neural networks.

5.3.4 Neural Network Model

We used a three-layer neural network model to create the base model and SRMs for our experiments. Neural networks is an algorithm that seeks to identify feature relationships and related patterns embedded in the input feature set to each other and to the output. The algorithm takes inspiration from the human brain in identifying different pieces of information that influence each other and to use them to come up with an outcome. Just like the brain, neural networks solve the problem in a highly distributed fashion and adjust to the input features as their values vary. Neural networks learn these patterns via multiple neural nodes that may or may not be fully interconnected and alter each other's activation state.

We tuned several hyperparameters such as number of neurons, number of layers, batch size, activation, optimization to reach the highest accuracy in predicting recidivism when using all crimes data for offenders from all races from the [60] dataset. We reached the highest accuracy with a batch size of 256 and 80 epochs in a three-layered topology neural structure with 70 neurons in the input and in the hidden layer, and finally 1 neuron in the output layer. The network architecture used is shown in Figure 5.4.

Figure 5.4: Artificial Neural Networks Model for Base and Singular Race Models



We used an 80:20 ratio to divide the dataset into training and testing subsets. We took a 10% validation split from the 80% training data subset and used validation to find the best model.

Overfitting [68] occurs in Neural Networks when the algorithm captures noise embedded in the training data. In order to prevent overfitting, we used the dropout regularization technique with a dropout rate of 10%. This means that during each of the training cycles, the network drops 10% of the connections randomly from otherwise fully connected layers as the neural network uses back propagation to learn the weights related to each of the neural units [65]. We used Keras wrapper for TensorFlow [63], the Adam optimizer [64] instead of the classical stochastic gradient descent, logarithmic loss function binary_crossentropy, the ReLU activation function in the hidden layers, and a Sigmoid function in the output layer to implement our neural networks. We used this artificial neural network to implement and predict for our binary classification problem of predicting possible recidivists.

5.3.5 Assessing Bias in the Results

We recorded prediction accuracy using the artificial neural network. We assessed the bias embedded in the results by tabulating recidivism rate, nonrecidivism rate in the data, null accuracy, accuracy, FPR (False Positive Rate), and FNR (False Negative Rate) in the predicted results for every model generated in each set of experiments. We listed all six metrics for each set of experiments and compared these values for various SRMs with those for the base model. Additionally, whenever possible, we compared the values of these metrics in a set of experiment in this study with the one form Chapter 4. This permitted us to compare prediction results and bias encompassed between the personal activity based Dataset 2 and the demographics information-based Dataset 1 using the exact same metrics.

5.3.6 Race-Based Bias

Underpredicting recidivism for a privileged class and overpredicting recidivism for a disadvantaged class constitutes bias. Results showing bias in recidivism predictions have been presented in several studies like [25], [10], [29], [15], and [18].

To observe race-based bias, we assess and observe FPR and FNR in two race-based groups simultaneously: Higher FPR in conjunction with lower FNR portends a negative bias faced by a disadvantaged group (an example with strong bias for Dataset 1 can be seen in Table 5.3) where more non-recidivist members will be left incarcerated and fewer recidivists will be mistakenly released. Lower FPR in conjunction with higher FNR signifies a positive bias is meted out to a favored group that will have fewer non-recidivist members left incarcerated and proportionately more recidivists released. Studies like [29], [10] and [25], found High FPR and low FNR associated with the African American race and low FPR and high FNR associated with the Caucasian race.

When FPR in both race based groups had the same rate and at the same time the FNR in the two groups had approximately the same rate, this meant absence of bias in the results (an example of this can be seen in the FNR and FPR in the two SRMs in Tables 5.12 and 5.6 for Dataset 2).

When, as compared to the base models, FNR decreased in both sub-population's SRMs, it was good for society - as fewer recidivists would be released erroneously - and certainly a step away from the traditional race-based bias, but a lower FNR rate in one of the SRMs compared to the other indicated some persisting bias. When FPR in both SRMs increased simultaneously, this was also a break from the full blown traditional race-based bias. However, having a higher FPR in one cohort than the other represented some bias.

5.4 Experiments and Results

In this section, we describe eight sets of experiments based on crime types. First, we described each experiment, then tabulate, compare and explain the results. Whenever possible, we compare each experiments' results with those of the similar experiment with the Dataset 1 from Chapter 4.

5.4.1 Experiment 1: All Crimes

Increasing accuracy, transparency, interpretability, fairness, and decreasing bias in decision support systems for making predictions are some of the most pertinent goals for today's researchers. In the current work we worked to increase prediction accuracy and decrease race-based bias while predicting prisoner recidivism.

During the first round of selecting a classifier, we split the dataset using an 80-20 ratio for training-testing purposes. Since artificial neural network-based models had the highest accuracy, neural networks were used for creating SRMs. There-

after, we aligned our experiments by crime type. People with similar characteristics and activities will act similarly after release on parole - we hypothesized. So we labeled our records with 8 possible outcome variables - laterCNV, laterFATAL, laterSEXUAL, laterGENERAL, laterPROPERTY, laterDRUG, laterPUBLIC, and laterOTHER. These were binary labels that indicated whether a person would ever commit a certain kind of crime after release on parole. This meant that if a fatal crime offender were to commit a fatal crime after two arrest cycles, laterFATAL indicated one. If during subsequent release and arrest cycles, an offender committed general and drug crime, the record indicated laterCNV (indicated conviction for any kind in any of the following arrest cycles), laterGENERAL and laterDRUG to be one while laterFATAL, laterSEXUAL, laterPROPERTY, laterPUBLIC, and laterOTHER, were labeled 0, even when the offender had committed and was convicted of a fatal, sexual, property, or other crime in the current arrest cycle. We used SRMs for two reasons: to add the information of race without disadvantaging one race over another and to track and compare race-based bias with that in the base model.

In all crimes experiment, we trained three models - all races, Caucasian and African American races - with data from all types of crimes. The outcome variable for this category was laterCNV and indicated a 1 if the offender was convicted of any kind of crime in a subsequent arrest cycle. The input feature vector included all the input variables mentioned in Section 3.3 “Feature selection”. We converted the specific integer values for categorical data to a more homogeneous system using the one-hot vector technique.

5.4.1.1 Results of Experiment 1: All Crimes

In the all crimes experiments (shown in Table 5.2), the all races-based model achieved a prediction accuracy of 67.6%. The recidivism prediction accuracy was 68%

Table 5.2: Experiment results for All crimes (Dataset 2)

All Crimes Since 94	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.519	0.494	0.492	0.492	0.525
Non-recidivists	0.481	0.506	0.508	0.508	0.475
Null Accuracy	0.519	0.506	0.508	0.508	0.525
Accuracy	0.676	0.680	0.674	0.688	0.676
FPR	0.291	0.281	0.297	0.278	0.397
FNR	0.354	0.359	0.351	0.293	0.259

Table 5.3: Experiment Results for All Crimes (Dataset 1).

All Crimes	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.406	0.341	0.457	0.341	0.457
Non-recidivists	0.594	0.659	0.543	0.659	0.543
Null Accuracy	0.594	0.659	0.543	0.659	0.543
Accuracy	0.652	0.675	0.633	0.685	0.638
FPR	0.235	0.180	0.290	0.121	0.345
FNR	0.513	0.603	0.458	0.692	0.383

and 67.4% for Caucasian and African American races respectively. The accuracy for these two races improved to 68.8% and 67.6% respectively as we moved to the SRMs. However, more significantly, FNR for both Caucasian and African American models went to 29.3% and 25.9% respectively down from 35.4% FNR of the all race model. This meant that fewer recidivists would be released from both races and that both races were treated similarly in the all race model and in SRMs. FPR for African American SRM was higher as compared to that for the Caucasian SRM. This meant that in general - for all possible crimes- a model trained on African American race related data will falsely label more non-recidivists as recidivists if they belonged to the African American race. It was interesting to notice that the all race, all crime model

showed less bias for the two races than the two SRMs. Bias in the African American SRM was indicated by higher FPR and lower FNR as compared to lower FPR with higher FNR for the Caucasian SRM. In contrast, the all race based model treated the two races fairly similarly in the presence of the activity based input feature set.

5.4.1.2 Dataset 2 vs Dataset 1 Results

In contrast with Dataset 2, the demographical information based Dataset 1 showed a stronger disparity in the treatment of the races in both the base all race model and the SRMs in Chapter 4. In the study with Dataset 1 (results for which are repeated here in Table 5.3 for convenience), the FPR for all races was 23.5% in general but 18% for Caucasians and decreased further to 12% for Caucasian SRMs, while it was 29% for African American race in the all race model and increased further to 34.5% in the African American SRM. Similarly, FNR was 51.3% in the all race model, but 60.3% FNR for Caucasians in the all race model and increased further to 69.2% in the Caucasian SRM. FNR was 45.8% for African Americans in the all race model and decreased further to 38.3% in the African American SRM. This disparate treatment of races when using a demographic information-based input feature set constituted race-based bias introduced by the prediction algorithm. This was bias that showed up in the prediction results because offenders' criminal activity-based information was not available in the input set. This bias meant that more non-recidivistic African American offenders would be kept incarcerated and fewer recidivistic African American offenders would be released. It also meant that more recidivistic Caucasians would be released and fewer non-recidivistic Caucasian offenders would be mislabeled as recidivistic. The work in the current study using Dataset 2 [60] significantly reduces this biased treatment of the races. As we worked on the baseline case for this study and lumped all crimes and all races together (in the absence of a race feature

vector), the SRMs showed more bias in comparison with the treatment of races by a model trained on all races (Table 5.2), albeit to much less of an extent than a similar experiment with Dataset 1 Table 5.3).

5.4.2 Experiment 2: Fatal Crimes

For Fatal crimes, we did experiments just like Experiment 1 for all crimes - albeit this time for fatal crimes. If the offender committed a fatal crime in any of the subsequent arrest cycles, we labeled our records with laterFATAL = 1. If the offender never recommitted a fatal crime, despite being adjudicated for a fatal crime in the current arrest cycle, laterFATAL was set to 0.

5.4.2.1 Results of Experiment 2: Fatal Crimes

The data in this crime category was very imbalanced with 2% recidivist test cases with slightly higher percentage of African American offenders being non-recidivists than of the Caucasian offenders. As shown in Table 5.4, despite the imbalance the neural network models were able to increase the accuracy beyond null accuracy in the all race models and SRM. FPR stayed close to 0% in all cases. FNR was close to 15% for Caucasian SRM and 42% for African American SRM, which seems to indicate positive bias in favor of African Americans in the related SRM. However, one must pay close attention to the high prediction accuracy and realize that in the two models, the number of false positives were rather low with 53 FN out of a total of 13,582 records (281 TP) in the Caucasian model and 78 FN out of 13,870 total records (108 TP) in the African American model. This meant that by using offender activity-based input vectors, we could distinguish most recidivists from the others. Of course, as researchers, we should ultimately strive to reach 100% accuracy.

Table 5.4: Experiment Results for Fatal crimes (Dataset 2)

Fatal Crimes Since 94	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.020	0.026	0.014	0.025	0.013
Non-recidivists	0.980	0.974	0.986	0.975	0.987
Null Accuracy	0.980	0.974	0.986	0.975	0.987
Accuracy	0.994	0.996	0.994	0.995	0.994
FPR	0.001	0.000	0.001	0.001	0.000
FNR	0.248	0.148	0.419	0.159	0.419

Table 5.5: Experiment Results for Violent Crimes (Dataset 1).

Violent Crimes	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.314	0.232	0.381	0.232	0.381
Non-recidivists	0.686	0.768	0.619	0.768	0.619
Null Accuracy	0.686	0.768	0.619	0.768	0.619
Accuracy	0.702	0.758	0.656	0.769	0.654
FPR	0.113	0.073	0.154	0.031	0.167
FNR	0.700	0.802	0.652	0.894	0.637

5.4.2.2 Dataset 2 vs Dataset 1 Results

The experiments included in the violent crime category in Chapter 4 from the demographic information based Dataset 1 indicated lower accuracy for African American SRMs and bias that manifested itself in the form of higher FPR for Africans Americans subpopulation/SRM, and proportionately lower FNR for Caucasian sub population and SRM as shown in Table 5.5.

The Dataset 2 experiments for fatal crimes included here, in contrast, showed increasing accuracy in SRMs for the two races along with a lower FPR for African American races and higher FNR for Caucasians. Additionally, even when FNR for

African American race was higher in all race model, it did not increase any further in the SRM. This was going against the grain of bias in the previous results.

5.4.3 Experiment 3: Sexual Crimes

In the Sexual crime category too, our experiments were again set up like Experiment 1 for all crimes - though with sexual crimes this time. We assigned the laterSEXUAL outcome variable to be 1 in our records if the parolee committed a sexual crime in any of the subsequent cycles after release and 0 otherwise - even if a sexual crime was committed in the current arrest cycle.

As in previous experiments, three models were trained for offenders - for all races combined and for the two SRMs. The all race model was tested with all races, Caucasian, and African American data. The two SRMs were tested using their respective race related test data. All three models included input variables described in Section 3.3 "Feature selection" but did not include race information explicitly.

Table 5.6: Experiment Results for Sexual Crimes (Dataset 2)

Sexual Crimes Since 94	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.030	0.032	0.029	0.034	0.028
Non-recidivists	0.970	0.968	0.971	0.966	0.972
Null Accuracy	0.970	0.968	0.971	0.961	0.972
Accuracy	0.977	0.976	0.977	0.975	0.980
FPR	0.001	0.001	0.001	0.002	0.002
FNR	0.735	0.711	0.764	0.665	0.649

5.4.3.1 Results of Experiment 3: Sexual Crimes

The results of this experiment are shown in Table 5.6. There were 3% recidivists in the imbalanced sexual crime category of the prepared dataset derived from Dataset 2. Accuracy increased in African American SRMs. FPR went up minimally for both SRMs but FNR went down for both SRMs. This was important - particularly as the accuracy increased beyond the null accuracy in the highly imbalanced crime category. The simultaneous movement of FNR and FPR in the same direction and with nearly similar magnitude in the two SRMs meant strong reduction of race-based bias in the results. Lower FNR in both SRMs meant that fewer recidivists would be labeled as non-recidivists for both races. Having FPR and FNR move in tandem in the same direction and with similar magnitude meant the race-based bias had been lowered.

5.4.4 Experiment 4: General Crimes

For General crimes, we trained three models - one for all the races combined and one each for the two SRMs - Caucasian and African American. We tested the all-race model with all the races, Caucasian, and African American race data and the two SRMs with their respective race related test data. The models did not have explicit access to the race feature. We used the same input variables to the classifier as the ones specified in the input variables described in Section 3.3 “Feature selection”. The outcome variable laterGENERAL was set to 0 if the offender did not commit a ’general’ category crime after the current cycle and was set to 1 if they ever committed such a crime even with no prior history. The one-hot vector technique for categorical data to reduce bias related to their integer values.

Table 5.7: Experiment Results for General crimes (Dataset 2)

General Crimes Since 94	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.189	0.148	0.227	0.150	0.219
Non-recidivists	0.811	0.852	0.773	0.850	0.781
Null Accuracy	0.811	0.852	0.773	0.850	0.781
Accuracy	0.827	0.861	0.796	0.867	0.813
FPR	0.038	0.030	0.047	0.025	0.063
FNR	0.750	0.765	0.740	0.747	0.626

5.4.4.1 Results of Experiment 4: General Crimes

There were significantly (proportionally) more non-recidivists than recidivists (81% versus 19%) in this General crime category in the data derived from Dataset 2 and more Caucasian non-recidivists than African American non-recidivists (85% vs 78%) in the two SRMs as indicated in Table 5.7. The African American model learned more than the Caucasian one as one compares the accuracy and null accuracy of the two SRMs individually. The predictive accuracy of two SRMs were better than that of the mixed model for the two races (86.7% and 81.3% instead of 86.6% and 79.6% for Caucasian and African American races in the all race model). Additionally, FPR for the Caucasian SRM decreased (good change) and went up for the African American SRM (not good). FNR decreased for both Caucasian and African American SRM (good change). Typically, race-based bias in predicting recidivism is represented by a higher FPR and a lower FNR for the disadvantaged subpopulation and lower FPR with a higher FNR for the favored group as compared to the disadvantaged group. In the current results, not only did the accuracy go up, FNR went down for both SRMs. This meant that fewer recidivistic offenders of both races would

be erroneously released when SRMs are used in conjunction with offenders' activity-based input vectors.

5.4.5 Experiment 5: Property Crimes

We trained three models - all-races, Caucasian and African American SRMs. We tested the former with all the races, Caucasian, and African American race records and the two SRMs with Caucasian and African American race related test data. Race feature vector was not explicitly available to any of the models - merely the data from one race or another or from all races was available to the models to train and test on. There were proportionally more non-recidivists than recidivists in the Property crime category in the Dataset 2.

5.4.5.1 Results of Experiment 5: Property Crimes

The accuracy of predicting recidivism was 74.9% for the all-race base model in this category and improved for both races individually when we used SRMs as shown in Table 5.8.

In the baseline model, prediction accuracy decreased slightly for African American race when we trained using data for all races together and tested for individual races. However, both SRMs predicted property crime recidivism with higher accuracy as compared to how the mixed-race base model treated each of the races individually.

The hallmark sign of results incorporating race-based bias show higher FPR and lower FNR for the disadvantaged class while having relatively lower FPR and higher FNR for the favored class. The results in this experiment did not show race-based bias in the base model trained on all races as the individual races were treated fairly equally - indicated by similar FPRs and FNRs for the two races in the base models. However, in the two SRMs, as the accuracy went up, FPR went up slightly

Table 5.8: Experiment Results for Property Crimes (Dataset 2)

Property Crimes Since 94	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.308	0.277	0.340	0.273	0.326
Non-recidivists	0.692	0.723	0.660	0.727	0.674
Null Accuracy	0.692	0.723	0.660	0.727	0.674
Accuracy	0.764	0.637	0.623	0.786	0.767
FPR	0.095	0.228	0.201	0.110	0.123
FNR	0.553	0.714	0.718	0.491	0.458

Table 5.9: Experiment Results for Property Crimes (Dataset 1).

Property Crimes	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.472	0.420	0.538	0.420	0.538
Non-recidivists	0.528	0.580	0.462	0.580	0.462
Null Accuracy	0.528	0.580	0.538	0.580	0.538
Accuracy	0.612	0.613	0.612	0.618	0.618
FPR	0.345	0.311	0.399	0.273	0.518
FNR	0.436	0.493	0.379	0.534	0.266

more for African American SRM than for Caucasian SRM. Similarly, FNR went down for both SRMs(good) but slightly more for the African American SRM than for the Caucasian SRM. Thus, one can say that in the presence of activity based features, the base model trained on all races, treated all races similarly with reduced bias. The SRMs that trained the models on one race at a time, increased the accuracy but also added a small amount of bias in the result.

5.4.5.2 Dataset 2 vs Dataset 1 Results

Recidivists and non-recidivists were roughly equally distributed for property crime in Dataset 1 [58], the demographic information-based dataset used in Chap-

ter 4 for the property crime set of experiments. In those experiments, accuracy of the mixed model was almost the same for the two races individually and for the two related SRMs. However, the FPR indicated bias in results as it was lower for Caucasians than for African Americans in the mixed model (31.1% vs 39.9%) (see Table 5.9, repeated here for convenience). SRMs also further heightened this difference - decreasing the FPR for the Caucasian model further while increasing it for the African American model (to 27.3% and 51.8% respectively). FNR in the Dataset 1 property crime experiments emphasized bias in the results as it was higher for Caucasians and lower for African Americans (49.3% and 37.9% respectively) in the all-race model and increased further for Caucasian SRM while decreasing further for the African American SRM (53.4% and 26.6% respectively) In contrast to the results based on Dataset 1, the current work shows significantly less bias in all models. Here both FNR and FPR moved in the same directions in both SRMs (Table 5.8). Even as the accuracy increased for both races in the SRM, FNR decreased significantly in both SRMs as the FPR increased slightly.

5.4.6 Experiment 6: Drug Crimes

For drug crimes too we trained three models - one each for all races, Caucasians and African Americans. We then tested all race model with all race data, Caucasian data, and African American data. The two Singular Race Models trained with Caucasian and African American races were then tested using the same races. without access to a race input feature vector. Eighty percent of the dataset - mixed and of individual races - was used for training purposes while 20% was used for testing the models. There were proportionally more non-recidivists than recidivists 76% vs 24% overall in the drug crime category in the Dataset 2.

Table 5.10: Experiment Results for Drug Crimes (Dataset 2)

Drug Crimes Since 94	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.236	0.181	0.290	0.182	0.292
Non-recidivists	0.764	0.819	0.710	0.818	0.708
Null Accuracy	0.764	0.819	0.710	0.818	0.708
Accuracy	0.796	0.828	0.763	0.846	0.764
FPR	0.062	0.046	0.080	0.043	0.124
FNR	0.664	0.742	0.619	0.654	0.505

Table 5.11: Experiment Results for Drug Crimes (Dataset 1).

Drug Crimes	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.441	0.327	0.482	0.327	0.482
Non-recidivists	0.559	0.673	0.518	0.673	0.518
Null Accuracy	0.559	0.673	0.518	0.673	0.518
Accuracy	0.639	0.690	0.621	0.700	0.626
FPR	0.244	0.119	0.301	0.100	0.341
FNR	0.508	0.702	0.463	0.712	0.409

5.4.6.1 Results of Experiment 6: Drug Crimes

As compared to the base model trained on all-race data, the two SRMs in this experiment had better predictive accuracies than the individual accuracies for the races by the all-race model for drug related crimes in Dataset 2: 84.6% vs 82.8% for the Caucasians and 76.4% vs 76.3% for the African Americans (see Table 5.10). Another way to look at it is that the all-race model had lower predictive accuracy for both races than the coressponding SRMs.

Presence of higher FPR in conjunction with lower FNR for the disadvantaged group represents negative bias while lower FPR and higher FNR for the privileged

group represents positive bias towards the privileged group. In the drug crimes related experiments, the base model showed this difference in the two races. FNR went down for both SRMs (good), as one compares the FNR in the all-race base model meted out to the two races. This means that bias in the result went down. However, the FPR went down in the Caucasian SRM and up in the African SRM. This characterizes continued race-based bias - however much less in magnitude as compared to the one shown in the predictions from the demographic information-based Dataset 1. This meant that even in the presence of activity based input features, drug related crimes still put African Americans in a disadvantageous position compared to the Caucasian group.

5.4.6.2 Dataset 2 vs Dataset 1 Results

Recidivists and non-recidivists were roughly equally distributed for Drug crimes in Dataset 1. The predictive accuracies of the two SRMs were better than the individual accuracies of the races by the all-race model for drug crimes by all races: 70% vs 69% for the Caucasians and 62.6% vs 62.1% for the African Americans as shown in Table 5.11 (again replicated here from the experiments in Chapter 4 for convenience). FPR indicated bias in SRM results as it was lower for Caucasians than for African Americans in the mixed model (11.9% vs 30.1%). SRMs also further heightened this difference - decreasing the FPR for the Caucasian model further while increasing it for the African American model (to 10% and 34.1% respectively). FNR emphasized race-based bias in the results as it was higher for Caucasians and lower for African Americans (70.2% and 46.3% respectively) in the mixed model and increased further for the Caucasian SRM while decreasing further for the African SRM (71.2% and 40.9% respectively) In contrast to the results in Dataset 1, the drug related crimes in the current study based on the augmented feature space derived from Dataset 2

show a decrease in FNR in both SRMs. Similarly the disparity in the FPR is low in both the SRM and in how individual races are treated in the all-race model. Thus the drug crime based SRM models for prisoner activity based Dataset 2 show lower race-based bias than the ones made from demographical Dataset 1.

5.4.7 Experiment 7: Public Crimes

We conducted the experiments also for public related crimes. Here too, we trained models for all races, Caucasians and African Americans and tested the all race model with all race data, Caucasian data, and African American data; then continued to test the two SRMs with the related singular race data without the explicit usage of the race feature vector. Training and test data were in the 80-20 ratio while the overall recidivists to non-recidivists showed a 31-69 ratio.

Table 5.12: Experiment Results for Public Crimes (Dataset 2)

Public Crimes Since 94	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.311	0.302	0.319	0.298	0.306
Non-recidivists	0.689	0.698	0.681	0.702	0.694
Null Accuracy	0.689	0.698	0.681	0.702	0.694
Accuracy	0.746	0.750	0.744	0.762	0.760
FPR	0.075	0.078	0.071	0.120	0.120
FNR	0.650	0.649	0.651	0.515	0.514

5.4.7.1 Results of Experiment 7: Public Crimes

The two SRMs predicted recidivism with higher accuracy than the all-race base model and improved upon the accuracies for the individual races by the all-race

model for public crimes. Having disparate growth in FPR and FNR based on a group membership - race in this case - reveals bias in the prediction. However, in this study, as a result of more personal activity-based input feature vector, we found higher predictive accuracy (approximately 76% in the two SRMs) versus 74.6% in the all race model for public crime related experiments (Table 5.12). FNR decreased further for both SRMs and remained almost the same, thus treating both groups similarly - and therefore, not exhibiting increased racial bias.

5.4.7.2 Experiment 8: Other Crimes

"Other" was the last crime category of crimes defined in Dataset 2 that did not fit in any other category. Here too, we trained three models that were completely unaware of the races but were exposed to data from all races, Caucasians and African Americans races respectively. We tested the first model with all race data, Caucasian data, and African American data and the latter two SRMs with the same race data. As in all previous experiments, training and test data was split at an 80:20 ratio. Recidivists to non-recidivists were in a 4:96 ratio in the test dataset.

5.4.7.3 Results of Experiment 8: Other Crimes

In the current work, the null accuracy was very high for the race based SRMs (95.7% for Caucasians and 96.6% for African Americans) (Table 5.13). The accuracy in the two SRMs increased further to 96.5% and 97.7%. Bias is represented by proportionally higher FPR and a lower FNR for the underprivileged class while the privileged class enjoys lower FPR coupled with higher FNR. This was not the case in this set of experiments too - as the prediction accuracy increased beyond the high null accuracy and both FPR and FNR for both races decreased relative to their corresponding values in the all race model.

Table 5.13: Experiment Results for Other Crimes (Dataset 2)

Other Crimes Since 94	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.040	0.042	0.038	0.043	0.034
Non-recidivists	0.960	0.958	0.962	0.957	0.966
Null Accuracy	0.960	0.958	0.962	0.957	0.966
Accuracy	0.968	0.963	0.971	0.965	0.977
FPR	0.004	0.005	0.004	0.004	0.003
FNR	0.716	0.766	0.664	0.729	0.624

Table 5.14: Experiment Results for Other Crimes (Dataset 1).

Other Crimes	All Races Train/Test	All Races Train, Test Caucasian	All Races Train, Test African American	Train/Test Caucasians	Train/Test African American
Recidivists	0.415	0.348	0.478	0.348	0.478
Non-recidivists	0.585	0.652	0.522	0.652	0.522
Null Accuracy	0.585	0.652	0.522	0.652	0.522
Accuracy	0.631	0.654	0.610	0.654	0.613
FPR	0.263	0.189	0.350	0.157	0.409
FNR	0.518	0.640	0.425	0.714	0.363

5.4.7.4 Dataset 2 vs Dataset 1 Results

In Dataset 1 recidivists and non-recidivists were in roughly 42:58 ratio (Table 5.14). Accuracy of the two SRMs was approximately the same as that of the individual race in the mixed model - a little higher overall for the Caucasian race and a little lower for the African American race. However, both FPR and FNR indicated bias as FPR was higher for the African race and lower for the Caucasian race as compared to the values of the all races model. This bias was exacerbated in the SRMs as the FPR increased further for the African American model and decreased for the Caucasian SRMs with respect to the FNR values for individual races in the mixed model. FNR,

on the other hand indicated bias in the results by showing higher values for the Caucasian SRM and lower value for the African American SRM with respect to the FNR values for individual races in the mixed model. The current work (Table 5.13) does not show this increased bias in the results.

5.5 Interpretation and Discussion of the Results

Ubiquitousness of machine learning based decision support systems has made the need to examine the accuracy, transparency, interpretability, fairness, and bias of such decisions rather urgent. In the current work, we focused on increasing accuracy and reducing bias.

The dataset used in this work had up to 99 arrest cycles associated with each 1994 arrest cycle of each offender tracked in the dataset. Each arrest cycle stored up to three most serious offenses committed during each arrest cycle. This allowed us to split each record into a maximum of 100 arrest cycles. Having 100 arrest cycles with at most three most serious crimes recorded in each of them gave us the ability to calculate a sum of crimes committed in each of the seven broad crime categories in any of the the previous arrest cycles. Even though the 1994 arrest cycles had several substance abuse related features, very few offenders had these recorded for them. By using all arrest cycles since 1994 (Figure 5.1) and removing the prisoner id, date of birth etc from each arrest cycle record, we could remove correlating features of the records and gain many records which did have several unique substance abuse related features with different numbers of crimes committed in the previous cycle at different ages. Fortified with a set of offender criminal activity and substance abuse-based input feature vectors, we used SRM (Singular Race Models) for two purposes: to increase accuracy without pitting one race against another, and to monitor race-related bias in the results. The eight sets of experiments tracked all crime types

together and seven types of broad crime categories committed. The base models in each set of experiments used the entire data set and hence used data from all races.

A high FPR coupled with low FNR for a disadvantaged class, while concurrently having low FPR and high FNR for the privileged class are the typical traits of biased decisions and have been observed in many other studies [25], [15], [18] [10], and [29] hereto. Singular Race Models used in the current study allowed the classifiers to gain additional information and increase accuracy in all sets of experiments. Having access to activity-based information like number of similar crimes committed in all previous cycles and substance abuse information along with treatments started or completed reduced the race-based bias in all but the all-crimes recidivism prediction results. These derived activity based input features helped the neural network learn more about the offenders and reduce bias learnt from solely demographic information-based features. In the all-crime category, the all-race model with all crimes treated both races similarly with very similar FPR, FNR, and accuracy. The SRMs in this category barely increased the accuracy while increasing bias - which meant lower FPR and higher FNR for Caucasian SRM while having higher FPR and lower FNR for African American SRM.

In the seven broad crime related categories, as the neural network was exposed to the crime specific outcome variable and required to predict whether an offender would commit a specific crime any time after release, divergence of FNR and FPR in the all race models was much lower for all races as compared to that shown in similar crime categories in the demographic information based Dataset 1. Results in most experiments improved the accuracy for both races while significantly reducing race-based bias. For example, in Fatal crimes, the accuracy improved beyond the null accuracy to 99.4% or beyond and the FPR stayed close to 1% for both races; FNR stayed at 41.9% for the African American cohort. Even though this seemed large,

the accuracy of 99.5% for the African American SRM meant that this represented a small number of offenders.

Race-based bias in predictions in the current work was reduced in general as compared to that in the similar experiments in Chapter 4 using Dataset 1. In the current work, FPR and FNR often had similar values for both races in the all-race model, while the values for FPR and FNR moved in tandem in similar directions for both SRMs. For example, in the sexual crime category, where data was skewed in favor of non-recidivism, the accuracy of predicting recidivism improved further to 98% for the African American SRM. FPR increased minimally for both SRMs and FNRs decreased significantly for both SRMs. In the general crime category, FNR decreased in both SRMs as the accuracy increased in both SRMs. In property crimes, public crimes, and other crimes category where all-race models had similar FPR and FNR for both races, the accuracy increased in both SRMs while the FPR increased slightly for both SRMs. Additionally, FNR decreased much more for both SRMs. In drug crimes, and general crimes even though FPR for African American SRM increased, FNR decreased for both Caucasian and African American models. In the fatal crimes category, the accuracy, FNR, and FPR stayed the same across all categories.

In the current work, accuracy, FPR, and FNR values were similar for both races in most all-race models. This meant that the all-races-based model with activity-based input variables was able to learn from offenders' data irrespective of the race. So, in almost all of the experiments, SRMs added accuracy without dramatically adding race-based bias. The fact that all-race based base models treated both races similarly is very significant. It has the potential to be the harbinger of the era where the true minorities have a shot at getting a fair treatment even in the absence of same-race offenders prior to their fate being decided by a decision support system. Since the race-based bias decreased in the presence of offender activity-based information,

we believe that it portends a fairer system than today. In this paper, we had grouped 26 different crime categories into 7 different categories, so it seems that maintaining all 26 crime categories' information in the input variables could potentially reduce bias further while increasing accuracy.

Like in the previous study on Dataset 1, the Caucasian based SRMs showed the highest accuracy in most experiments like those of public crimes, drug crimes, property crimes, general crimes, Fatal crimes, and All crimes. In other crimes and Sexual crimes though, the African American SRMs showed the highest accuracy (Figures 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11, 5.12 reprise the results in more graphical form by showing bar graphs of the accuracy, FPR, and FNR for each of the experiments).

This dataset enabled us to use offender activity information in each arrest cycle directly and also derive additional variables like the total number of specific types of crimes before each arrest cycle. This helped neural networks learn activity-based patterns and to move away from race-based patterns. Splitting records by arrest cycles also helped neural networks learn from each cycle and improve accuracy while reducing bias.

Figure 5.5: Results for All Crimes Experiment (Dataset 2).

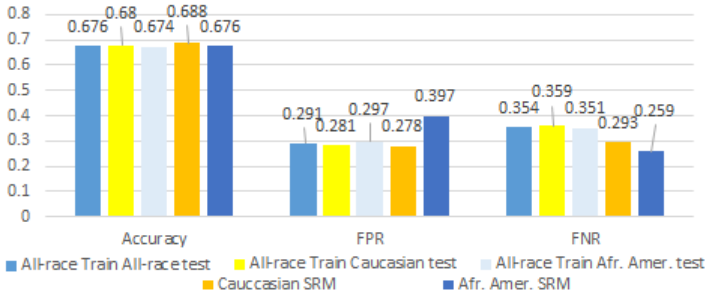


Figure 5.6: Results for Fatal Crimes Experiment (Dataset 2).

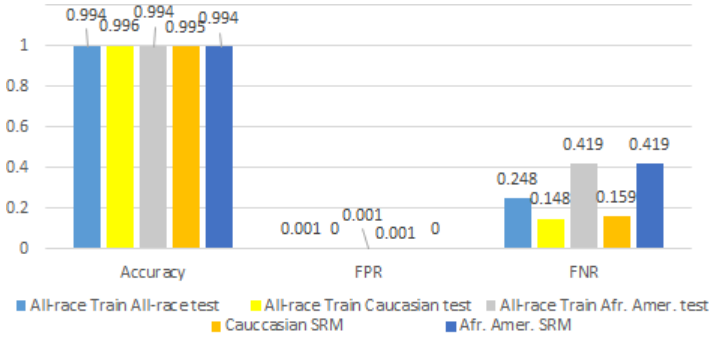


Figure 5.7: Results for Sexual Crimes Experiment (Dataset 2).

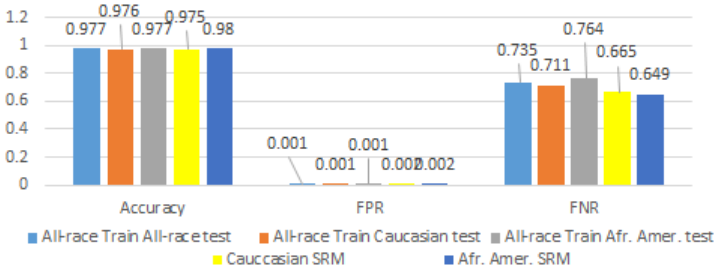


Figure 5.8: Results for General Crimes Experiment (Dataset 2).

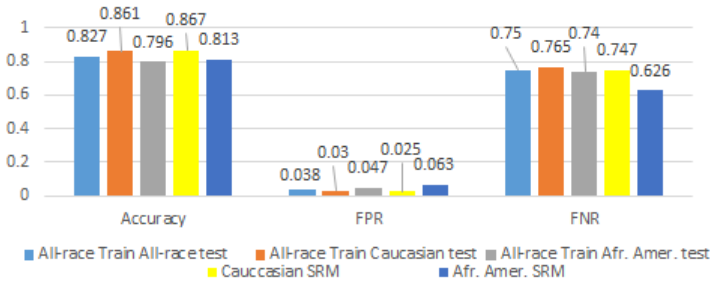


Figure 5.9: Results for Property Crimes Experiment (Dataset 2).

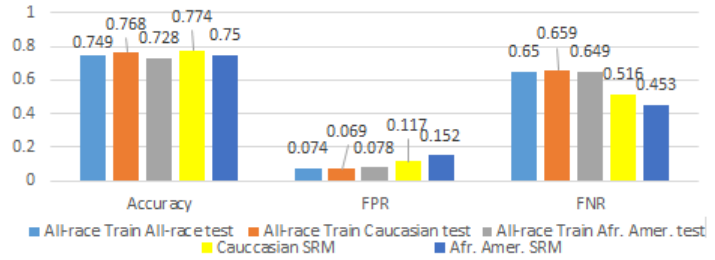


Figure 5.10: Results for Drug Crimes experiment (Dataset 2).

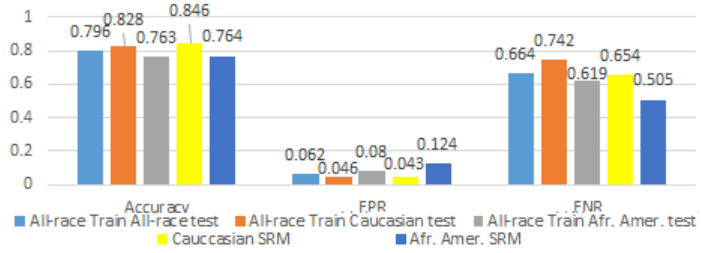


Figure 5.11: Results for Public Crimes Experiment (Dataset 2).

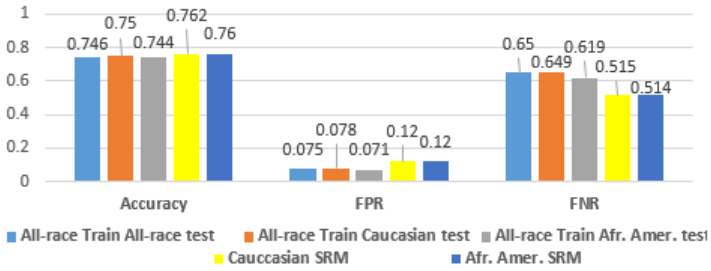
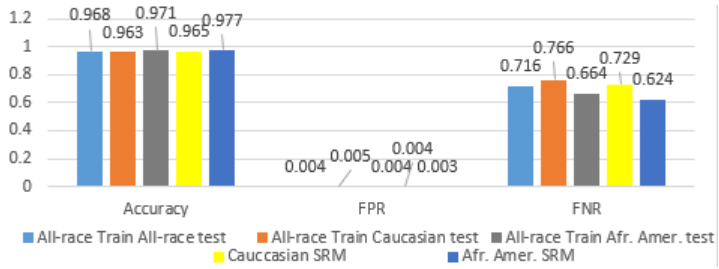


Figure 5.12: Results for Other Crimes Experiment (Dataset 2).



5.6 Conclusion

In this work we focused on increasing accuracy and reducing race-based bias in predicting prisoner's recidivism. To achieve this, we studied the effect of adding several criminal activity and substance-abuse based variables to the input feature set. Splitting each offender's record by arrest cycles permitted us to learn from many more arrest cycles and predict for many more cycles than merely the 1994 release cycles. As a result of these split records, we could not only use the criminal activity-based in-

formation from these records but could also derive some additional ones - for example age at each arrest, number of crimes committed in each of the broad crime categories before each cycle etc. Having many records reduced neural networks from overfitting, which lead to a higher accuracy on the test set. Having access to several records of post 1994 nature, provided us with a way to include substance abuse variables in several post 1994 records. Criminal activity and substance abuse-based information provided neural networks with other means beyond demographical information - to search for similarities amongst offenders who committed similar crimes.

Furthermore, using SRMs to train and test one race at a time was particularly advantageous in using race-based information without race-based discrimination by increasing the predictive accuracy and in observing the reduced bias in the results. Our experiments showed that in the presence of the activity based information, the base models treated both races similarly. SRMs increased the accuracy and often reduced the bias - though not always. This is a huge step in the direction of fairness for the two dominant races. This also leads us to our future work - namely how does the personal activity-based input feature set treat minorities.

CHAPTER 6

Reducing Race-Based Bias and Increasing Recidivism Prediction Accuracy by using Past Criminal History Details

A recent survey paper on bias and fairness [43] found two potential sources of unfairness in prediction results: those arising from the bias in the data and those arising from the algorithms. Our previous work in Chapter 4 found bias in prediction results and observed that the dataset was of a demographical nature and that investigation into a personal information-based dataset will be a good next step to pursue. As a result, we started to work with Dataset 2 which contains more personal information in Chapter 5 and showed that using such data could not only increase accuracy but, more importantly, could also reduce racial bias in the prediction results. In this chapter, we extend this by more carefully analyzing the effects of preprocessing the data set and in particular the effects of including historic information related to previous crime histories as well as to treatment and education activities pursued during incarceration. For this, we preprocess the data further to create and use, among other things, a prisoner's criminal activity, time between readmission, substance abuse, vocational/educational courses, and past criminal history as personal features for prediction purposes.

The first objective in this work is to study the effect of an offender's activity-related personal data and the influence of prior arrest history on the accuracy of predicting future recidivism and on the bias embedded in the prediction results.

The second objective in this work is to lay out steps to use our approach to select a prediction model using False Positive Rate Parity to obtain the one that

achieves high accuracy and least bias amongst the generated models. We compare these models with each other and with the base model produced without considering any prior arrest history.

The third objective in this work is to use our approach to predict recidivism for the All crimes and Sexual crime category.

The fourth objective in this work is to analyze the results produced in this research and compare them to results from other research to assess the benefits of the used data augmentation technique.

6.1 Dataset

For this work, we again used Dataset 2, a dataset from "Recidivism of Prisoners Released in 1994" study [60]. It is described in more detail in Section 3.2

6.2 Methodology

In this work, we use a unique approach to reduce bias while increasing recidivism prediction accuracy. The previous work in Chapter 4 (and published in [29]), as well as the work by Ozkan [18] that worked with numerous statistical models to improve recidivism prediction accuracy found neural networks to deliver better results amongst the classifiers used. Therefore, in this work, as with the work in Chapter 5, we have again used neural networks to build our models. We performed seven sets of experiments each for all crimes and for the sexual crime categories by varying the used prior crime history to the previous 0, 10, 20, 40, 80, or 100 arrest cycles. The goal here was to study the impact of including prior history in more detail and to evaluate its impact on prediction accuracy and prediction bias. The overview of our experiments' methodological framework is shown in Figure 6.1.

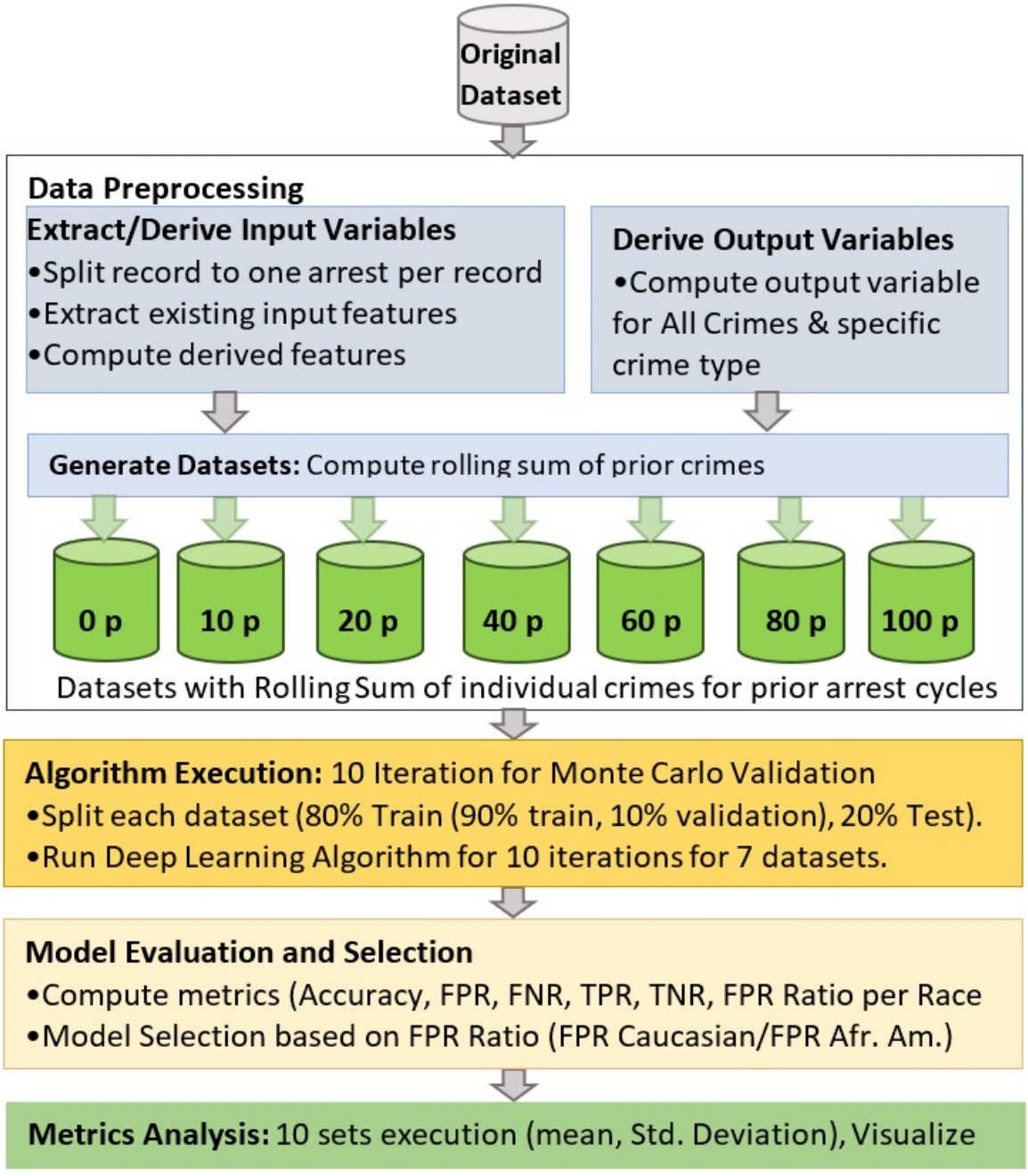


Figure 6.1: Overview of the Methodological Framework. 'p' stands for prior arrest cycles in the datasets.

We divide our approach into 4 steps: Data preprocessing, Algorithm execution (neural networks), Model evaluation/selection, and Visualization/Metrics Analysis. The details of the four steps in our approach are as follows:

Step 1 is **Data Preprocessing**. This is comprised of Input Feature Selection/Derivation, Output Variable derivation (crime based), and Datasets generation. Step 1 includes the following:

- Use the raw dataset.
- Split each offender's record. Since each original record is composed of the 1994 arrest cycle and several other arrest cycles before and after the 1994 cycle for a given offender, split these to generate one arrest cycle per record. This increases the number of records from approximately 38 thousand to 442 thousand, each with a unique history.
- Add 26 individual crime detail features to each record.
- Generate 7 baseline and experiment datasets with different crime history considerations using a rolling sum of 26 individual crimes in the previous 0,10, 20, 40, 60, 80, and 100 prior arrest cycles for a given individual's records.
- Prevent overfitting by removing details such as date of birth and rounding features like age at arrest and age at first arrest.
- Convert categorical features such as VOCAT (vocational courses related features) into one hot vectors to prevent integer values from biasing the prediction.
- Compute crime-based output variables laterCNV and laterSEXUAL to represent binary classification of whether an offender committed any crime or a sexual crime, respectively, in any of the cycles succeeding a given one.

Step 2 is **Algorithm Execution**. This is comprised of data splitting into train-

ing, test, and validation datasets, and Monte Carlo Validation. Step 2 includes the following:

- Split each dataset into training (80%), test (20%), and validation (10% of training) datasets.
- Tune hyperparameters to use neural networks with a batch size of 256 and 80 epochs.
- Run 10 iterations for Monte Carlo Cross-Validation for each of these 7 datasets - once for predicting re-conviction for any crime in a later cycle (laterCNV output variable) and once for predicting re-conviction for a sexual crime in a later cycle (laterSEXUAL output variable). This yields a total of 140 sets of experiments (10x7x2).

Step 3 is **Model Evaluation and Selection**. This is comprised of metrics generation and model evaluation based on given metrics. Step 3 includes the following:

- Generate Accuracy, FPR, FNR, TPR, TNR, and Ratio of FPR African American and FPR Caucasian for the test datasets. See Chapter 2 for definitions.
- Select a model based on False Positive Rate Parity. This means that we select the model for which the ratio of FPR African American and FPR Caucasian is closest to 1. This model treats the two races most similarly. See Chapter 2 for definitions of various bias metrics.

Step 4 is **Visualization, Metrics Analysis**. Step 4 includes the following:

- Compute the average for Accuracy, FPR, FNR, TPR, TNR, and FPR ratio from the 10 Monte Carlo Validation iterations in the previous step.

- Visualize the metrics generated from the 10 iterations using each of the 7 datasets with rolling sum of individual crimes for the corresponding number of prior arrests.

6.2.1 Notation

We use the following notation to formulate the problem of predicting recidivism with a single sensitive attribute of race. We need to find the model that helps predict with high accuracy and lowest bias possible amongst the given models each one of which is developed with the rolling sum of prior individual crimes from N prior arrest records. In this chapter, we will use the following notation to characterize aspects of the datasets:

$X \in R^d$: quantified features of the dataset.

$A \in \{0, 1\}$: Race-based binary sensitive attribute (African American,Caucasian)

$C = c(X, A) \in \{0, 1\}$:prediction of the binary target variable (not reconvicted/reconvicted)

$Y \in \{0, 1\}$: target variable from the data set (will not reoffend/will reoffend).

Assumption: (X,A,Y) are generated from a distribution D denoted

as $(X, A, Y) \sim D$.

6.2.2 Fairness-Based Model Selection using FPR Ratio

We use the Fairness through Unawareness technique [69] by not including the sensitive attribute of race to minimize disparate treatment and use False Positive Rate parity to select the model that minimizes disparate impact for different sub-populations.

This assumes that the prediction results are the same with or without A, the sensitive attribute of race.

C:= c(X,A) = c(X)

We seek to prevent disparate treatment of races by removing the sensitive attribute from the input features. We measure the disparate impact by testing our models in two different ways - testing with all the test data and testing with only test data pertaining to one race at a time without the race information.

False Positive Rate Parity: In the recidivism domain, this means that the rate at which a non-recidivist member of either subpopulation may be mislabeled as a recidivist should be the same. This definition embraces the fact that the two subgroups may have different rates of recidivism. So FPR parity allows us to use the race-based or group fairness [11] criteria of having a similar False Positive Rate for the two sub groups while prioritizing fairness to individuals. When a decision erroneously labels an individual to be a recidivist, the decision is unfair for the offender. Thus, by minimizing the FPR and working to converge the FPR ratio of the two races, we seek to minimize bias meted out to an individual in a disadvantaged class in the results.

In this work we want to focus not only on the recidivism prediction accuracy but also on minimizing the prediction misclassification cost for a subpopulation. When we compare the FPR of the two subpopulations, the one with higher FPR is having its members mislabeled as recidivists at a higher rate. This means the subpopulation with a higher FPR will be kept incarcerated at a higher rate than the privileged group. Therefore, we compare the FPR for the two subpopulation by using the higher FPR as the numerator and the lower FPR as the denominator for each of the models and select a model that has False Positive Rate parity for the two subpopulations. We assume that A=0 is the disadvantaged class as members of a disadvantaged class will

have a higher rate of labeling non-recidivists as recidivists. A = 1 represents the privileged class.

To achieve perfect error rate parity, the following must be true for the two sub populations:

$$P_0[C = c|Y \neq c] = P_1[C = c|Y \neq c] \qquad \forall c$$

To achieve False Positive Rate ϵ parity, the following must be true:

$$\frac{P_0[C = 1|Y = 0, A = 0]}{P_1[C = 1|Y = 0, A = 1]} \leq 1 + \epsilon, where\ \epsilon,\ is\ p/100$$

p is the maximum percentage by which the ratio of labeling mistakes among the two groups can deviate in order for of labeling mistakes the system to be willing to accept the result while still calling it fair. This is the ratio of FPR of the disadvantaged class and FPR of the favored class.

6.3 Experiments

In this work, we increased the accuracy of predicting recidivism and decreased race-based bias by using offenders' activities and crime-based personal history. In order to empower neural networks to find crime-based similarities amongst offenders and ignore their membership in a race-based group, we excluded race information and included several available and derived input variables described later in this section.

We conducted experiments for two different types of crimes - one all encompassing "All Crimes" category and another looking at a specific "Sexual Crimes" category. We chose the All crimes category for prediction purposes as this was used in another work [18] on the same dataset [60]. We chose the sexual crime category because it was a violent crime that had almost similar recidivism rates for the two subpopulations – 29% and 26% recidivism rates for African Americans' and Caucasians' records, respectively. The All crimes category in contrast had a similar recidivism rate of almost

84% for both Caucasian and African American records in our generated datasets with rolling sums of 26 individual crimes for different numbers of prior arrest cycles (0, 10, 20, 40, 60, 80, and 100) .

For each of the two crime categories, we set up our experiments to use 80% of the records from each of the seven datasets to train and validate (10% of training dataset for validation) the model, and 20% for the test dataset to predict whether the offender would commit "All Crime" or a "Sexual Crime" in any of the following cycles. Two derived binary variables were used to indicate the future outcome for a given offender record.

For our experiments, we trained the model using all the races without ever using the race information. We tested the model performance using test data of all races but lacking race information. In order to asses how the model treated each of the race, we extracted Caucasian data and African American data from the test data and used it for the model after removing the race-based information from each subset.

6.3.1 Model Evaluation

We evaluated our model's performance by using ten independent iterations of Monte Carlo cross-validation [70]. We randomly selected an 80-20 ratio to split the data into training and test sets, while 10% of the training data was used for validating the model by Keras' deep learning model. We repeated this process 10 times for Monte-Carlo cross-validation. We partitioned the data (80-20 split) independently for each run. After tabulating all values for the ten iterations, we computed the average values for Accuracy, FPR, FNR, TPR, TNR, and the ratio of FPR African American to FPR Caucasian for all model evaluation and bias metrics. Even though we use Accuracy and FPR parity in our approach, we calculated all the mentioned metrics as researchers often use some of these to evaluate and compare the performance of

their models with those of others. Their graphical representation is presented in this work. See Figure 6.3 and Figure 6.4.

6.3.2 Input Variables

For input variables, we used several features that were explicitly present and created several derived ones. We chose the input features that seemed subjectively to be the most relevant ones. We excluded the correlated features as they did not increase the accuracy of prediction and included the most informative ones that added information about offenders' activities during the prison time and criminal history.

Available input variables: We used vocational courses (VOCAT), Educational courses (EDUCAT), HIV positive or not (HIV), substance abuse-related variables (DRUGAB, DRUGTRT, ALCABUS, ALCTRT), behavior modification treatment participation variables (SEXTRT), number of state prisoners represented by a convict (WEIGHT), dead or undergoing life sentence (DeadOrLifeCnf), Admission Age for each cycle (AgeC), Admission Age for the first arrest cycle (AdAgeC1), released prisoner's 1994 release offense (SMPOFF26), felony or misdemeanor (J00NFM), conviction for adjudication offense (J00NCNV), confinement for adjudication offense (J00NCNF), confinement length for the most serious adjudication charge (J00NPMX) and crimes adjudicated for (or not) in an arrest cycle (convictions, fatal, sexual, general, property, drug, public, and other category)

Derived input variables: We included several derived variables based on an offender's current activities and past criminal history as these affect what they do after release on parole. Furthermore, individuals tend to repeat the kind of crimes they committed earlier and often do not switch to crime categories very different from what they have committed earlier. We included 26 broad crime categories related binary variable to indicate the crime types committed during a cycle in every record. We also

included sums of each of the specific 26 crime categories committed in the previous N cycles (cumCR_01, cumCR_02, ... cumCR_26) followed by normalization of these variables. We then encoded the values of the included categorical attributes to binary variables. We included other derived variables such as the number of years between admission of past and current arrest cycles (Years_To_LastCyc), sum of crimes adjudicated for in the N previous arrest cycles (CUMcnv , CUMfatal, CUMsexual, CUMgeneral, CUMproperty, CUMdrug, CUMpublic, and CUMother categories) crime count in the N previous arrest cycles of confinement(CUMJ001CNF), involvement in domestic violence (CUMJ001DMV), conviction (CUMJ001CNV), involvement with Fire Arms (CUMJ001FIR) and whether the arrest record was from the 1994 arrest cycle or a later cycle (after94R).

6.3.3 Output Variables

In this work, we formulate two recidivism prediction problems by generating two output variables: laterCNV and laterSEXUAL. These are binary variables and indicate whether an offender is reconvicted or adjudicated for any crime or sexual crimes in any of the ensuant arrest cycles after the arrest cycle for which recidivism is predicted.

6.3.4 Bias Metrics for Experiment Results

We computed five metrics for each of our models: Accuracy, FPR (False Positive Rate), FNR (False Negative Rate), TPR (True Positive Rate), and TNR(True Negative Rate). We evaluated bias by comparing FPR and FNR for the two major subpopulations. We computed FPR2/FPR1 for both subpopulations and marked it for various models. See second graphs in Figure 6.3 and Figure 6.4. We selected a prediction model based on the least FPR ratio. The remaining metrics enable us to

compare our model performance with those of models developed differently than in this work.

6.3.5 Artificial Neural Networks

In this set of experiments, we employed a three-layer Neural Network Model. A Neural Network is an algorithm that detects the pattern buried in the input feature set of a given dataset and relates it to the output. The algorithm is not very different from how the human brain recognizes different pieces of input features that impact each other and relates them to an outcome. Since the neural networks can learn these patterns even as they change, we do not need to reprogram them.

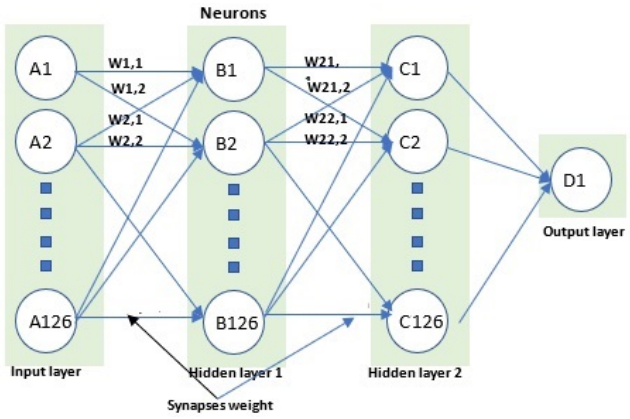
By learning the patterns embedded in the input by using several nodes that may not be fully connected, the network is able to influence the activation state of the other nodes [29].

As we tuned the possible hyperparameters such as number of neurons, number of layers, batch size, activation function etc., to achieve the highest accuracy with the lowest race-based bias possible, a three-layered neural network with 256 batch size and 80 epochs with 126 neurons in the input and in each of the hidden layers and 1 neuron in the output layer were found to be ideal. See Figure 6.2.

6.4 Results and Discussion

Our results presented in Figures 6.3 for the All cries category and in Figure 6.4 for the Sexual crimes category, showed that with the increase in the number of prior arrest cycles used to calculate the cumulative sum of individual crimes in each arrest cycle, the accuracy increased while the FPR and FNR decreased - though not always equally for each rolling sum of crime datasets. The Caucasian subpopulation in the test dataset showed a lower FPR and a higher FNR rate than the corresponding rates

Figure 6.2: Artificial Neural Networks with Neurons, synapses and Weighted Sum for Each Layer Used to Create Rolling Sum of Individual Crimes for Prior Arrest Cycles' Models.



for the African American subpopulation – a hall mark of bias in results - to some degree in all models - though much less than that in [18]. A more detailed comparison with their results will be presented in Section 6.5. In our dataset, a higher percentage of Caucasians had fewer than 10 records than African Americans. Most convicts had under 40 arrest cycles. At the 40 prior record level for "All Crimes", 91% of all African American records and 94% of all Caucasian records are taken into consideration. 61% of African American records and 66% of Caucasian records were in the 20 or fewer records category.

We found that our models achieved the highest accuracy of 89.8% and 90.4% for African Americans and Caucasians, respectively, with 40 prior dataset models. However, at a minimal loss of accuracy, the least biased model was achieved using the 20 prior dataset. Here an FPR African American/FPR Caucasian ratio of 1.21 was achieved. The FPR ratio of 1.32 for the base model with 0 priors was considerably improved to 1.21 by using the 20 priors dataset. Considering more arrest cycles than 20, deteriorated (increased) this ratio for the "All Crimes" category and hence increased the bias in the result.

Figure 6.3: Monte Carlo cross validation Average Accuracy and Bias Metric Values for All Crimes models with rolling sum of individual prior crimes for 0, 10, 20, 40, 60, 80, and 100 prior arrest cycles.

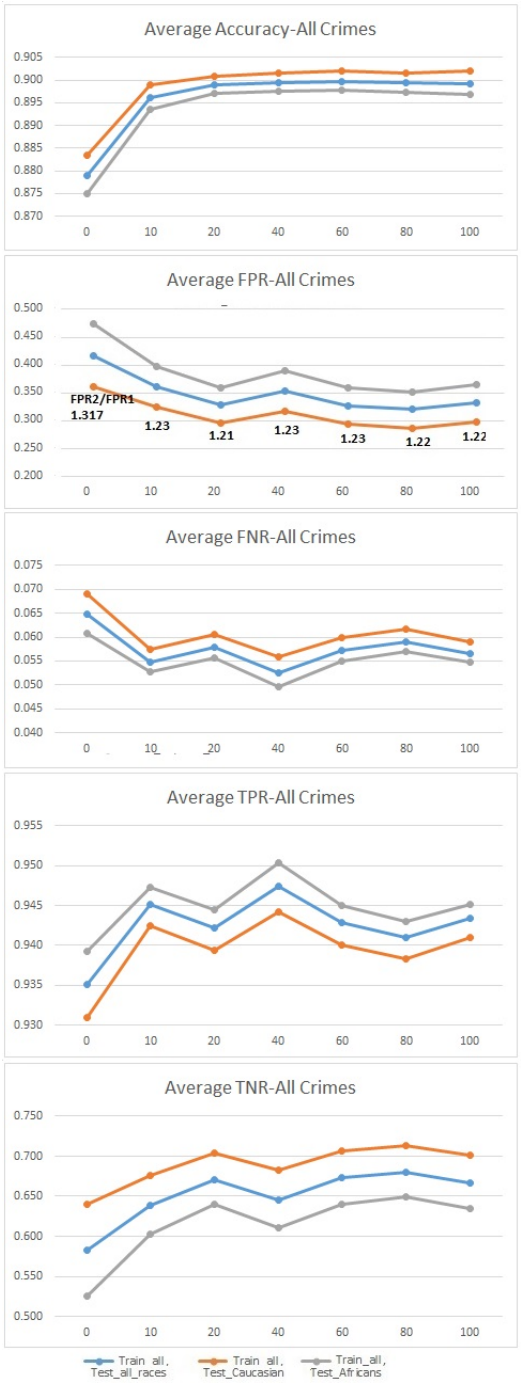


Figure 6.4: Monte Carlo Cross-Validation Average Accuracy and Bias Metric Values for Sexual Crimes models with Rolling Sum of Individual Prior Crimes for 0, 10, 20, 40, 60, 80, and 100 Prior Arrest Cycles.

Similarly, for the "Sexual Crimes" category, an accuracy of 86% was possible using any of the 20 or more prior arrest datasets for both race-based cohorts but the least FPR ratio of 1.07 spurs selection of the 40 priors dataset model.

Thus our experiment results for "All Crimes" indicated that selecting models generated with 20 prior arrests provided us with an average accuracy of 89.7% and 90.1%; FPR of 36% and 30%; FNR of 5.6% and 6.1%; TPR of 94.4% and 93.9%; and TNR of 64% and 70.3% for African Americans and Caucasians, respectively.

Similarly, our experiment results for "Sexual Crimes" indicated that selecting models generated with 40 or 60 prior arrests provided us with an average accuracy of 86%; FPR of almost 8% for both groups; FNR of almost 8%; TPR of 72% and 67% for African American and Caucasian cohorts; and TNR of almost 90% for both groups.

Selecting the model with the lowest FPR ratio allows us to minimize the bias meted out to the disadvantaged group - thus following a group-based criterion. By focusing on FPR we focus on the individual fairness as a lower FPR means fewer non-recidivists mislabeled as recidivists and languishing needlessly behind bars.

6.5 Performance Comparison with Other Works

An earlier work by Ozkan [18] with the same dataset employed six classifiers - namely, Logistic Regression, Random Forests, XGBoost, Support Vector Machines, Neural Networks, and Search algorithms. They used 1994 records to train and test their models. They achieved the best Accuracy /FPR /FNR /TPR /TNR of 0.778 /0.406 /0.054 /0.946 /0.594 using XGBoost /logistic Regression /SVM /SVM /Logistic Regression, respectively - while overall XGBoost and Neural Networks worked the best for them. We, in contrast, used all arrest records for training and testing purposes and could achieve 0.899/0.329/0.058/0.942/0.671 and much closer FPR parity

for the two races from the same model using the dataset with a rolling sum of 20 prior arrest records for All Crimes. We selected this model because it had the best FPR ratio for the two race-based cohorts amongst all models. See Figure 6.3. In other words, we could achieve a higher accuracy, lower FPR and FNR, same TPR and a higher TNR for a model that offered FPR parity or reduced bias to the two race-based cohorts.

Another work by Zeng et al. [15] on the same dataset developed the "SLIM scoring system for arrest", used only 1994 records for training and testing purposes, and used future arrests as representative of recidivism. This is not the same as our approach where we use all arrest records and future reconviction to indicate recidivism. This work [15] does not publish its accuracy. Additionally, their emphasis was on the interpretability of results while ours was on increasing accuracy and reducing bias. They had a mean fivefold CV TPR/FPR of 78.3%/46.5%. Our values for the mean of 10 Monte Carlo Validation for sexual crimes (see Figure 6.4) for the two metrics are 94.2%/32.9%. Additionally, their mean fivefold CV TPR/FPR for Sexual Crimes was 43.7%/19.9%, while ours for reconviction with all arrest records and mean of 10 Monte Carlo Validation is 70.2%/8.1%. Thus our model achieved significantly better predictions with much lower error rates.

6.6 Conclusion

In this work we recommend a new approach to increase accuracy in predicting recidivism while decreasing race-based bias in the results. The algorithm trains several models using a race blind dataset. It increases the number of derived features based on offenders' current offenses/activities and prior criminal history. The approach trains models on different numbers of prior arrests. It tests each of these models to check how each model treats the all-race population and different subpopulations,

without ever using a race input feature set. In this approach, we select a model that increases the accuracy while minimizing race-based bias. The results revealed that the method delivers a model that is both more accurate and fairer than without adding personal activities and criminal history information. Our approach can be used for both predicting future general "All Crimes" or specific crime categories like future "Sexual Crimes" convictions. This approach may also be used for applications such as loan application acceptance, dating, hiring, etc where personal history may be emphasized and considered for better predictions.

In future work it would be interesting to study accuracy and bias in other types of crimes and in domains other than recidivism. Varied datasets with varying numbers of prior personal history may be considered for each model. In the recidivism domain, since very few offenders had substance-abuse and treatment related information, it would be interesting to study the effect of more of such information on accuracy and bias. Furthermore, it would be interesting to prod the model during the training period to further reduce the bias using FPR and FNR parity.

CHAPTER 7

Using Bias Parity Score to Find Feature-Rich Models with Least Relative Bias

7.1 Introduction

In this chapter, we extend the work on machine learning (ML) classification-related problems in the criminal justice and recidivism domain. As in the previous chapters, we focus on increasing the accuracy and measurably reducing the bias by using offenders' criminal history, substance abuse, and treatments taken. We further show that using offender history helps us negate the traditional view of the accuracy-versus-fairness trade off by both increasing accuracy and reducing bias while attempting to address the limitations in the previous chapters with respect to the absence of a generic way to quantitatively measure different aspects of fairness. For this, we introduce a new measure called Bias Parity Score (BP score or BPS) that allows us to use one numeric value to compare models based on the bias still embedded in them. In the recidivism classification problem, we again used neural networks on the dataset from the "Recidivism of Prisoners Released in 1994" study [60] to predict recidivism in released offenders. To increase the capabilities of the trained models and to permit selection among multiple models to optimize fairness, we again created models using different numbers of past arrest cycles, evaluated them for bias using the BPS score of several fairness measures, and present the results in this chapter. As we tabulated our results using the various fairness measures and accuracy in conjunction with BPS, we observe that BPS helps us see the quantitative reduction in bias in different bias metrics. Our main contributions here are three-fold: (i) to use the offender history to reduce bias and increase accuracy, (ii) to introduce and define BP score, and (iii) to

use BP score to measure and compare bias in the models generated in our experiments and to facilitate selection of the least biased model without sacrificing significantly in terms of accuracy.

7.2 Bias, Bias Parity Score and Statistical Measures

In this section, we describe bias in the recidivism domain and state the definitions and formula for Bias Parity Score. The definitions of various statistical measures [11, 43] that we will use in our work are included in Section 7.2.2. The statistical measures such as False Positive (FP), False Negative (FN), True Positive (TP), True Negative (TN), False Positive rate (FPR), False Negative Rate (FNR), True Positive Rate (TPR), True Negative Rate (TNR), Positive Predicted Value (PPV), Negative Predictive Value (NPV), False Discovery rate (FDR), False Omission Rate (FOR), etc. are based on the confusion matrix, a table that presents predicted and actual classes in a machine learning classification model. In our representation, the positive class is the true labeling of future recidivistic behavior of an offender, while the negative class is the true labeling of future non-recidivism of an offender.

Bias: In the recidivism domain, underpredicting recidivism for a privileged class and overpredicting recidivism for a disadvantaged class represents bias [25, 10, 15, 29, 35, 18].

Bias Parity Score (BPS): In this work we define a new measure, BP Score, that helps us quantify the bias in a given model. BP Score can be computed as follows, where $measure(A = 0)$ and $measure(A = 1)$ are the values of a given metric for the sensitive and non-sensitive subpopulations, respectively:

$$BPS - measure = \frac{min(measure(A = 1), measure(A = 0)) * 100}{max(measure(A = 1), measure(A = 0))}. \tag{7.1}$$

BPS is 100 when there is no bias as measured by the metric that has the values $measure(A = 0)$ and $measure(A = 1)$ for the two subpopulations represented by the sensitive attribute values of 0 and 1, respectively. BP Score will enable us to see how much more a model needs to improve to be bias-free. As per the work by Won et al. [22], the cost of mislabeling non-recidivistic people is different from mislabeling recidivistic offenders. A practitioner may decide that a minimum BP Score of 90 for FNR and 95 for FPR, for example, is acceptable to strike a balance between releasing a potentially non-recidivating offender and keeping society safe from a future recidivating offender.

BP Score for each statistical measure can be computed by plugging in the values of $measure(A = 0)$ and $measure(A = 1)$ in the BPS formula stated above. In our experiments, the best BPS for most metrics was achieved with a model generated with five prior arrest cycles. If not indicated differently, we state the metric and BPS values for this model.

7.2.1 Interpreting Statistical Measures and Parity in Recidivism

One can observe that the best values of FPR, FDR, PPV, and TNR are 0, 0, 1, and 1, respectively and are achieved when FP is 0 or minimal. This is achieved when the algorithm can identify the negative class easily. In our dataset, this means that the non-recidivists need to be easily identifiable to have optimal values for FPR, FDR, PPV, and TNR. It should be noted that the best values of FOR, FNR, NPV, and TPR are 0, 0, 1, and 1, respectively and are achieved when FN is 0 or minimal, or when the algorithm can distinguish between the positive and the negative classes accurately. In our dataset, this means that the recidivists need to be easily identifiable to have optimal values for FOR, FNR, NPV, and TPR. The best value of Equal Opportunity is 1 and is achieved when both FP and FN are zero.

Furthermore, when non-recidivists and recidivists can be identified accurately, the accuracy of prediction goes up. Therefore, in our methodology, as we included information regarding the history of past criminal activities, substance abuse, and treatments taken, the neural network can improve its ability to distinguish between the positive and the negative class (i.e., the recidivists and the non-recidivists). In our experiments we show that this leads to both an increase in prediction accuracy and a decrease in race-based bias.

7.2.2 Notation and Statistical Measures for Recidivism Prediction

We use the notation included in this section to formulate the problem of predicting recidivism with a single sensitive attribute of race. We need to find the model that helps predict with high accuracy and lowest bias possible amongst the given models, each one of which is developed with the rolling sum of prior individual crimes from N prior arrest records.

$X \in R^d$: quantified features of each elemnt in the dataset.
$A \in \{0, 1\}$: Race-based binary sensitive attribute (African American, Caucasian)
$C \in \{0, 1\}$: predicted variable (not reconvicted/reconvicted)
$Y \in \{0, 1\}$: target variable (will not reoffend/will reoffend).
Assumption: (X, A, Y) are generated from a distribution d denoted as $(X, A, Y) \sim d$.
measure(A') : value of a metric for subpopulation with A = 0.
measure(A) : value of a metric for subpopulation with A = 1.

TP: True Positive (TP) is a correct positive prediction. In our work it means that a future recidivist was correctly forecasted to recidivate.

TN: True Negative (TN) is a correct negative prediction. In our work it means that a future non-recidivist was correctly forecasted to not recidivate.

FP: False Positive (FP) is an incorrect positive prediction, when a future non-recidivist was falsely forecasted to recidivate

FN: False Negative (FN) is an incorrect negative prediction. In our work it means that a future recidivist was erroneously labeled as nonrecidivist.

Positive Predictive Value: PPV, also referred to as Precision, is the total number of True Positive cases divided by a total of all predicted positive cases. The best possible value of PPV is 1 and is achieved when FP becomes zero, i.e., when none of the future non-recidivists are wrongly accused of being a recidivist. The worst value of PPV is 0 and happens when none of the individuals predicted to recidivate are actually recidivists, i.e., when TP is zero. Therefore, it is desirable to have PPV as close to 1 as possible. PPV also refers to the probability of an offender to truly belong to the positive class, $P(Y = 1 \mid C = 1)$. $measure_{PPV}(A = 0) = P(Y = 1 \mid C = 1, A = 0)$ and $measure_{PPV}(A = 1) = P(Y = 1 \mid C = 1, A = 1)$. PPV parity [42] is achieved when $measure_{PPV}(A = 0) = measure_{PPV}(A = 1)$. Average PPV and BPS for the model with 5 past arrest cycles and all race data were 0.853 and 97.9, respectively.

$$PPV = \frac{TP}{TP + FP} \tag{7.2}$$

False Positive Rate: FPR is the total number of incorrect positive predictions divided by a total of all the non-recidivists (FP + TN) in the dataset. The best possible value of FPR is 0 and is achieved when FP becomes zero, i.e., none of the future non-recidivists are wrongly accused of being a recidivist. The worst value of FPR is 1 and happens when all future non-recidivists are falsely labeled as recidivists such that TN becomes 0. Therefore, it is desirable to have FPR as close to 0 as

possible. FPR is represented as $P(C = 1 \mid Y = 0)$. $measure_{FPR}(A = 0) = P(C = 1 \mid Y = 0, A = 0)$ and $measure_{FPR}(A = 1) = P(C = 1 \mid Y = 1, A = 1)$.

FPR parity or False Positive Error Rate Balance [42] or predictive equality [71] is achieved when $measure_{FPR}(A = 0) = measure_{FPR}(A = 1)$. Average FPR and BPS for the model with 5 past arrest cycles and all race data were 0.378 and 83.7, respectively.

$$FPR = \frac{FP}{FP + TN} \tag{7.3}$$

False Negative Rate: FNR is the total number of incorrect negative predictions divided by a total of all the recidivists (FN + TP) in the dataset. The best possible value of FNR is 0 and is achieved when FN becomes zero, i.e., none of the future recidivists are erroneously labeled as a non-recidivist. The worst value of FNR is 1 and happens when all future recidivists are falsely labeled as non-recidivists such that TP becomes 0. FNR is represented as $P(C = 0 \mid Y = 1)$. $measure_{FNR}(A = 0) = P(C = 0 \mid Y = 1, A = 0)$ and $measure_{FNR}(A = 1) = P(C = 0 \mid Y = 1, A = 1)$. FNR parity is achieved when $measure_{FNR}(A = 0) = measure_{FNR}(A = 1)$. Average FNR and BPS for the model with 5 past arrest cycles and all race data were 0.056 and 93.9, respectively.

$$FNR = \frac{FN}{FN + TP} \tag{7.4}$$

It is desirable to have both FPR and FNR as close to 0 as possible. A high FPR represents many future non-recidivists wasting behind bars while a high FNR means many future recidivists let lose to commit many needless crimes in the world. Both FPR and FNR mean different types of error and associated cost to society. Both higher FPR and higher FNR cause a decrease in the predictive accuracy.

False Discovery Rate: FDR is the total number of incorrect positive predictions (FP) divided by a total of all positive predictions (TP + FP). The best possible value of FDR is 0 and is achieved when FP becomes zero, i.e., none of the future non-recidivists are mislabeled as recidivist. The worst value of FOR is 1 and happens when all future recidivists are falsely labeled as non-recidivists such that TP becomes 0. Therefore, it is desirable to have FDR as close to 0 as possible. FDR refers to the probability of a positively labeled individual to actually belong to the negative class, ($P(Y = 0 \mid C = 1)$, or the probability of a person kept incarcerated to be a non-recidivist. $measure_{FDR}(A = 0) = P(Y = 0 \mid C = 1, A = 0)$ and $measure_{FDR}(A = 1) = P(Y = 0 \mid C = 1, A = 1)$. FDR parity is achieved when $measure_{FDR}(A = 0) = measure_{FDR}(A = 1)$. Average FDR and BPS for the model with 5 past arrest cycles and all race data were 0.071 and 89.8, respectively.

$$FDR = \frac{FP}{TP + FP} \tag{7.5}$$

False Omission Rate: FOR is the total number of incorrect negative predictions (FN) divided by a total of all predicted non-recidivists (TN + FN). The best possible value of FOR is 0 and is achieved when FN becomes zero, i.e., none of the future recidivists are mislabeled as non-recidivist and let go. The worst value of FOR is 1 and happens when all future non-recidivists are falsely labeled as recidivists such that TN becomes 0. Therefore, it is desirable to have FOR as close to 0 as possible. FOR refers to the probability of a positive class to be labeled negatively, (P(Y = 1 |C = 0), or the probability of a someone who is let go to be a recidivist. $measure_{FOR}(A = 0) = P(Y = 1 \mid C = 0, A = 0)$ and $measure_{FOR}(A = 1) = P(Y = 1 \mid C = 0, A = 1)$. FOR parity is achieved when $measure_{FOR}(A = 0) = measure_{FOR}(A = 1)$. Average

FOR and BPS for the model with 5 past arrest cycles and all race data were 0.323 and 93.4, respectively.

$$FOR = \frac{FN}{TN + FN} \tag{7.6}$$

Negative Predictive Value: NPV is the total number of True Negatives divided by the total number of negative predictions. The best possible value of NPV is 1. The worst value of NPV is 0 when all negative class are predicted to be positive class such that FN =0. NPV refers to the probability of a negative prediction to truly belong to the negative class, $P(Y = 0 \mid C = 0)$, or the probability of someone predicted to be a non-recidivist to actually be a non-recidivist. $measure_{NPV}(A = 0) = P(Y = 0 \mid C = 0, A = 0)$ and $measure_{NPV}(A = 1) = P(Y = 0 \mid C = 0, A = 1)$. NPV parity is achieved when $measure_{NPV}(A = 0) = measure_{NPV}(A = 1)$. Average NPV and BPS for the model with 5 past arrest cycles and all race data were 0.677 and 96.8, respectively.

$$NPV = \frac{TN}{TN + FN} \tag{7.7}$$

True Positive Rate: TPR is the total number of True Positive cases identified divided by the total number of positive cases. The best possible value of TPR is 1 and is achieved when FN is equal to 0. TPR, also known as sensitivity or recall, is the probability of the positive class to be labeled as positive, $P(C = 1 \mid Y = 1)$ or the probability of a recidivist to be labeled as one. $measure_{TPR}(A = 0) = P(C = 1 \mid Y = 1, A = 0)$ and $measure_{TPR}(A = 1) = P(C = 1 \mid Y = 1, A = 1)$. TPR parity is achieved when $measure_{TPR}(A = 0) = measure_{TPR}(A = 1)$. Average TPR and

BPS for the model with 5 past arrest cycles and all race data were 0.944 and 99.6, respectively.

$$TPR = \frac{TP}{TP + FN} \tag{7.8}$$

True Negative Rate: TNR is the total number of correctly labeled negative predictions divided by all the negative cases. The best possible value of TNR is 1 and is achieved when FP is 0. TNR refers to the probability of a negative class being labeled negative, $P(C = 0 \mid Y = 0)$. This in our predictions is the probability of non-recidivists being labeled a non-recidivist. $measure_{TNR}(A = 0) = P(C = 0 \mid Y = 0, A = 0)$ and $measure_{TNR}(A = 1) = P(C = 0 \mid Y = 0, A = 1)$. TNR parity is achieved when $measure_{TNR}(A = 0) = measure_{TNR}(A = 1)$. Average TNR and BPS for the model with 5 past arrest cycles and all race data were 0.622 and 89.8, respectively.

$$TNR = \frac{TN}{FP + TN} \tag{7.9}$$

Accuracy: Accuracy is the total number of appropriately labeled predictions divided by the number of all the cases. The best possible value of Accuracy is 1 and is achieved when both FP and FN are 0. Accuracy refers to the probability of accurate labeling for both positive and negative classes, P(C = Y). This in our predictions is the probability of being correctly labeled as a recidivist or non-recidivist as the case is. $measure_{ACC}(A = 0) = P(Y = C, A = 0)$ and $measure_{ACC}(A = 1) = P(C = Y, A = 1)$. Accuracy parity is achieved when $measure_{ACC}(A = 0) = measure_{ACC}(A = 1)$.

Average Accuracy and BPS for the model with 5 past arrest cycles and all race data were 0.892 and 99.4 respectively.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7.10}$$

Statistical Parity or Group Fairness or Demographical Parity [72] or Equal Acceptance Rate [73] is the property that the demographics of those labeled with a positive (or negative) classifications is the same as the demographics of the population as a whole [72]. In other words, this is true if all subgroups have equal probability to be labeled as the positive class $P(C = 1 \mid A = 0) = P(C = 1 \mid A = 1)$ [11]. $measure_{S.P.}(A = 0) = P(C = 1, A = 0)$ and $measure_{S.P.}(A = 1) = P(C = 1, A = 1)$ and the condition $measure_{S.P}(A = 0) = measure_{S.P.}(A = 1)$ is satisfied. The fraction of all positive predicted cases from all cases and BPS (statistical parity) for the model with 5 past arrest cycles and all race data were 0.854 and 97.9, respectively.

The situation may be unfair for an individual even as statistical parity is accomplished. As per Dwork et al. [72], this can provide fair affirmative action but may be insufficient in other situations, e.g., if one subgroup has in fact more members with positive class than the other group or when more unqualified members are chosen to fulfill the condition [72]. In our dataset, it is used to give members of different races a possibility of parole but this may not be fair if one group reoffends more.

$$Statistical\ Parity \implies \frac{TP1 + FP1}{TP1 + FP1 + TN1 + FN1} = \frac{TP2 + FP2}{TP2 + FP2 + TN2 + FN2} \tag{7.11}$$

Treatment Equality [74] or False Negative to False Positive Ratio is achieved when the ratio of errors, i.e., False Negatives and False Positives for the subpopulations are equal, such that $measure_{T.E.}(A = 0) = FN1/FP1$ and $measure_{T.E.}(A =$

1) = $FN2/FP2$ and the condition $measure_{T.E.}(A = 0) = measure_{T.E.}(A = 1)$ is satisfied. In our dataset this means that for both African-Americans and Caucasians, the ratio of recidivists labeled as non-recidivists (FN) to non-recidivists labeled as recidivists should be the same. The average fraction of all False Negative cases to False Positive cases from all cases and BPS (Treatment Equality) for the model with five past arrest cycles and all race data were 0.798 and 83.8, respectively.

$$Treatment\ Equality \implies \frac{FN1}{FP1} = \frac{FN2}{FP2}. \tag{7.12}$$

7.3 Experiments

In the current chapter, we show that BP Score can quantify the bias in a given model and represent it using one number per statistical measure for different subpopulations in the dataset. We increase the accuracy and decrease the bias in the generated models by using different numbers of past arrest cycles to include the past criminal activity, substance abuse, and courses taken during incarceration. This increased the accuracy and reduced the bias as the neural network's decision is now based on individuals' crime and activity similarity in the absence of race information from the training dataset.

We used a three-layered neural network model and divided the data using an 80 to 20 ratio for training and testing purposes and utilized 10% of the training dataset for validation purposes. We tuned various hyper parameters to finally select a batch size of 256 training for 80 epochs with a network comprised of 127 neurons in the input and hidden layers and 1 neuron in the output layer as this is a binary classification problem. We used the Keras wrapper for the TensorFlow library to build and train the network. Furthermore, we used Adam as the optimization algorithm

for stochastic gradient descent for training our deep learning model. We excluded race while training the model using the input feature set described in the following section. We evaluated the model for biased results by testing it three ways: once with data from all the races, and twice using the two sets of individual race data while excluding race from the test sets and calculating BP Scores for various measures. We pre-processed our dataset to use some of the existing raw features, ran 10 iterations of each experiment for Monte Carlo cross-validation, computed the average for each metric, tabulated it, and graphed the values with each model's BP Score for each statistical measure. We used a total of 127 normalized numerical and binary input features (from categorical attributes) in our feature vectors to run our experiments.

7.3.1 Input and Output Variables

Input Variables: Dataset 2, the dataset from the "Recidivism of Prisoners Released in 1994" study [60] as described in Section 3.2, provided us with several features that we used as is, for example, date of release, and we derived many others such as the total number of convictions prior to an arrest–release cycle. All direct and derived features could be categorized into demographic features, personal activity during incarceration features, test results before 1994 release features and past criminal activity features. We included all of these while eliminating any related features that did not improve algorithm performance.

Demographic features: This group is comprised of features like date of release or number of state prisoners represented by a convict (WEIGHT). We combined dead or undergoing life sentence into DeadOrLifeCnf, and multiple attributes like the birth day, birth month, birth year and arrest year to derive Arrest Cycle Admission Age (AgeC) and Admission Age for the first arrest cycle (AdAgeC1).

Personal Activity features: This group was comprised of features with information on vocational courses attended, completed or not (VOCAT), educational courses attended, completed, or not (EDUCAT), and behavior modification treatment participated in, completed, or not (SEXTRT). Since these features were not available for any of the arrest-release cycles prior to 1994 release cycles, we recorded -1 for these features for pre-1994 cycles.

The 1994 Test results features: This group encompassed tests taken and their results just prior to the 1994 release cycle. These are comprised of HIV-positive or not (HIV) and substance abuse-related results (DRUGAB, DRUGTRT, ALCABUS, ALCTRT). Since these features were not available for any of the arrest-release cycles prior to 1994 release cycles, we recorded -1 for these features for pre-1994 cycles.

Criminal Activity features: This group included features like the released prisoner's 1994 release offense (SMPOFF26), felony or misdemeanor (J00NFM), conviction for adjudication offense (J00NCNV), confinement for adjudication offense (J00NCNF), confinement length for the most serious adjudication charge (J00NPMX) and crimes adjudicated for (or not) in an arrest cycle (convictions, fatal, sexual, general, property, drug, public, and other categories). In this group we used several features that are derived variables based on an offender's current activities and past criminal history as these affect what they do after being released from prison. Furthermore, individuals tend to repeat the kind of crimes they committed earlier and often do not switch to crime categories very different from what they have committed earlier. Here, we included 26 broad crime categories using binary variable to indicate the crime types committed during a cycle in every record. We included sums of each of the specific 26 crime categories committed in the previous N cycles (cumCR_01, cumCR_02, ... cumCR_26) followed by normalization of these variables. We then encoded the values of the included categorical attributes to binary variables. We

included other derived variables such as the number of years between the admission of past and current arrest cycles (Years_To_LastCyc), sum of crimes adjudicated for in the N previous arrest cycles (CUMcnv , CUMfatal, CUMsexual, CUMgeneral, CUMproperty, CUMdrug, CUMpublic, and CUMother categories), crime count in the N previous arrest cycles of confinement(CUMJ001CNF), involvement in domestic violence (CUMJ001DMV), conviction (CUMJ001CNV), involvement with Fire Arms (CUMJ001FIR), and whether the arrest record was from the 1994 arrest cycle or a later cycle (after94R).

Output Variable: For each arrest cycle, we computed a binary variable indicating whether the offender was adjudicated in a subsequent arrest cycle.

7.3.2 Experiment Setup

Our objective is to increase the accuracy of prediction while reducing bias. We measure bias embedded in predictive models using BPS for various statistical measures. We reduced bias by adding personal history from different numbers of past arrest-release cycles. We split each offender record into multiple records, which contained multiple arrest and release records of each offender along with 99 pre and post arrest–release cycles of the 1994 release cycle. We set up multiple experiments such that each arrest-release cycle incorporated a rolling sum of the prior 0, 1, 3, 5, 7, 10, 15, 20, 40, 60, 80, and 100 crime cycles' criminal activity, substance abuse behavior, and courses taken. Each of our experiments was set up to use records generated using a fixed number of past arrest cycles at a time. This means that in each experiment we used the influence of the past 0 or 1 or 3 or 5 or 7 or 10 or 15 or 20 or 40 or 60 or 80 or 100 arrest cycles. In each experiment, the training data consisted of records from all races. The test data was comprised of three sets: records of offenders of all races, records of offenders of only the Caucasian race, and records

of offenders of only the African American race. We used the records from the two dominant races in the dataset, i.e., Caucasian and African American, to compute the BPS while ignoring the data from other races, such as Asian, as the number of corresponding records was very small.

Using this setup we then studied the effect of different numbers of past arrest cycles in the dataset on both accuracy and bias for various measures. Evaluating the BP scores for different arrest cycle histories then provides a means of selecting from among these models the one that provides the predictions with the highest degree of fairness. This model, in turn, is then compared agains other techniques that were applied to the same dataset both in terms of prediction accuracy and prediction bias.

The benefit of this data augmentation and model selection process, as illustrated by the results presented below, is that we can obtain a model that not only achieves significantly higher accuracy as compared to previous work, but that we can do so while also improving fairness by reducing bias. Most of this is achieved through the data augmentation which results in datasets that are multiple times larger than the original dataset. In this case, the dataset grew from 38,624 records to more than 442,000 records. However, since we are using neural networks for our classifier, this increase in data set does not directly reflect in the required training time (and thus computational complexity) for each model. In our experience, training with the extended set did not significantly increase training time for each model. The main computational cost of our approach thus results from the need to train and evaluate multiple models with different history length in order to be able to select the lowest bias model for the desired BP scores. The model selection approach presented here thus increases training effort linearly in the number of history lengths used, a cost that is easily compensated by the improvement in performance and the reduction in bias achieved.

7.3.3 Bias Metrics for Experiment Results

Tables 7.1–7.6 show average accuracy, several bias metric values, and their BP Scores for All Crimes models with the rolling sum of individual prior crimes for $0, 1, 3, 5, 7, 10, 15, 20, 40, 60, 80,$ and 100 prior arrest cycles using 10-fold Monte Carlo cross-validation, In each table, the results for all races, Caucasians, and African Americans are listed for each arrest cycle number as well as the BPS score indicating the level of parity between Caucasian and African American populations. Best results in each of the categories are indicated in bold. In recidivism, ACC, FPR, and FNR are frequently used to measure bias. Since some of the prior works on this dataset (See Section 7.4) that we compare our results with use TPR and TNR to measure results, we have included these, too. The results in these tables show consistently that the incorporation of past release cycles in the dataset improves fairness with the highest fairness achieved with either 5 or 20 release cycles, depending on the statistical measure (as indicated by the bolded entries for the BP Scores).

To further study the effect of the number of arrest cycles on the BP scores for the different statistical measures, Figures 7.1 and 7.2 show the BP scores for different statistical measures plotted against the number of arrest cycles used.

These graphs show two important observations, namely (i) that for all measures the inclusion of past arrest records in the data significantly improved the fairness (i.e., the BP score), and (ii) that for different measures the detailed relation of arrest cycles to BP score behaved differently. In particular, behavior seemed to fall into one of three groups, indicated with different colored graphs in the figures. In the first of these groups (comprising ACC, NPV, and FOR), fairness increased until around 40 arrest

Table 7.1: Average Accuracy.

Arrest Cycles	**All Races Train/Test**	**All Races Train, Test Caucasian**	**All Races Train, Test Afr. Amer.**	**BPS**
0	0.879	0.883	0.875	99.032
1	0.888	0.893	0.884	99.062
3	0.891	0.894	0.888	99.383
5	0.892	0.895	0.890	99.419
7	0.894	0.897	0.892	99.420
10	0.896	0.899	0.894	99.399
15	0.898	0.900	0.896	99.484
20	0.899	0.901	0.897	**99.583**
40	0.899	**0.902**	0.898	99.557
60	**0.900**	0.902	**0.898**	99.530
80	0.899	0.902	0.897	99.536
100	0.899	0.902	0.897	99.440

Table 7.2: Average Positive Predictive Rate.

Arrest Cycles	**All Races Train/Test**	**All Races Train, Test Caucasian**	**All Races Train, Test Afr. Amer.**	**BPS**
0	0.853	0.838	**0.867**	96.645
1	0.857	0.843	0.865	97.412
3	0.848	0.838	0.858	97.740
5	**0.853**	0.844	0.862	**97.915**
7	0.849	0.838	0.858	97.657
10	0.852	0.842	0.862	97.648
15	0.847	0.835	0.855	97.717
20	0.844	0.835	0.854	97.779
40	0.853	**0.842**	0.863	97.544
60	0.845	0.835	0.854	97.716
80	0.842	0.832	0.851	97.773
100	0.846	0.836	0.855	97.804

Table 7.3: Average False Positive Rate.

Arrest Cycles	**All Races Train/Test**	**All Races Train, Test Caucasian**	**All Races Train, Test Afr. Amer.**	**BPS**
0	0.417	0.360	0.474	75.938
1	0.393	0.347	0.438	79.134
3	0.366	0.330	0.401	82.150
5	0.378	0.344	0.411	**83.742**
7	0.357	0.321	0.393	81.542
10	0.362	0.324	0.398	81.433
15	0.333	0.300	0.364	82.448
20	0.329	0.297	0.360	82.481
40	0.354	0.317	0.389	81.496
60	0.327	0.293	0.359	81.591
80	**0.320**	**0.287**	**0.351**	81.798
100	0.333	0.299	0.365	81.751

Table 7.4: Average True Negative Rate.

Arrest Cycles	**All Races Train/Test**	**All Races Train, Test Caucasian**	**All Races Train, Test Afr. Amer.**	**BPS**
0	0..583	0.640	0.526	82.156
1	0.607	0.653	0.562	86.016
3	0.634	0.670	0.599	89.308
5	0.622	0.656	0.589	89.810
7	0.643	0.679	0.607	89.320
10	0.638	0.676	0.602	89.072
15	0.664	0.700	0.636	90.881
20	0.671	0.703	0.640	**91.041**
40	0.646	0.683	0.611	89.456
60	0.673	0.707	0.641	90.637
80	**0.680**	**0.713**	**0.649**	91.041
100	0.667	0.701	0.635	90.492

Table 7.5: Average False Negative Rate.

Arrest Cycles	**All Races Train/Test**	**All Races Train, Test Caucasian**	**All Races Train, Test Afr. Amer.**	**BPS**
0	0.059	0.069	0.061	88.059
1	0.058	0.061	0.056	92.904
3	0.060	0.063	0.059	93.262
5	0.56	0.059	0.055	**93.938**
7	0.058	0.061	0.056	92.058
10	0.055	**0.057**	0.053	91.688
15	0.058	0.060	0.056	93.258
20	0.058	0.061	0.056	91.740
40	**0.053**	0.056	**0.050**	88.990
60	0.057	0.060	0.055	91.598
80	0.059	0.062	0.057	92.288
100	0.057	0.059	0.055	93.019

Table 7.6: Average True Positive Rate.

Arrest Cycles	**All Races Train/Test**	**All Races Train, Test Caucasian**	**All Races Train, Test Afr. Amer.**	**BPS**
0	0.935	0.931	0.939	99.123
1	0.942	0.939	0.944	99.544
3	0.940	0.937	0.941	99.551
5	0.944	0.941	0.945	**99.624**
7	0.942	0.939	0.944	99.488
10	0.945	**0.943**	0.947	99.496
15	0.942	0.940	0.944	99.568
20	0.942	0.939	0.944	99.470
40	**0.947**	0.944	**0.950**	99.353
60	0.943	0.940	0.945	99.467
80	0.941	0.938	0.943	99.495
100	0.943	0.941	0.945	99.565

Figure 7.1: Bias Parity Score (BPS) by number of past arrest-release cycles (0, 1, 3, 5, 7, 10, 15, 20, 40, 60, 80, 100).
Group 1: BPS-Avg. Accuracy, BPS-Avg. Negative Predicted Value, BPS-Avg. False Omission Rate.
Group 2: BPS-Avg. False Negative rate, BPS-Avg. True Positive Rate, BPS-Avg. FN-to-FP-ratio. **Averages**: Computed using 10 iterations of Monte Carlo cross-validation.

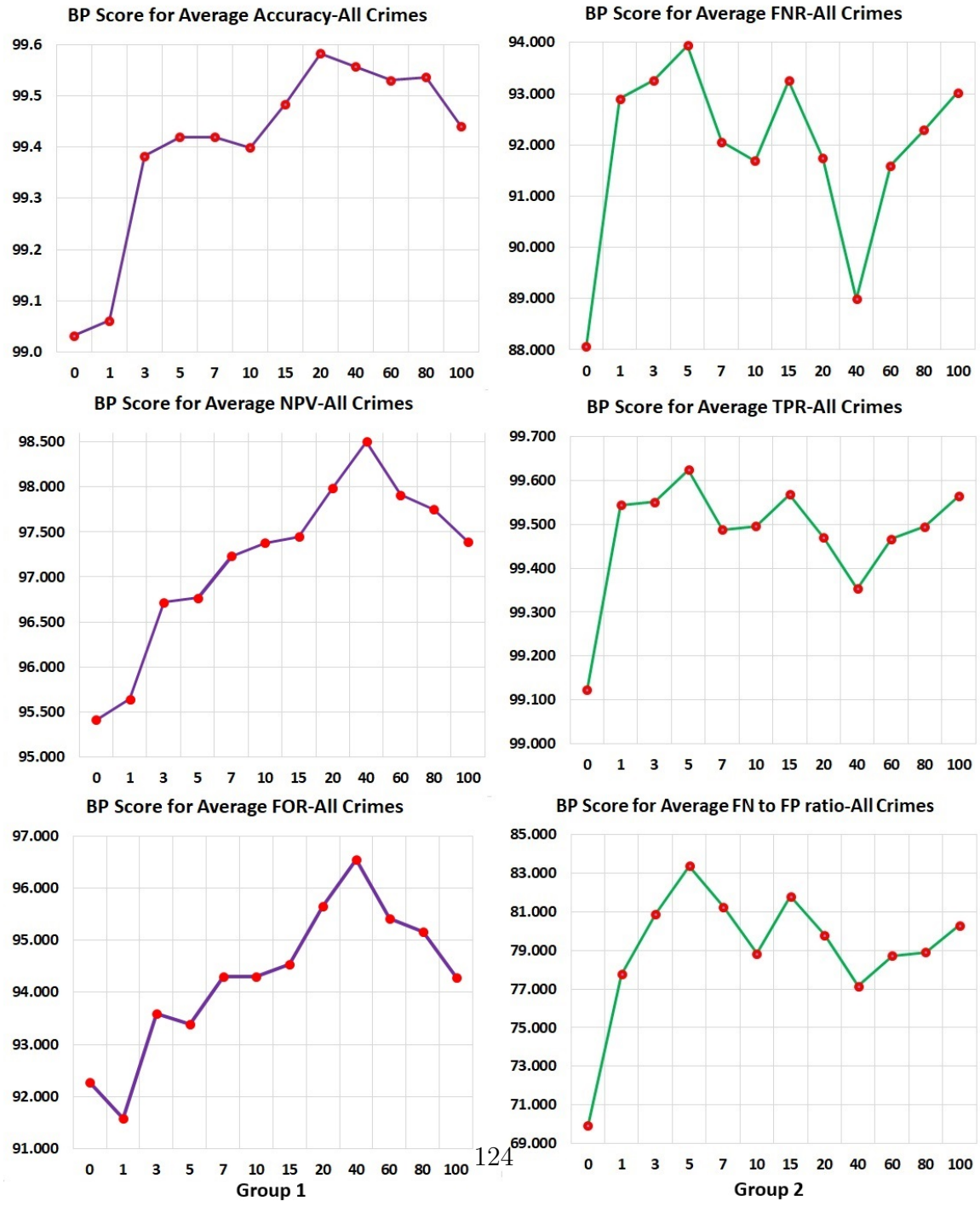
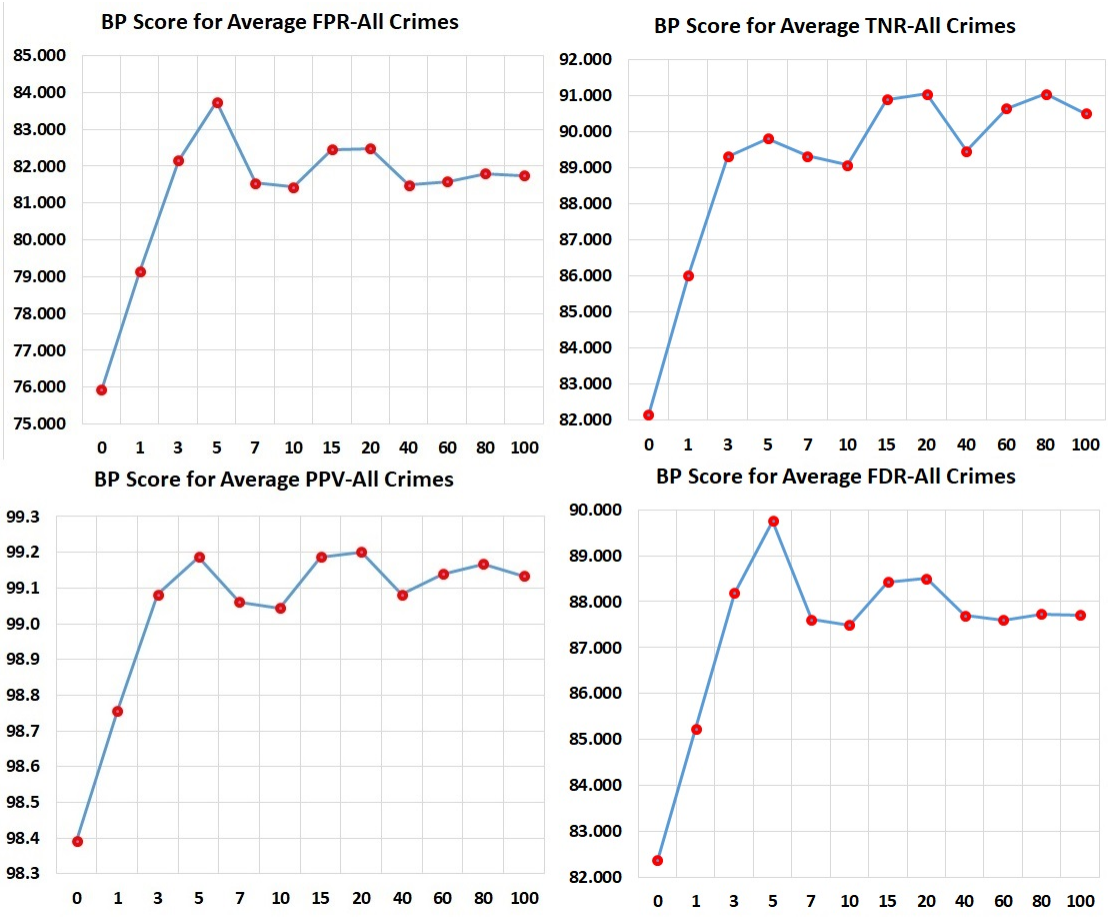


Figure 7.2: BP Scores by number of past arrest-release cycles (0, 1, 3, 5, 7, 10, 15, 20, 40, 60, 80, 100).
Group 3: BPS-Avg.False Positive Rate, BPS-Avg. True Negative Rate, BPS-Avg. Positive Predictive Value, Avg.False Discovery Rate. **Averages**: Computed using 10 iterations of Monte Carlo cross-validation



cycles and then dropped. In the second group (comprising FNR, TPR, and FN to FP ratio), fairness (BP score) peaked at five arrest cycles and then dropped slightly, while in the third group (comprising FPR, TNR, PPV, and FDR), fairness reached close to highest performance again at five cycles and then stayed relatively constant. These results suggest that we can use the desired balance of measures to pick the least biased model for later use, In this case, since we will compare against systems that largely evaluated their models in terms of FPR, FNR, TPR, and TNR, we selected

our model with five prior arrest cycles for the comparison shown in Table 7.7. More detailed discussion of the results and observations are provided in Sections 7.4 and 7.5.

Table 7.7: Performance evaluation comparative results with the same dataset.

Work	**ACC**	**FPR**	**FNR**	**TPR**	**TNR**
Ozkan. et al. [18]	77.8% XGBoost	40.6% Log.Reg.	05.4% SVM	94.6% SVM	59.4% Log.Reg.
Zeng et al. [15]	-	46.5%	-	78.3 %	-
Current Work	89.2% BPS 99.4	37.8% BPS 83.7	5.6% BPS 93.9	94.4% BPS 99.6	62.2% BPS 89.8

7.4 Performance Evaluation

To evaluate the overall performance of our model, we compare the performance of our lowest bias model with the results of two previous papers that used the same dataset.

The work by Ozkan [18] utilized six classifiers, namely, Logistic Regression, Random Forests, XG-Boost, Support Vector Machines, Neural Networks, and Search algorithms on the same dataset as used in this work. Just like Zeng et al. [15], they used only the 1994 records for training and test purposes while we used all arrests cycles for training and testing purposes. Zeng et al. [15] used rearrests as representative of recidivism while [18] the current work uses reconvictions as indicators of recidivism. Though the previous two works used some history, we use a more detailed history that takes the rolling sum of 26 different types of crimes, substance abuse and courses taken into consideration for each cycle.

The work by Ozkan [18] achieved the best results for Accuracy using XGBoost at 0.778, for FPR and TNR using Logistic Regression at 0.406 and 0.594, respectively, and for FNR and TPR using SVM at 0.054 and 0.946, respectively. For our comparison, we will use these best scores even though they are not achieved by one consistent model and can thus not be achieved simultaneously by their system.

In contrast to the comparison data from Oskan, we used a single model for all of the performance results for our model by picking the model with the lowest average bias in all the comparison categories which turned out to be the one considering five past arrest cycles. The classifier trained here is a single neural network as described in Section 7.3 and thus achieves all comparison results simultaneously. This model achieved an accuracy of 0.892, FPR of 0.378, FNR of 0.056, TPR of 0.944, and TNR of 0.622. Comparing this with Ozkan's results shows that our model achieves significantly higher accuracy while obtaining better FPR and TNR, with approximately equivalent FNR and TPR compared to their best values in each of the categories.

Zeng et al. [15], who were pursuing transparency rather than accuracy or reduction of bias, achieved a mean five-fold cross-validation TPR and FPR of 78.3% and 46.5%, respectively, while we achieved 94.4% and 37.8% for these metrics using a 10-fold Monte Carlo cross-validation, thus significantly improving in both measures.

Neither of the comparisons directly evaluated a single fairness score and their results are thus optimized irrespective of fairness across groups. To test that our model can not only achieve better performance but also fairer predictions, we evaluated the BP scores. The achieved BP scores of 99.4 for Accuracy, 83.7 for FPR, 93.9 for FNR, 99.6 for TPR, and 89.8 for TNR show that our system succeeded at obtaining high levels of parity in all of the metrics. See Table 7.7 for the complete comparison.

7.5 Results and Discussion

We used a three-layer deep learning neural network on a recidivism dataset. The dataset had more Caucasian records than African American records. However, the African American offenders had more rearrests than Caucasians. As a result, when we split the original offender 1994 release records that included up to 99 prior and/or subsequent arrest records to create multiple records, each with a single arrest and subsequent release information, there were more African American records than Caucasian records. When we used neural networks to predict recidivism without each prior arrest's history being included in the records, the results had significant bias embedded in them. The inclusion of information from past arrests increased accuracy and reduced bias. We conducted experiments by including history information from 0, 1, 3, 5, 7, 10, 15, 20, 40, 60, 80, and 100 past arrest cycle records. This increased the accuracy and decreased the bias using the various given metrics. We used BP Score for each metric for each of the given sets of experiments and found that for most metrics, the model generated with the data using the past five arrest cycles' information had the least bias. See Tables 7.1–7.6 for the detailed results of the different models. If it was not the least bias, it was close to the least bias amongst the given BP scores.

The results from processing the raw data without a prior history, even in the absence of race, designates more recidivism labels to non-recidivating African Americans (higher FPR) than to non-recidivating Caucasians (lower FPR). Furthermore, it assigns fewer non-recidivating labels to recidivating African Americans (lower FNR) and more non-recidivating labels to recidivating Caucasians (higher FNR). This trend was reduced as we included a rolling sum of 26 types of individual crimes, substance abuse, and courses and treatments taken during incarceration for 0, 1, 3, 5, 7, 15, 20, 40, 60, 80, and 100 past arrest records. We can determine if a classifier is fair

by using the BP Score that compares metric values for the different subpopulations. Only a practitioner can decide whether having a given BP Score, for example 90, for all or a subset of the fairness metrics can be considered fair or the model needs to be improved further. As the quality of data improves and the ethical standards of our society are raised, we assume that society will continue to demand an ever-increasing value of the BP score, until it is close to a near-perfect 100. The BP score can allow a practitioner to identify bias in a model using different statistical measures, where 100 and 0 are the best and the worst values, respectively. In recidivism, high FPR means more non-recidivists will have to languish in jail while a high FNR means that society is at risk, because recidivists are being let out of jail. A parole officer may decide to use BPS for both Accuracy and FPR above 90 and for FNR above 80. On the other hand, a loan officer may use a minimum BPS for False Negative Error Rate Balance (Equal Opportunity) [42, 72] for a model to determine the loan worthiness of a client. Any new fairness measures that may evolve after this paper can also use BP Scores to determine the bias embedded in models and choose the one with the highest BPS or the least bias amongst the given models.

We plotted BP Scores for different metrics based on models created using different numbers of past arrest cycle details (0, 1, 3, 5, 7, 10, 15, 20, 40, 60, 80, 100 prior arrests). We saw our BP Scores fall into three groups as shown in Figure 7.1 and 7.2: Group 1 is comprised of BPS-Average Accuracy, BPS-Average Negative Predicted Value, and BPS-Average False Omission Rate. This group experiences a steady increase in BPS until around 40 past arrest cycles of information. This means that considering long histories improve these metric values until only very few individuals are in the long history category, upon which fairness suffers (BPS goes down). Group 2 consists of BPS-Average False Negative Rate, BPS-Average True Positive Rate, and BPS-Average FN-to-FP-ratio (Treatment Equality). For these metrics, short histo-

ries are beneficial but then have significant dips in BP scores with additional history. This means that long histories can produce significant bias issues in Average False Negative Rates, Average True Positive Rates, and Average False Negative to False Positive Rates (Treatment Equality). Group 3 is comprised of BPS-Average False Positive Rate, BPS-Average True -Negative Rate, BPS-Average Positive Predictive Value, and Average False Discovery Rate. These show similar effects as Group 2 early on. This group shows a rapid increase in BP score by adding a little bit of history but is then largely unaffected when increasing the history further. As a result of looking at the BP scores, especially for Group 2, we chose the model created with information from five past arrest cycles. This model minimizes bias reflected in most metrics used in this work. A decision maker can use BPS similarly. Averages have been calculated using 10 fold Monte Carlo cross-validation.

7.6 Limitations

We have used the "Recidivism of Prisoners Released in 1994" [60] dataset and our work is limited by the data provided in it, which in turn is limited by the time frame for which data is captured. Prisoner records for up to 1997 (and in some cases beyond 1997) have been included in the dataset. We have utilized all of these in our experiments. It is possible that a number of individuals committed crimes after 1997 but their criminal activities were not included in the dataset. Additionally, race is stored in our dataset as a discrete value. We have accordingly assumed racial disjointness of the offenders, as done in substantial amounts of the related literature, including [25, 15, 18, 10] and many more. However, demographic groups intersect and it is likely that mixed-race offenders are recorded as only one of the races. Furthermore, we use re-conviction as an indicator of recidivism, as opposed to using re-arrest,

as frequent re-arrest could be indicative of human bias. However, if there was human bias in the reconviction decisions, algorithmic decisions will embody them, too.

7.7 Conclusions

In this chapter, we used a three-layered neural network classifier to train a recidivism predictor and computed several statistical measures (defined in Section 7.2) for a recidivism dataset and demonstrated how these can be used to compute the fairness of a prediction model. We introduced a new metric called BP score, which quantifies the bias in models and allows easy comparison of model results to identify the one with the least bias. The original recidivism dataset was comprised of one record per offender, where each record was a 1994 release record and included up to 99 prior and/or subsequent arrest-release records in each record. We used each record to mint multiple records, each comprised of a single arrest and subsequent release event. We then made multiple models, enriched with features containing different numbers of past arrest cycles and related offenders' criminal history, substance abuse history and any courses or treatments taken during incarceration. We demonstrate that by adding personal history we could increase prediction accuracy and measurably decrease bias. We demonstrate how to use BP score to measure reduction in bias using any of the statistical measures described in the paper. Using the metric BP score for the demonstration of the quantitative measure of the reduction of bias using one number is an important contribution of this paper because BP score can be used to compare bias embedded in multiple models and can be used with any underlying statistical measure. Using past history to increase prediction accuracy and decrease bias is important because similar techniques can be applied to applications in many domains like dating, loan application acceptance, hiring, granting parole, insurance coverage, and medical diagnosis as each of these have related history.

7.8 Future Work

As machine learning-based decision support systems continue to be deployed in many domains, the bias and fairness of their results have become significantly pertinent for our world. As a result, the need to alleviate bias in the resulting predictive models has become increasingly important to keep our world fair. The impact of our work is very broad and can be applied in multiple areas such as medical diagnosis, loan application acceptance, dating, hiring, etc. Furthermore, many sensitive attributes such as gender, race, age, income, weight, zip code, mental health status, etc. [10, 9] can influence the unfairness in prediction results. Their effects in different domains can be studied and mitigated with the methodology used in this work. In the current chapter, we developed an approach to adding history to the input vector that increased accuracy of prediction models. Yet another contribution of the current chapter is the fairness measure BP score that can be used to quantify bias in the prediction models for different statistical measures. Both of these lead us to multiple directions of future work.

So far we have applied the approach to predict total crimes. Using the enriched recidivism dataset we could predict the three most serious crimes committed after release on parole as the current dataset stores three most serious crimes committed in each arrest-release cycle. This would allow more detailed predictions and should be followed up by measuring bias in these enhanced prediction models.

When making more complex predictions for different crime categories, it might become important to develop new measures to measure the accuracy of multiple types of concurrent crimes (outcomes) in the prediction model that could also consider the severity of the individual crimes. This should be followed up by creating corresponding models, verifying their accuracy and efficacy, and evaluating the presence of bias in them.

While the current work uses the BP scores to select a model, it does not use them to modify the training of the model. To address this we will in the next chapter develop different BP Score-based loss functions, namely FPR-BPS, FNR-BPS, TPR-BPS, TNR-BPS, FPR-FNR-BPS, and TPR-TNS-BPS, to train models with multiple recidivism-based datasets. The bias in the resulting models will be measured using BP Scores for the pertinent statistical measures to see if even better models can be achieved when explicitly pursuing fairness during training.

We could apply similar techniques to first enrich datasets from dating, hiring, medical diagnosis, and other domains, followed by measuring bias in these models with BP score. Similarly, we could consider applying these techniques in the context of a broader range of potential factors for bias, using them to reduce age, income, weight, zip code, or mental health status-based bias in other domains.

CHAPTER 8

Increasing Fairness in Predictions Using Bias Parity Score Based Loss Function Regularization

8.1 Introduction

The use of automated decision support and decision-making systems (ADM) [53] in applications with direct impact on people's lives has increasingly become a fact of life, e,g. in criminal justice [75, 35, 10], medical diagnosis [75, 1], insurance [6], credit card fraud detection [76], electronic health record data [77], credit scoring [78] and many more diverse domains. This, in turn, has lead to an urgent need for study and scrutiny of the bias-magnifying effects of machine learning and Artificial Intelligence algorithms and thus their potential to introduce and emphasize social inequalities and systematic discrimination in our society. Appropriately, much research is being done currently to mitigate bias in AI-based decision support systems [1, 75, 54, 49, 52, 38].

Bias in Decision Support Systems. As our increasingly digitizing world stows away more data with the passing of each day, more and more decision makers are using AI based decision support systems. With this, the need to keep the decisions of these systems fair for people of diverse backgrounds becomes essential. Groups of interest are often characterized by sensitive attributes such as race, gender, affluence level, weight, and age to name a few. While machine learning based decision support systems do often not consider these attributes explicitly, biases in the data sets, coupled with the used performance measures can nevertheless lead to significant discrepancies in the system's decisions. For example, many minorities have

traditionally not participated in many domains such as loans, education, employment in high paying jobs, receipt of health care. This can lead to unbalanced datasets as the minority-based data may be combined with the majority sensitive attribute-based data. Similarly, some domains like homeland security, refugee status determination, incarceration, parole, loan repayment etc., may be already riddled with bias against certain subpopulations, even in the absence of AI based decision support system. Human bias seeps into the datasets used for AI based prediction systems which, in turn, amplify it further. Thus, as we begin to use AI based decision support system, it becomes important to ensure fairness for all who are affected by these decisions.

Contributions. We propose a technique that uses Bias Parity Score (BPS) measures to characterize fairness and develops a family of corresponding loss functions that are used as regularizers during training of Neural Networks to enhance fairness of the trained models. The goal here is to permit the system to actively pursue fair solutions during training while maintaining as high a performance on the task as possible. We apply the approach in the context of six fairness measures based on parity across groups and investigate multiple loss function formulations and regularization weights in order to study the performance as well as potential drawbacks and deployment considerations. In these experiments we show that, if used with appropriate settings, the technique reduces bias in the models and outdoes the cross-entropy alone loss function mode on two recidivism related datasets as well as on a commonly used Adult Income dataset in terms of accuracy, false positive rate, false negative rate, true positive rate, and true negative rate. By using these Loss functions, we were able to measurably reduce bias in the recidivism results for the two race-based cohorts and demonstrate on the gender-based Adult Income dataset that the proposed method, for appropriate choice of regularization loss function, can outperform state-of-the art techniques aimed at more targeted aspects of bias and

fairness. In addition, we investigate potential divergence and stability issues that can arise when using these fairness loss functions, in particular when shifting significant weight from accuracy to fairness.

In the work presented here, we utilize a quantitative measure of parity as our concept of fairness in the form of Bias Parity Score applied to metrics that include FPR, FNR, TPR, TNR, and FPR, as well as combinations thereof.

8.2 Notation

In this chapter we will present the proposed approach to actively improve fairness in the context of trained Neural Network models largely in the context of class prediction tasks with groups of interest indicated by a binary sensitive attribute. However, the approach can easily be extended to multi-class sensitive attributes across which fairness is to be enforced as well as to regression tasks, both of which will be briefly discussed. For the description of the approach in the context of our primary prediction domain, we use the following notation to describe the classification problem and measures to capture group performance measures to be used to assess prediction bias:

For each element of the dataset, $D = \{(X, A, Y)_i\}$, each data point is represented as an attribute vector (X, A), where $X \in R^d$ represents the quantified features of each element in the dataset, and $A \in \{0, 1\}$ is the binary sensitive attribute indicating the group the data item belongs to. $C \in \{0, 1\}$ where C indicates the value of the predicted variable, and $Y \in \{0, 1\}$ is the true value of the target variable in the training set.

To obtain fairness characteristics, we use parity over statistical performance measures, m, across the two groups indicated by the sensitive attribute, where $m(A)$

indicates the measure for the subpopulation where the sensitive attribute $A = 1$, and $m(A')$ representing the measure for the subpopulation where $A = 0$.

Since this chapter will introduce the proposed framework in the context of binary classification prediction, the fairness related performance measures used here will center around different elements of the confusion matrix and in particular False Positive Rates, False Negative Rates, True Positive Rates, and True Negative Rates. In the notation used here, these metrics as well as the corresponding measures can be represented as:

False Positive Rate: $P(C = 1 \mid Y = 0)$.

$$m_{FPR}(A = 0) = P(C = 1 \mid Y = 0, A = 0)$$

$$m_{FPR}(A = 1) = P(C = 1 \mid Y = 0, A = 1).$$

False Negative Rate: $P(C = 0 \mid Y = 1)$.

$$m_{FNR}(A = 0) = P(C = 0 \mid Y = 1, A = 0)$$

$$m_{FNR}(A = 1) = P(C = 0 \mid Y = 1, A = 1).$$

True Positive Rate: $P(C = 1 \mid Y = 1)$.

$$m_{TPR}(A = 0) = P(C = 1 \mid Y = 1, A = 0)$$

$$m_{TPR}(A = 1) = P(C = 1 \mid Y = 1, A = 1).$$

True Negative Rate: $P(C = 0 \mid Y = 0)$.

$$m_{TNR}(A = 0) = P(C = 0 \mid Y = 0, A = 0).$$

$$m_{TNR}(A = 1) = P(C = 0 \mid Y = 0, A = 1).$$

Strict parity in any of these or similar measures is achieved when the measure for both groups are identical, i.e. if $m(A) = m(A')$.

Definitions of Fairness. Current literature on fairness recommends several formal concepts of fairness which require that one or more demographic or statistical properties are held across multiple subpopulations in the corpus. Demographic parity, also referred to as statistical parity, mandates that the decision rates are inde-

pendent of the values of a sensitive attribute that represents membership of different subgroups [54, 48, 55, 56]. For binary classification problems this is often mathematically represented as $P(C = 1|A = 0) = P(C = 1|A = 1)$, where $C \in \{0, 1\}$ is the decision made by the system. This, criterion, however, makes an underlying equality assumption between the subpopulations which might not hold for all problems. To address this, several recent works [53] focus on error rate balance where fairness requires subpopulations to have equal False Positive Rates (FPR) or equal False Negative Rates (FNR) or both. Another commonly used parity condition is equality of odds which commands equal True Positive Rate (TPR) and equal True Negative Rate (TNR). While perfect parity as a constraint would be desirable, it often is not achievable and thus quantitative measures representing the degree of parity have to be used [57]. Refer to [43, 40, 11] for a more complete recent survey of computational fairness metrics.

In the work presented here, we utilize a quantitative measure of parity as our concept of fairness in the form of Bias Parity Score applied to metrics that include FPR, FNR, TPR, TNR, and FPR, as well as combinations thereof.

8.3 Approach

As indicated above, achieving increased fairness or reduced bias in decision support systems is an important property and a range of techniques have been proposed to measure fairness. The work presented here is aimed at providing a framework to allow a deep learning-based prediction system to learn fairer models for known sensitive attributes by actively pursuing improved fairness during training. For this, we first need a quantitative measure of fairness that can be widely addressed. We propose to use the Bias Parity Score (BPS) which evaluates the degree to which a common measure in the subpopulations described by the sensitive attribute is the

same. Based on this fairness measure we then derive a family of corresponding loss functions that are differentiable and can thus be used as part of the training loss function as a regularization term in addition to the original task performance loss.

8.3.1 Bias Parity Score (BPS)

In machine learning based predictions, bias is the differential treatment of "similarly situated" individuals. It manifests itself in unfairly benefitting some groups while posing others at an unfair disadvantage [79, 25, 10, 15, 29, 35, 18]. Bias in recidivism, for example, may be observed in a higher FPR accompanied by lower FNR for one subgroup which implies that individuals in this subgroups are more often incorrectly predicted to reoffend and thus denied parole, putting them at an unfair disadvantage. Similar bias may be observed when predicting whether to invite job candidates for interviews or making salary decisions, resulting in preferential hiring opportunities or salary offers to one cohort and the opposite for another cohort.

To capture this, the relevant property can be captured as a measure and parity between subgroups can be defined as equality for this measure. In this work we capture the similarity of the measures for the subgroups quantitatively in the form of Bias Parity Score (BPS) as introduced in Chapter 7 and published in [57]. BPS is a fairness measure that helps us quantify the bias in a predictive model in a single number. Given that $m_s(A = 0)$ and $m_s(A = 1)$ are the values of any given statistical measure, s, for the sensitive and non-sensitive subpopulations, respectively, BPS can be computed as:

$$BPS_s = 100\frac{min(m_s(A = 1), m_s(A = 0))}{max(m_s(A = 1), m_s(A = 0))}. \tag{8.1}$$

where the multiplication factor of 100 leads to a percentage measure, making it easier to read.

A BPS of 100 represents perfect parity and thus the highest possible degree of fairness while a BPS of 0 represents maximal bias between the groups. BPS of any statistical measure is thus the ratio of that statistical measure between the two groups. The smaller and the greater statistical measure values are used in the numerator and the denominator, respectively, to compute a symmetric measure for two groups. BPS provides us with a universal fairness measure that can be applied to any property across subgroups and to evaluate fairness in any predictive model generated using any given machine learning classifier.

We here defined BPS in terms of a binary sensitive attribute, i.e. in the context of fairness between two groups. While this is the case we will investigate in this dissertation when evaluating the benefits of translating this to loss functions and using it to train a Neural Network predictor, this fairness measure can relatively easily be expanded to a situation with a multi-valued sensitive attribute $A \in \{a_1, ...a_k\}$ in a form such as:

$$BPS_s = \sum_{a_i} \frac{100}{k} \frac{min(m_s(A = a_i), m_s())}{max(m_s(A = a_i), m_s())}$$

where $m_s()$ is the value of the underlying measure for the entire population and thus measures fairness as the average bias of all classes compared to the population average. Similarly, the worst bias of any class could be used instead of the average.

Even though, BPS can be used for many statistical measures as described in Chapter 7 [57] we will use it for evaluation purposes here with False Positive rate (FPR), False Negative rate (FNR), True Positive rate (TPR), and True Negative rate (TNR) as these are often used for classification systems. In addition we will utilize it on prediction rate to obtain a BPS equivalent to quantitative statistical parity in order to facilitate comparisons with previous approaches in the Adult Income domain.

8.3.2 BPS-Based Fairness Loss Functions

While BPS scores of statistical entities such as FPR represent a measure of fairness, it does not lend itself directly to training a deep learning system since it is not generally differentiable. To address this, we need to translate the underlying measure into a differentiable form and combine it into a differentiable version of the BPS score that can serve as a training loss function. As we are using FPR, FNR, TPR, and TNR here as measures we first have to define continuous versions of these functions. For this, we build our Neural Network classifier to have a logistic activation function as the output (or a softmax if using multi-attribute predictions), leading, when trained for accuracy using binary cross-entropy to the output, y, of the network to represent the probability of the positive class, $y = P(Y = 1|X, A)$. Note that in the experiments performed, we will withhold the sensitive attribute from the input of the classifier and thus practically $y = P(Y = 1|X)$ in our experiments. Using this continuous output, we can redefine a continuous measure approximation $mc_s()$ for FPR, FNR, TPR, and TNR as:

$$\begin{aligned} mc_{FPR}(A = k) &= \frac{\sum_{(X,A,Y)_i:A_i=k,Y_i=0} y_i}{\sum_{(X,A,Y)_i:A_i=k} y_i} \\ mc_{FNR}(A = k) &= \frac{\sum_{(X,A,Y)_i:A_i=k,Y_i=1} (1-y_i)}{\sum_{(X,A,Y)_i:A_i=k} (1-y_i)} \\ mc_{TPR}(A = k) &= \frac{\sum_{(X,A,Y)_i:A_i=k,Y_i=1} y_i}{\sum_{(X,A,Y)_i:A_i=k} y_i} \\ mc_{TNR}(A = k) &= \frac{\sum_{(X,A,Y)_i:A_i=k,Y_i=0} (1-y_i)}{\sum_{(X,A,Y)_i:A_i=k} (1-y_i)} \end{aligned} \tag{8.2}$$

It is important to note that this is not equal to $m_s()$ as it is sensitive to deviations in the exact prediction probability, for example if the output changes from 0.6 to 0.7 the continuous measure, $mc_s()$, changes while the full measure, $m_s()$, would not change since both would result in the positive class.

To reduce this discrepancy, we designed a second measure approximation, $ms_s()$, that uses a sigmoid function, $S(x) = \frac{1}{1+e^{-x}}$, to more closely approximate the full statistical measure by reducing the occurrences of intermediate prediction probabilities:

$$\begin{aligned} ms_{FPR}(A=k) &= \frac{\sum_{(X,A,Y)_i:A_i=k,Y_i=0} S(y_i-0.5)}{\sum_{(X,A,Y)_i:A_i=k} S(y_i-0.5)} \\ ms_{FNR}(A=k) &= \frac{\sum_{(X,A,Y)_i:A_i=k,Y_i=1} S(0.5-y_i)}{\sum_{(X,A,Y)_i:A_i=k} S(0.5-y_i)} \\ ms_{TPR}(A=k) &= \frac{\sum_{(X,A,Y)_i:A_i=k,Y_i=1} S(y_i-0.5)}{\sum_{(X,A,Y)_i:A_i=k} S(y_i-0.5)} \\ ms_{TNR}(A=k) &= \frac{\sum_{(X,A,Y)_i:A_i=k,Y_i=0} S(0.5-y_i)}{\sum_{(X,A,Y)_i:A_i=k} S(0.5-y_i)} \end{aligned} \tag{8.3}$$

In the same way sigmoided versions of the measures can be derived for the other statistics.

Once measures are defined, a continuous approximation for BPS fairness for both the continuous and the sigmoided measures can be defined as

$$\begin{aligned} BPSc_s &= \frac{min(mc_s(A=1), mc_s(A=0))}{max(mc_s(A=1), mc_s(A=0))} \\ BPSs_s &= \frac{min(ms_s(A=1), ms_s(A=0))}{max(ms_s(A=1), ms_s(A=0))} \end{aligned} \tag{8.4}$$

These, in turn can be translated into loss functions that can be used during training by inverting them, and further expanded by allowing to weigh the importance of small biases versus large biases by raising the loss to the k^{th} power which depresses the importance of fairness losses close to 0 (i.e. when the system is almost completely fair).

$$\begin{aligned} LFc_{(s,k)} &= (1 - BPSc_s)^k \\ LFs_{(s,k)} &= (1 - BPSs_s)^k \end{aligned} \tag{8.5}$$

These loss functions are continuous and differentiable in all but one point, namely the point where numerator and denominator are equal and thus in the minimum of the loss function. This, however, can be easily addressed in the training algorithm when optimizing the overall loss function.

8.3.3 Fairness Regularization for Neural Network Training

The approach proposed in this dissertation is aimed at training Neural Network-based deep learning classifiers to obtain more fair results in decision support systems. The tasks we are addressing here are classification tasks with a goal of achieving high accuracy predictions that are also fair. The underlying task is thus maximizing accuracy which is commonly encoded in terms of a binary cross entropy loss function, LF_{BCE}.

Starting from this, we utilize the fairness loss functions derived in the previous sections as regularization terms resulting in an overall loss function, LFc and LFs for continuous and sigmoided fairness losses, respectively:

$$\begin{aligned} LFc(\vec{\alpha}, \vec{k}) &= LF_{BCE} \ + \ \textstyle\sum_{s_i} \alpha_i LFc_{(s_i,k_1)} \\ LFs(\vec{\alpha}, \vec{k}) &= LF_{BCE} \ + \ \textstyle\sum_{s_i} \alpha_i LFs_{(s_i,k_1)} \end{aligned} \tag{8.6}$$

where $\vec{\alpha}$ is a weight vector determining the contribution of each of the different fairness loss functions to the fairness regularization, $\vec{k}$ is a vector of powers to be used for each of the fairness losses, and s is the vector of the 4 loss metrics, $< FPR, FNR, TPR, TNR >$. Setting an α_i to 0 effectively removes the corresponding fairness criterion from the loss function.

These loss functions can be used to train a Neural Network classifier where different values for $\vec{\alpha}$, $\vec{k}$, and the choice of sigmoided vs continuous loss puts different emphasis on different aspects of the underlying fairness characteristics.

8.4 Experiments

To study the applicability of the proposed use of fairness losses as regularization terms, we conducted experiments on three differnt datasets, two in the the recidivism domain and one in income prediction domain, and analyzed the behavior and effects of different function and weight choices.

8.4.1 Datasets

D1: Dataset 1 is the main dataset used in this study and has been previously used in our experiments in Chapter 4. It is the raw data from the study"Criminal Recidivism in a Large Cohort of Offenders Released from Prison in Florida, 2004-2008 (ICPSR 27781)" [58]. This dataset is described in Chapter 3 and available in the ICPSR repository. [1] It is based on the information provided by Florida Department of Corrections (FDOC) and the Florida Department of Law Enforcement (FDLE) [59]. It is comprised of 156,702 records distributed in a 41:59 ratio of recidivists to non-recidivists records. This ratio of our two underlying subpopulations is 34:66 for Caucasians and 46:54 for African Americans which makes it very unbalanced between these groups and leads to significant bias in traditional recidivism prediction approaches. In each crime category, the dataset has a higher proportion of non-recidivists Caucasians than African Americans. The dataset covers six crime categories and provides a large range of demographic features, including crime committed, age, time served, gender, etc. We employed one-hot encoding to treat categorical features and trained our system to predict whether an offender would be reconvicted within the next 3 years.

D2: Dataset 2 is a secondary dataset described in Chapter 3 and has been used in Chapters 5-7 in this dissertation. In this chapter we mainly utilized the less balanced Dataset 1 but used Dataset 2 for this study to validate results from Dataset 1. This dataset ensued from the "Recidivism of Prisoners Released in 1994" study [60]. It contains data from 38,624 offenders that were released in 1994 from one of 15 states in the USA. This is a very comprehensive dataset in the recidivism domain that contains up to 99 pre and post 1994 criminal history records , treatments and courses taken by offenders while they were still in prison. As described in Chapter 6

[1] `https://www.icpsr.umich.edu/icpsrweb/NACJD/studies/27781`

and 7 (and published in [35, 57]), each of the 38,624 records was split to create one record per arrest cycle and treat each arrest cycle as a point in time for a parole decision to be made. Any subsequent relapse into criminal ways was used to label records for recidivism. This process resulted in approximately 442,000 records that use demographic and history information to predict reconviction. This dataset is significantly more balanced between the two underlying subpopulations, allowing us to verify effects identified in Dataset 1 in a dataset with different characteristics.

D3: Dataset 3 is the UCI Adult dataset or "Census Income" dataset [62] described in Chapter 3, a dataset extracted from the 1994 census data includes input variables such as age, occupation, education, race, sex, marital-status, native-country, hours-per-week etc. and indicates income exceeding $50K/year or not. This dataset is also a demographic dataset and is used to evaluate bias mitigation techniques for gender-based inequities.

8.5 Performance Evaluation using Recidivism Data

8.5.1 Constructing Neural Networks

For the two recidivism datasets and all settings of the fairness regularization, we trained neural networks with 2 hidden layers with 41 units in each of them. Each input and hidden layer was followed by ReLU activation function and dropout [80] with 10% probability. This was followed by Batch Normalization of each layer. Batch Normalization makes the optimization smoother [81]. The output layer had 1 unit which used a logistic output function as indicated previously. We tuned various hyper parameters to select a batch size of 256 and 100 epochs and trained our neural network model using the Adam [64] optimization algorithm for stochastic gradient descent. While training each parametrized model, we saved it only if its accuracy

improved over the one created in the previous epoch iteration. The size, structure, and hyperparameter settings of the network was chosen based on experience in the work presented in the previous chapters. another work with these datasets.

8.5.2 Evaluation Study

To assess the operation of and evaluate the characteristics of the proposed fairness loss regularization in the context of the recidivism datasets with race-based fairness criteria, we conducted experiments with 6 setting for the measures used in the fairness for both continuous and sigmoided loss functions, employed 4 different exponents for the continuous case, and ran experiments for 10 different weight settings for the degree of influence of the fairness regularization. In particular, we chose settings that used each of the 4 statistics individually as well as ones that used FPR and FNR or TPR and TNR simultaneously with equal weights. Weights α were varied between 0.1 and 1 in steps of 0.1, and for the continuous loss functions, powers of 1, 2, 3, and 4 used to train networks. The same experiments, except only using a power of 1 was repeated with sigmoided loss functions. The goal here was to be able to compare the effects of different settings on the behavior of the system both in terms of accuracy achieved of the final model, the resulting fairness, and the stability of the solution. The Baseline model used only BCE loss and no fairness regularization.

To capture the impact of training variance on these models, Monte Carlos cross validation with 10 iterations was conducted for each of the setting and the means of the individual metrics for the resulting models are reported here to increase reliability of any conclusions drawn and reduce the chance of outliers. In addition, stability of the resulting solutions was evaluated using the variance across the 10 training runs. To study the effects, both BPS and loss function values were recorded and evaluated.

8.5.3 Results and Discussion

The goal of the set of experiments conducted here is to evaluate whether the proposed approach to active fairness training in deep learning systems through fairness regularization can achieve the desired goal and to evaluate the effect of different loss function and regularization weight choices on the performance both in terms of accuracy and fairness. To perform this study in the context of recidivism we utilized mainly Dataset 1 due to its higher imbalance and studied the effect of different aspects of the loss function design in Sections 8.5.3.1, 8.5.3.2, and 8.5.3.3. We then used Dataset 2 to validate some of the results on a more balanced dataset in Section 8.5.3.4.

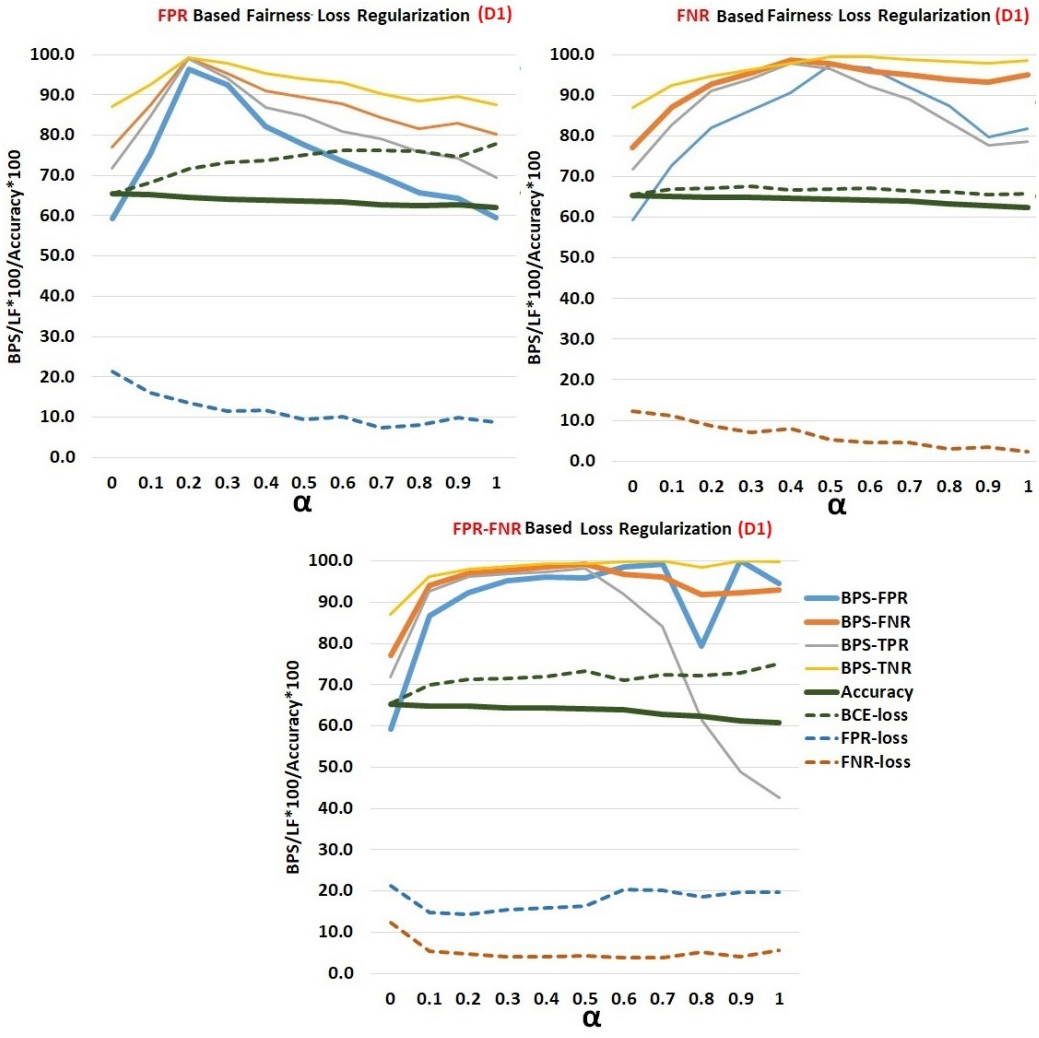
Measuring Bias in Results: After conducting the experiments, we investigated the effect of the loss functions by studying prediction results' characteristics for residual bias and measured these by computing the Bias Parity Score for the four main statistical measures, FPR, FNR, TPR, and TNR, as well as Bias Parity Score for Accuracy. In addition we recorded Accuracy as well as the values for BCE loss and for each of the fairness loss functions used in the respective experiment. This allowed us to analyze effects of loss function choice and parameter settings on performance both in terms of the internal operation of the approach (i.e. loss functions) and of the desired performance metrics (Accuracy and BPS scores).

8.5.3.1 Applicability and Effect of Regularization Weight

To study the basic performance of the system as well as the effect on fairness and accuracy, a set of experiments were conducted for each of the six continuous fairness measures in the linear case (i.e. with a power of 1). In these experiments, the regularization weight was increased step-wise, starting from the Baseline with no regularization ($\alpha = 0$) and moving to equal contributions of BCE loss and regularization ($\alpha = 1$). Figure 8.1 shows the average BPS scores, accuracy result, and loss

function values as a function of the regularization weight, α. Values for Accuracy and Loss functions are multiplied by 100 to be in the same scale as BPS scores. In this figure, results for $LFc_{(FPR,1)}$, $LFc_{(FNR,1)}$, and $LFc_{(FNR,1)} + LFc_{(FPR,1)}$ experiments are shown. Behavior for $LFc_{(TPR,1)}$, $LFc_{(TNR,1)}$, and $LFc_{(TPR,1)} + LFc_{(TNR,1)}$ was similar.

Figure 8.1: BPS Measures, Accuracy, and Loss function values as a function of regularization weight, α, for $LFc_{(FPR,1)}$ regularization (top-left), $LFc_{(FNR,1)}$ regularization (top-right), and $LFc_{(FPR,1)} + LFc_{(FNR,1)}$ regularization (bottom).



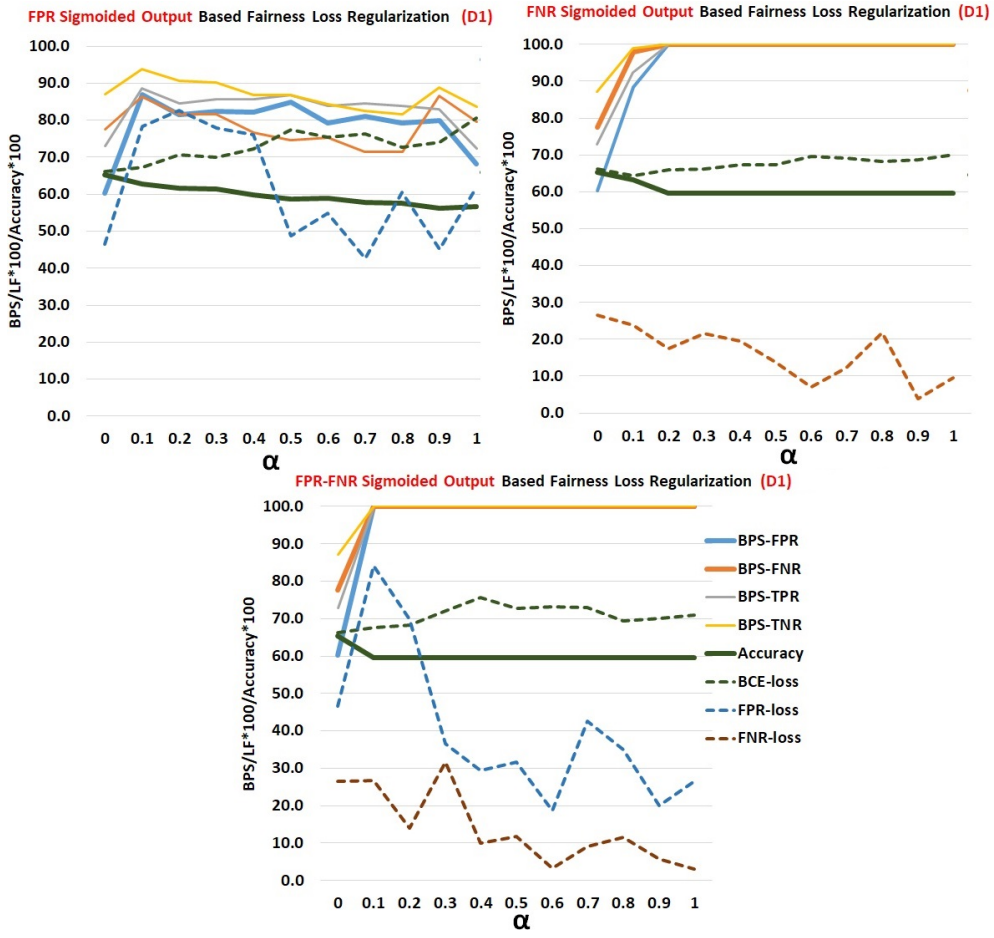
In all of the cases, these graphs show that introducing the fairness regularization loss immediately leads to an increase in the corresponding BPS score, and thus the corresponding fairness measure, while only very gradually decreasing accuracy of the prediction model as the influence of the fairness loss increases. This demonstrates the viability of the proposed technique to obtain more fair prediction models for decision support systems without destroying task performance. Moreover, in this case, regularization for one statistic also yields improvement in the other statistics, at least initially, which can be explained with the close relation of FPR, FNR, TPR, and TNR in the context of accuracy as a performance metric.

However, the experiments with the three different loss functions also show some important differences that give important information regarding important considerations when determining regularization weights. In particular, while the experiments with $LFc_{(FNR,1)}$ and $LFc_{(FNR,1)} + LFc_{(FPR,1)}$ show a relatively steady increase in fairness as the regularization weight is increased, the case of $LFc_{(FPR,1)}$ shows that after an initial strong increase in fairness, the fairness measure encoded in the regularization loss function, $BPSc_{FPR}$, starts to decrease and finally collapse once the regularization weight, α, exceeds 0.2. At the same time it can be observed that the regularization loss, $LFc_{(FPR,1)}$, continues to decrease, showing a decoupling between the core fairness measure and the loss function at this point. The reason for this discrepancy is that in order to obtain a differentiable loss function it was necessary to interpret the network output as a probabilistic prediction and thus the loss function can be improved by changing the output y_i from 0.4 to 0.3 which has no effect on the actual fairness measure. Decreasing the output for one negative item from 0.4 to 0.1 while simultaneously decreasing an output for a positive data item from 0.51 to 0.49 here represent an improvement in loss function value while reducing the corresponding fairness BPS as the second item classification became incorrect.

8.5.3.2 Sigmoided Loss Function

One way to address the decoupling of the loss function from the fairness measure is the use of the proposed sigmoided fairness loss function which aggressively moves metric values for the loss function towards 0 and 1. Figure 8.2 shows the results for the corresponding cases to the previous section but using the sigmoided version of the fairness loss function for $LFs_{(FNR,1)}$, $LFs_{(FPR,1)}$, and $LFs_{(FNR,1)} + LFs_{(FPR,1)}$. Again BPS values, Accuracy, and loss function values are shown.

Figure 8.2: BPS Measures, Accuracy, and Loss function values as a function of regularization weight, α, for sigmoided loss functions $LFs_{(FPR,1)}$ regularization (top-left), $LFs_{(FNR,1)}$ regularization (top-right), and $LFs_{(FPR,1)} + LFs_{(FNR,1)}$ regularization (bottom).



While these results when compared to the linear results in Figure 8.1 as expected show less of a decoupling between loss function and fairness (except between the baseline and the first result with regularization weight of 0.1), tend to impose higher levels of fairness earlier, and maintain fairness more reliably throughout the full range of regularization weights, the results also show a stronger degradation in accuracy and, when looking at the regularization loss function for higher weights also exhibit strong signs of instabilities, reflected in significantly increased variances across the 10 runs used in these results. This implies that while there are advantages in terms of how well the sigmoided loss function represents fairness, optimizing these regularization functions is significantly harder for the algorithm due to much steeper gradients, leading to less stable convergence. When choosing between these two options it is thus important to consider this tradeoff and to have ways to monitor gradient stability.

8.5.3.3 Effect of Loss Function Power

Another way to modify the effect of the regularization losses is to increase the power of the loss function. Raising the power will reduce the impact of small amounts of bias near a fair solution while increasing importance of the fairness measure if fairness losses are high. The goal here is to reduce small changes near the solution and thus to reduce the small scale adjustments most responsible for divergence between fairness and corresponding loss function. Figure 8.3 shows the effect of different powers for the continuous loss function $LFc_{(FPR,k)}$ for powers k or 1, 2, 3, and 4. This was the example in the initial, linear experiments where a strong decoupling between loss function and fairness metric occurred. This pattern can also be observed with Dataset 3 in 8.5 and 8.6 and discussed in more detail in Section 8.6 below.

These graphs show that as the power of the loss function increases, the improvement in fairness becomes smoother and the loss function and fairness measure diverge

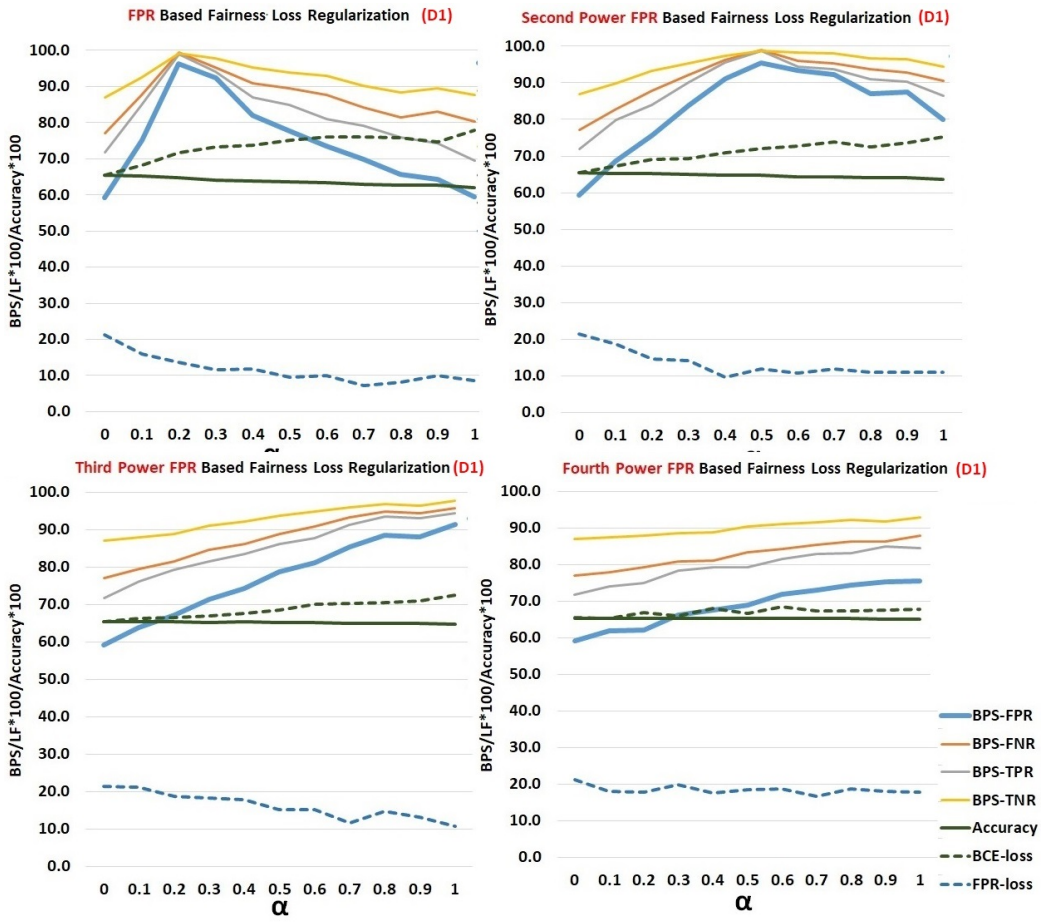


Figure 8.3: BPS Measures, Accuracy, and Loss function values as a function of regularization weight, α, for different powers of the continuous loss function functions $LFc_{(FPR,1)}$ regularization (top-left), $LFc_{(FPR,2)}$ regularization (top-right), $LFc_{(FPR,3)}$ regularization (bottom-left), and $LFc_{(FPR,4)}$ regularization (bottom-right).

less and significantly later. However, the graphs also show that a significantly larger regularization weight is required to optimize the function with a higher power loss function, with the power 4 experiment not reaching the best fairness even at a weight of $\alpha = 1$. While this is not a problem in this case as the fairness loss function is relatively low, it might become a problem in datasets where fairness loss is inherently larger, thus over-emphasizing the effect of fairness. The best performance here seems to be achieved with power 3.

8.5.3.4 Effects in More Balanced Data

Dataset 1, as described before, contains relatively biased data, reflected in the initially relatively low fairness scores in the range of 70 - 80. Also, this dataset only contains relatively basic attributes, thus limiting achievable accuracy to around 65%. Due to this, this dataset offers relatively significant room for improvement in fairness without significant deterioration in accuracy. To see if similar benefits can be achieved in the context of a less biased and significantly richer dataset, the experiments were repeated on the Dataset 2. In this dataset initial fairness in the baseline case is closer to 90 and accuracy reaches 89%. To see how similar loss function settings work in this case, Figure 8.4 shows the results for the best settings for Dataset 2 for both continuous and sigmoided loss functions for the FPR -based reguarization. In particular, it shows the continuous case with power 3, $LFs_{(FPR,3)}$ and the basic sigmoided case of $LFs_{(FPR,1)}$.

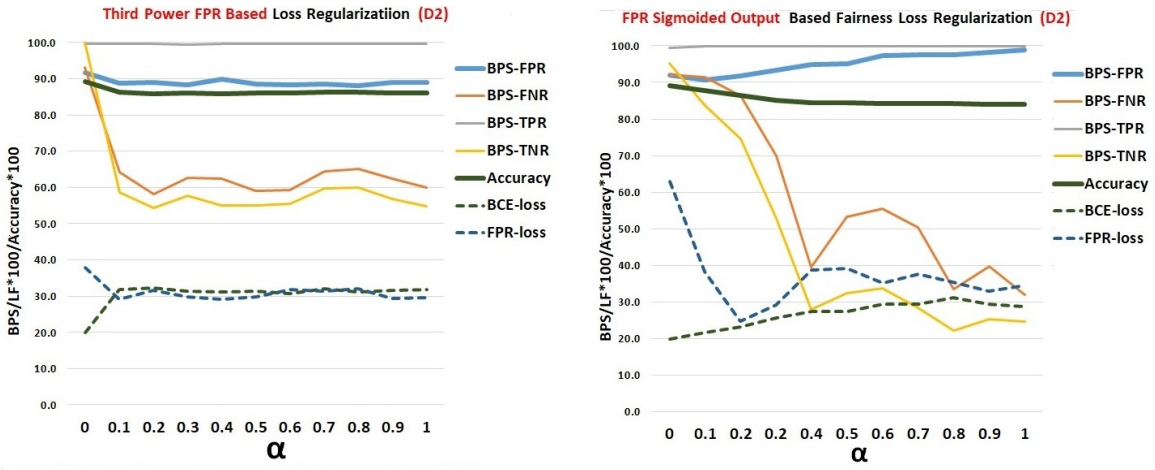


Figure 8.4: BPS Measures, Accuracy, and Loss function values as a function of regularization weight, α, for Dataset 2 with $LFc_{(FPR,3)}$ regularization (left), and sigmoided $LFs_{(FPR,1)}$ regularization (right).

These results show that even for the less biased dataset where improvement in fairness is more difficult to achieve, the proposed regularization method is successful at increasing the desired fairness characteristic without a dramatic drop in accuracy. However, while in the first dataset an increase in fairness in one metric tended to go hand in hand with increases in fairness in the other metrics even though they were not explicitly used as part of the training, this effect does not occur in Dataset 2. In order to improve fairness in one metric in this less biased dataset, application of the corresponding loss function, while leading to improvement in that fairness metric yields degradations in others. This is not entirely unexpected here as the smaller amount of improvement potential is bound to require more tradeoffs.

8.6 Performance Evaluation using Dataset 3

To demonstrate that the applicability and results for the proposed fairness approach extend beyond the recidivism domain and race-based bias, as well as to be able to compare overall performance with other state of the art approaches, we applied our approach on Dataset 3 and compared the results with those published in [44, 44, 82, 34, 83]. To do these comparisons we again evaluated effects of different settings in the loss functions but also experimented with multiple Neural Network architectures and certain data pre-processing methods, including data normalization.

8.6.1 Neural Network Architectures

To determine the best network architecture we experimented with a number of hyperparameters, varying the number of layers, the number of units in each layer, the activation functions used in the hidden layers, and the drop out rate to find the architecture with the highest accuracy. As a result of this, we developed two architectures that were used in the experiments.

Architecture 1: The first architecture was directly derived from the ones used for recidivism datasets and was comprised of two hidden layers, each with ReLU activation function and 108 neurons.

Architecture 2: The second architecture was modified based on experiences with the Dataset 3 and was comprised of two hidden layers, each with leaky ReLU activation function but 108 and 324 neurons respectively.

The main considerations in the design of the second architecture was the achieved accuracy when trained without any fairness considerations, representing a slight improvement in that condition from the baseline accuracy of 84.5% for Architecture 1 to a baseline accuracy of 84.747% for Architecture 2.

8.6.2 Results and Comparison

8.6.2.1 Effect of Loss Function Parameters

Varying loss function weights, continuous vs sigmoided loss functions, and loss function powers in the context of the Dataset 3 yielded very similar observations as in the case of the recidivism datasets as shown by the graphs in Figure 8.5 for Architecture 1 and Figure 8.6 for Architecture 2.

The figures here show the results for different powers (1, 2, 3, and 4) using positivity rate ($P(C = 1|X)$) as the underlying measure which, as a BPS score corresponds to a BPS version of the Statistical Parity fairness measure. As in previous experiments, increasing the power reduced decoupling between the loss function and the BPS score and led to smoother improvements. Additional studies with sigmoided loss function on Dataset 3 further demonstrated the same behavior as for the recidivism datasets, demonstrating the generality of the proposed framework. In this case, however, power 4 on continuous loss functions seemed to achieve the best results.

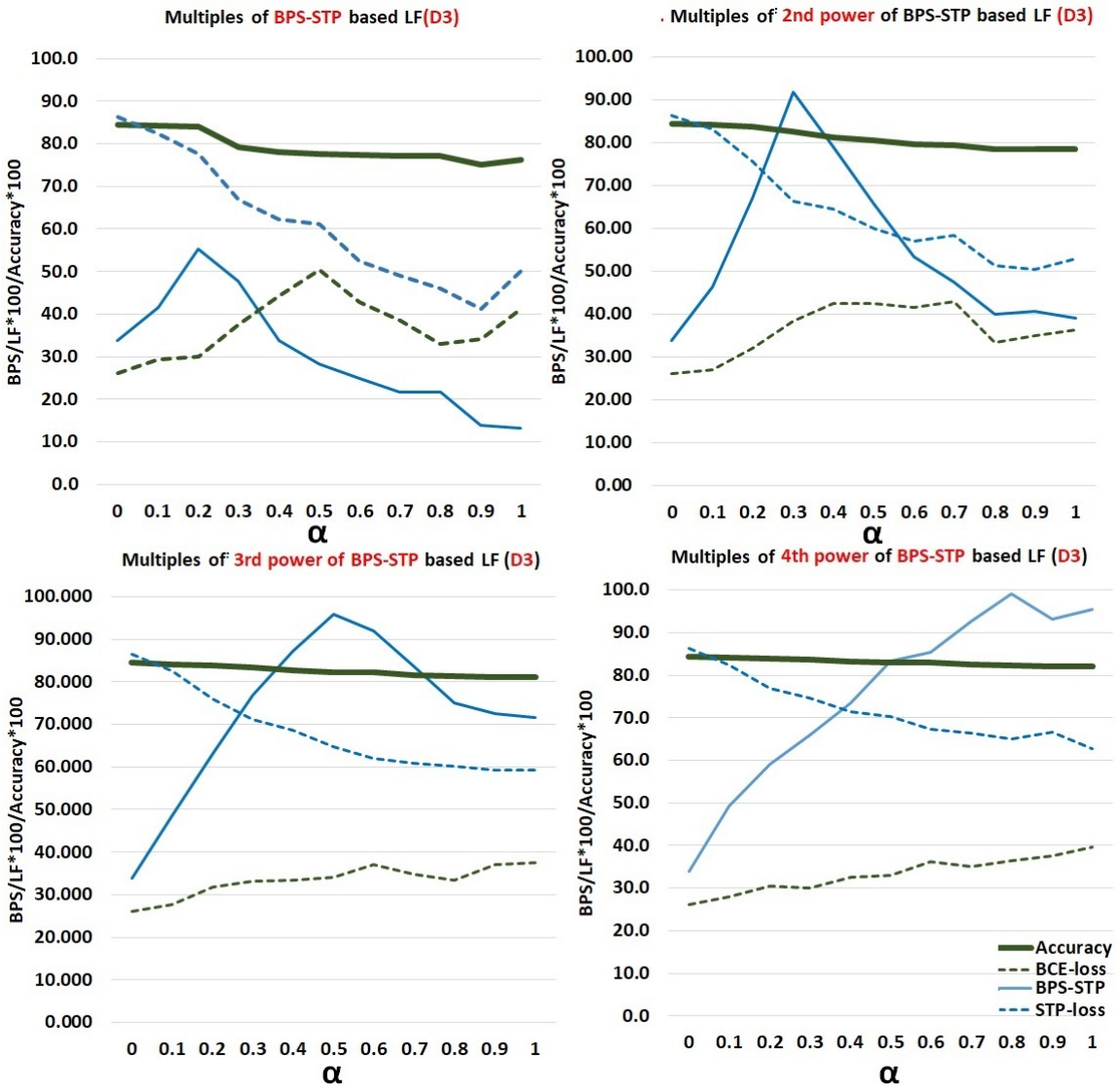


Figure 8.5: BPS Measures, Accuracy, and Loss function values as a function of regularization weight, α, for $LFc_{(STP,1)}$ regularization (top-left), $LFc_{(STP,2)}$ regularization (top-right), $LFc_{(STP,3)}$ regularization (bottom-left), and $LFc_{(STP,4)}$ regularization (bottom-right). Architecture 1.
Architecture 1: ReLU Activation Fn. in the 2 hidden layers having 108 neurons each.

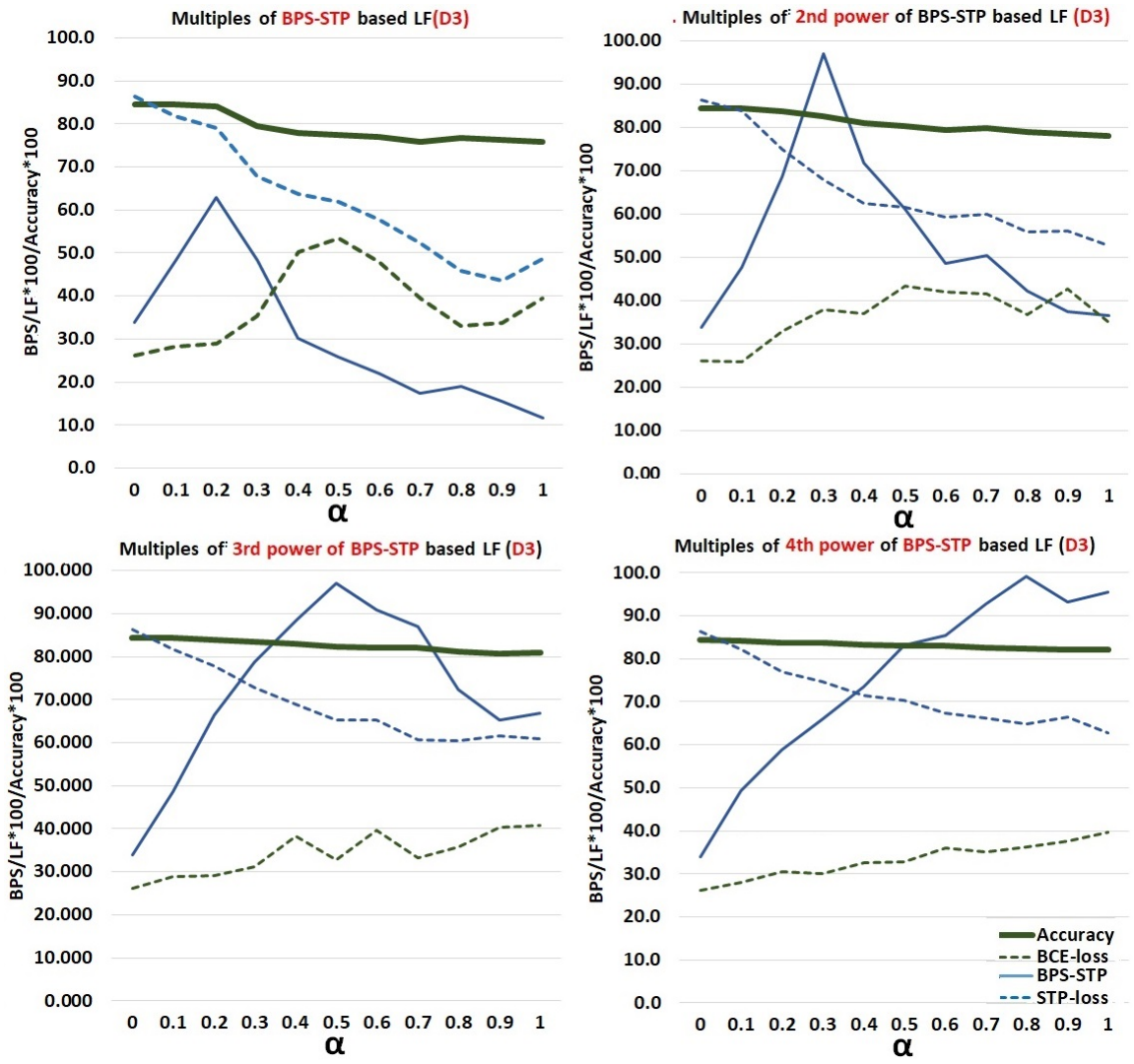


Figure 8.6: BPS Measures, Accuracy, and Loss function values as a function of regularization weight, α, for $LFc_{(STP,1)}$ regularization (top-left), $LFc_{(STP,2)}$ regularization (top-right), $LFc_{(STP,3)}$ regularization (bottom-left), and $LFc_{(STP,4)}$ regularization ((bottom-right)) Architecture 2: leaky ReLU Activation Fn. in the 2 hidden layers having 108 and 318 neurons respectively.

8.6.2.2 Effect of Architectures on D3 Dataset

. We used two architectures for our experiments: Architecture 1 and Architecture 2. These were selected after tuning hyperparameters like number of neurons, number of epochs, weights associated with the loss function and the loss function itself. Architecture 1, was comprised of two hidden layers, each with ReLU activation function and 108 neurons. Architecture 2, was comprised of two hidden layers, each with leaky ReLU activation function but 108 and 324 neurons. Both architectures produced very similar patterns as were shown with Dataset 1 and Dataset 2 too. However, Architecture 2 improved BPS scores further as shown in 8.5 and 8.6. In particular, Architecture 2 achieved slightly higher accuracy at the same level of BPS score and was thus chosen for comparison experiments.

8.6.2.3 Comparison with State of The Art Results

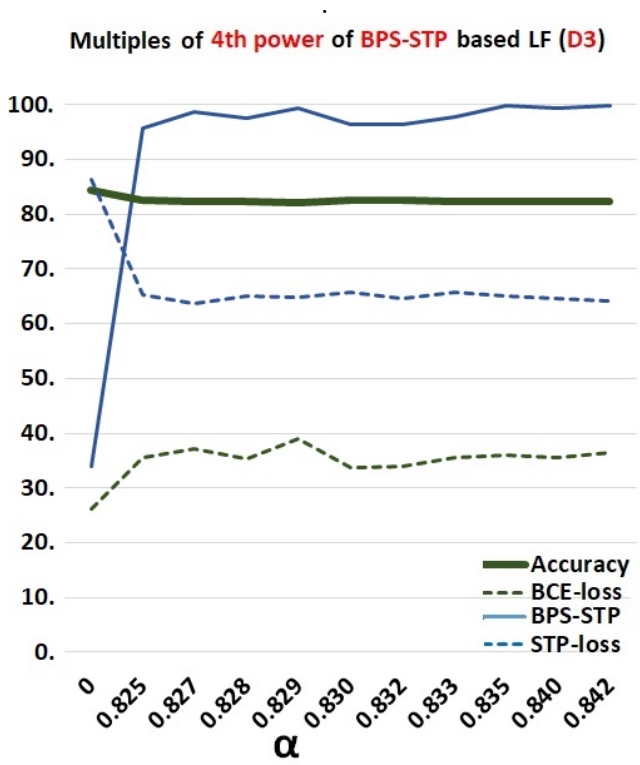
To compare the approach proposed here to previous work, we compared to the recent results to Zhang's work that optimized p-rule (i.e. statistical parity) [84], and the work by Krasanakis [44] that used adversarial training to achieve results with closely matching FPR and FNR across genders.

As [84] used p-rule as the fairness criterion, we utilized BPS on prediction rate to incorporate corresponding loss functions into our approach. As observed in the previous experiments on recidivism, applying BPS-based fairness loss in Dataset 3 again changed the overall accuracy only to small degree with an imposition of a very high degree of fairness loss to compare against prior p-rule work [84], yielding a drop from 84.5% to 82.3% as shown in Table 8.1 using Architecture 1, but while BPS-Acc improved from 86.6 to 89 and pRule changed from 33.9% to 99.128%. By further changing to Architecture 2 the accuracy could be further improved as shown

in Figure 8.6. In this case, Architecture 2 with a power of 4 seemed to perform the best with weights around a value of 0.8 and thus a finer scaled grid search was performed for α values between 0.825 and 0.842 as shown in Figure 8.7 to find the best results at $\alpha = 0.84$ with a BPS_{STP} (pRule) of 99.9% and an accuracy of 83%.

Thus our approach offered higher accuracy than [84] while maintaining a pRule of approximately 100%.

Figure 8.7: BPS Measures, Accuracy, and Loss function values as a function of regularization weight, α, for $LFc_{(STP,4)}$ regularization. Architecture 2,POW=4,α=0.84 Architecture 2: leaky ReLU Activation Fn. in the 2 hidden layers with 108 & 324 neurons respectively. Accuracy 82.618% pRule(BPS-STP) 99.858%



The comparison of our results with both Architecture 1 and Architecture 2 to previous work with pRule-based fairness is shown in Table 8.1, illustrating that the

proposed approach can outperform the more specialized approaches, including ones used specifically for the pRule.

Table 8.1: Adult Income dataset disparate impact elimination for BPS-FPR-FNR-based Loss Function. Accuracy and pRule Comparison with published results of other techniques.
Architecture 1 values with STP-Loss Function, POW=4, α=0.8
Architecture 2 values with STP-Loss Function,POW=4, α=0.84

Fairness Technique	Adult Income	
	pRule	acc
Krasanakis et.al. None 2018 [44]	27%	85%
Krasanakis et.al. 2018 [44]	100%	82%
Zafar et.al. 2017 [34]	94%	82%
kamishima et. al. 2012 [83]	85%	83%
None (Neural Networks Current work)	34%	85%
Architecture1	99%	82%
Architecture 2	**100**%	**83**%

To compare to the work in [44] who are aimed at achieving similar FNR and FPR values for both genders, we utilized a combined fairness loss function that employed BPS_{FNR} and BPS_{FPR} together in the training phase. The FPR and FNR for adult income dataset using our technique as shown in Table 8.2 were approximately equal across the two gender based groups at 0.0589 versus 0.0628 and at 0.4431 versus 0.5105, respectively. BPS_{FPR} and BPS_{FNR} for our results was a 94 and 87 respectively with Architecture 1, where a BPS of 100 means no unfairness while a BPS of 0 means complete unfairness. With Architecture 2 , we could further improve BPS_{FPR} and BPS_{FNR} while maintaining low absolute values of FPR and FNR of the two gender based cohorts and thus outperformed the BPS_{FPR} and BPS_{FNR} in [44] as shown in Table 8.2.

Table 8.2: Adult Income dataset: False Positive Rate (FPR) and False Negative Rate (FNR) for income bracket prediction for the two gender based groups, with and without debiasing.
Architecture 1:ReLU Activation Fn. in the 2 hidden layers having 108 neurons each.
Architecture 2:leaky ReLU Activation Fn. in the 2 hidden layers having 108 and 318 neurons respectively.
Architecture 1 values with FPR-FNR-Sigmoided-LF,POW=4,α1=0.05, α2=0.05
Architecture 2 values with FPR-FNR-Sigmoided-LF,POW=3,α1=0.1, α2=0.125

		Female		Male		BPS
		Without	With	Without	With	
Beutel et al. [82]	FPR	0.1875	0.0308	0.1200	0.1778	17.3
	FNR	0.0651	0.0822	0.1828	0.1520	52.6
Zhang et al. [84]	FPR	0.0248	0.0647	0.0917	0.0701	92.0
	FNR	0.4492	**0.4458**	0.3667	**0.4349**	97.5
Architecture 1	FPR	0.0319	0.0589	0.1203	**0.0628**	93.9
	FNR	0.4098	**0.4431**	0.3739	0.5105	86.8
Architecture 2.	FPR	0.0319	0.0610	0.1203	0.0708	**93.3**
	FNR	0.4098	0.4785	0.3739	0.4862	**98.4**

8.7 Conclusions

In this work we have proposed elements and considerations to impose fairness on Neural Networks during the training phase. In particular we proposed to translate Bias Parity Score-based fairness metrics into corresponding loss functions that can then be used as regularization terms during training to actively achieve improved fairness and reduced bias between subpopulations in the data. For this, we introduced a family of derived fairness loss functions and conducted experiments on recidivism prediction data where we investigated different regularization weights and fairness loss function settings that are added to the task function which represents accuracy through a binary cross entropy loss function. In these experiments we demonstrated how to use the loss functions to bring measurable improvement in equity to predictions and hence to the cohorts involved. In contrast to some of the other previous works, this work does not depend on changing input or output labels to make fair

recommendations while simultaneously not forsaking accuracy. By building our loss functions and yet dropping the sensitive attribute information from the input feature vector of the neural networks model, we ensure that the this work is also guided by fairness by unawareness guidelines.

This work illustrates that by concurrently using one or more BPS measure-based loss functions in concurrence with binary cross entropy can design automated decision support systems that can optimize for social objectives such as fairness. However, this work also shows that a correct pick of the regularization weight and the fairness loss function form can be essential to address convergence stability and to address specifics challenges with a dataset, such as different levels of bias in the data and varying levels of improvement potential due to limitations in available attributes or initial performance. This work also shows that besides a correct weight, choice of power of the loss function or the use of a sigmoided loss introduces both benefits and challenges, where use of sigmoided losses, for example, can decrease the loss more rapidly and avoid divergence between fairness metric and loss function, but can also destabilize the convergence.

8.8 Future Directions

In this chapter we presented a family of loss functions and showed potential benefits and detriments of different choices. In future work we would like to investigate methods that could automatically adjust fairness regularization terms according to properties of the dataset used.

Beyond Neural Networks: There are many machine learning algorithms suitable for numerous applications in different domains and there is much room in those to increase the fairness in the model itself.

Sensitive Attribute Disjointedness: Not unlike most of the fairness literature, we assume that the race attributes are accurate. There is an assumption that there is a disjointedness in the race attribute, however demographic groups intersect. There needs to be research on such a racial intersection.

CHAPTER 9

Conclusion

9.1 Summary of Contributions

This dissertation uses three socially relevant datasets from two different domains, namely recidivism and income prediction. Two of these are Department of Justice (DOJ) datasets while one is a Census-based adult income dataset. One of the DOJ datasets and the adult income dataset hold mostly demographic information. The second DOJ dataset has information related to offenders' criminal history, substance-abuse, and treatments taken during incarceration. In Chapter 4 and Chapter 8, we work with two datasets with largely demographical features. Our prediction results indicated presence of race-based and gender-based bias respectively in them that are much in-line with the published literature on similar datasets. In Chapter 5 we demonstrate our work with a personal history based dataset and demonstrate that such a dataset produces prediction results with much less bias than the result based on a demographical feature based dataset.

In Chapter 6, we demonstrate an approach that uses personal data such as criminal history along with treatment and education activities availed during incarceration to further enrich the dataset before it is used for making predictions. In our approach, we demonstrate how to preprocess the dataset and derive additional features from the given data. This is followed by selecting models created using data with 0, 10, 20, 40, 60, 80, and 100 past arrest cycles, computing the values of various statistical measures for each model and using the ratio of FPRs for the two race-

based cohorts to find the model having the FPRs as close to parity as possible. This approach improved both accuracy and fairness.

In Chapter 7, we improve our techniques to derive additional features from the temporal information in the data. Rolling sums of 26 past individual crimes committed in the past 0, 1, 3, 5, 7, 10, 15, 20, 40, 60, 80, and 100 arrest cycles further improve fairness and accuracy of predictions. The dataset with 5 past arrest cycle history length produced the fairest predictions. We introduce a new fairness measure called Bias Parity Score (BPS) that summarizes bias in a single measure and facilitates model selection for improved fairness and accuracy. We illustrate its usage and application. BPS is a versatile measure that divides the values of the same statistical measure in two cohorts, multiplies it by 100, and represents the relative bias as a percentage with 100 indicating no bias and 0 indicating full bias. This in turn offers a very easy way for a layperson and experts alike to select most appropriate model.

In Chapter 8, we propose a group of BPS inspired loss functions and utilize them as regularization component in neural network to reign in race based bias in the recidivism datasets and gender based bias in the census-based adult income dataset. We analyze the effects of loss function choice, and various loss function parameter settings on accuracy and fairness. We demonstrate that with pertinent sets of fairness loss functions, corresponding weights, and powers to which the BPS-inspired loss functions are subjected, we can reduce any specific kind of bias in a model even in unbalanced datasets. We establish that for recidivism Dataset 1 by using a combination of binary cross entropy, BPS_{FPR} and BPS_{FNR} based loss functions with different weights associated with each component, we could achieve a BPS of 93.3 and 98.4 for FPR and FNR with a minimal drop in accuracy. This is better than any to-date published results. We illustrate that by using a STP-based loss function

we could achieve a BPS_{STP} (pRule) of 99.9% and an accuracy of 83% for income dataset. Thus, this approach outperforms the more specialized ones used for pRule.

Thus, in this dissertation we identify bias in prediction results using traditional statistical measures, enrich datasets, and show that richer datasets with derived attributes can help produce results with improved fairness and accuracy. We then introduce a metric called BPS to measure bias in one number and then use BPS based loss functions to mitigate bias in the training phase of a neural network. We demonstrate that the prediction results using our approach outperform the ones published so far using the same datasets by other studies. Furthermore, our technique can be applied in datasets with a different sensitive attribute and from different domains

9.2 Future Work

Our work leads to many future directions. Here are some of them:

- Find ways to automate exploration of sources of bias in diverse domains and mitigate bias in these
- Find ways to completely automate the process of finding appropriate neural network architecture, requisite loss functions, and hyper-parameters to generate models that mitigate bias in different domains for different sensitive attributes
- Find ways to mitigate bias using other machine learning classifiers. Investigate how BPS can be applied in other algorithms
- Establish techniques to identify and mitigate bias as the underlying data undergoes content drift while working in a dynamic environment with data stream evolving over time
- Find techniques to adjust the hyper-parameters to the changing distribution of sensitive attributes using an evolving data stream where predictions need to be made in a live environment

- Use and create better datasets that go beyond ones with simple demographic features like attributes and establish frameworks that can automate feature extraction and preprocessing of data
- Use and create datasets that acknowledge the overlap of sensitive attributes. For example, our current datasets and related work assume disjointedness of race feature vector values. Demographic groups intersect and both the data and the requisite research needs to examine this racial intersection as well as the presence of multiple sensitive attributes

REFERENCES

[1] M. E. Ahsen, M. U. S. Ayvaci, and S. Raghunathan, "When algorithmic predictions use human-generated data: A bias-aware classification algorithm for breast cancer diagnosis," *Information Systems Research*, vol. 30, no. 1, pp. 97–116, 2019.

[2] H. He and K. Nawata, "The application of machine learning algorithms in predicting the borrower's default risk in online peer-to-peer lending," 2019.

[3] G. Suarez-Tangil, M. Edwards, C. Peersman, G. Stringhini, A. Rashid, and M. Whitty, "Automatically dismantling online dating fraud," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1128–1137, 2019.

[4] A. A. Mahmoud, T. A. Shawabkeh, W. A. Salameh, and I. Al Amro, "Performance predicting in hiring process and performance appraisals using machine learning," in *2019 10th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2019, pp. 110–115.

[5] J. E. Johndrow, K. Lum, *et al.*, "An algorithm for removing sensitive information: application to race-independent recidivism prediction," *The Annals of Applied Statistics*, vol. 13, no. 1, pp. 189–220, 2019.

[6] M. Baudry and C. Y. Robert, "A machine learning approach for individual claims reserving in insurance," *Applied Stochastic Models in Business and Industry*, vol. 35, no. 5, pp. 1127–1155, 2019.

[7] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *European conference on machine learning*. Springer, 2004, pp. 39–50.

[8] S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: From discrimination discovery to fairness-aware data mining," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 2125–2126.

[9] C. FitzGerald and S. Hurst, "Implicit bias in healthcare professionals: a systematic review," *BMC medical ethics*, vol. 18, no. 1, p. 19, 2017.

[10] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, vol. 4, no. 1, p. eaao5580, 2018.

[11] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 2018, pp. 1–7.

[12] A. S. Bhati, "Estimating the number of crimes averted by incapacitation: an information theoretic approach," *Journal of Quantitative Criminology*, vol. 23, no. 4, pp. 355–375, 2007.

[13] A. S. Bhati and A. R. Piquero, "Estimating the impact of incarceration on subsequent offending trajectories: Deterrent, criminogenic, or null effect," *J. Crim. L. & Criminology*, vol. 98, p. 207, 2007.

[14] Y. Zhang, L. Zhang, and M. S. Vaughn, "Indeterminate and determinate sentencing models: a state-specific analysis of their effects on recidivism," *Crime & Delinquency*, vol. 60, no. 5, pp. 693–715, 2014.

[15] J. Zeng, B. Ustun, and C. Rudin, "Interpretable classification models for recidivism prediction," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 180, no. 3, pp. 689–722, 2017.

[16] P. A. Langan and D. J. Levin, "Recidivism of prisoners released in 1994," *Fed. Sent. R.*, vol. 15, p. 58, 2002.

[17] M. Alper, M. R. Durose, and J. Markman, *2018 update on prisoner recidivism: A 9-year follow-up period (2005-2014)*. US Department of Justice, Office of Justice Programs, Bureau of Justice . . . , 2018.

[18] T. Ozkan, "Predicting recidivism through machine learning," Ph.D. dissertation, 2017.

[19] A. D. Tiedt and W. J. Sabol, "Sentence length and recidivism among prisoners released across 30 states in 2005: Accounting for individual histories and state clustering effects," *Justice Research and Policy*, vol. 16, no. 1, pp. 50–64, 2015.

[20] C. Rudin, "Please stop explaining black box models for high stakes decisions," *arXiv preprint arXiv:1811.10154*, 2018.

[21] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, p. 93, 2019.

[22] H.-R. Won, J.-S. Shim, and H. Ahn, "A recidivism prediction model based on xgboost considering asymmetric error costs," *Journal of Intelligence and Information Systems*, vol. 25, no. 1, pp. 127–137, 2019.

[23] M. R. Durose, A. D. Cooper, and H. N. Snyder, *Recidivism of prisoners released in 30 states in 2005: Patterns from 2005 to 2010*.

[24] H. Jung, S. Spjeldnes, and H. Yamatani, "Recidivism and survival time: Racial disparity among jail ex-inmates," *Social Work Research*, vol. 34, no. 3, pp. 181–189, 2010.

[25] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals," *And it's biased against blacks. ProPublica*, 2016.

[26] F. S. Taxman, A. Pattavina, M. S. Caudy, J. Byrne, and J. Durso, "The empirical basis for the rnr model with an updated rnr conceptual framework," in *Simulation strategies to reduce recidivism*. Springer, 2013, pp. 73–111.

[27] R. Berk, *Criminal Justice Forecasts of Risk: a Machine Learning Approach*. Springer-Verlag New York Inc, 2012.

[28] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[29] B. Jain, M. Huber, L. Fegaras, and R. A. Elmasri, "Singular race models: addressing bias and accuracy in predicting prisoner recidivism," in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 2019, pp. 599–607.

[30] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019, pp. 329–338.

[31] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[32] M. Miron, S. Tolan, E. Gómez, and C. Castillo, "Evaluating causes of algorithmic bias in juvenile criminal recidivism," *Artificial Intelligence and Law*, pp. 1–37, 2020.

[33] J. Jung, S. Goel, J. Skeem, *et al.*, "The limits of human predictions of recidivism," *Science advances*, vol. 6, no. 7, p. eaaz0652, 2020.

[34] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without

disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1171–1180.

[35] B. Jain, M. Huber, R. A. Elmasri, and L. Fegaras, "Reducing race-based bias and increasing recidivism prediction accuracy by using past criminal history details," in *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 2020, pp. 1–8.

[36] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[37] B. T. Luong, S. Ruggieri, and F. Turini, "k-nn as an implementation of situation testing for discrimination discovery and prevention," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 502–510.

[38] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013, pp. 325–333.

[39] A. Biswas, M. Kolczynska, S. Rantanen, and P. Rozenshtein, "The role of in-group bias and balanced data: A comparison of human and machine recidivism risk predictions," in *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*, 2020, pp. 97–104.

[40] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," *arXiv preprint arXiv:1810.08810*, 2018.

[41] I. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?" in *Advances in Neural Information Processing Systems*, 2018, pp. 3539–3550.

[42] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[43] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *arXiv preprint arXiv:1908.09635*, 2019.

[44] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos, and Y. Kompatsiaris, "Adaptive sensitive reweighting to mitigate bias in fairness-aware classification," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 853–862.

[45] D. Biddle, *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd., 2006.

[46] U. E. O. E. Commission, "Questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures," *Federal Register*, March 1979.

[47] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, *et al.*, "Bias in data-driven artificial intelligence systems—an introductory survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1356, 2020.

[48] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 2009, pp. 13–18.

[49] M. Feldman, "Computational fairness: Preventing machine-learned discrimination," Ph.D. dissertation, 2015.

[50] V. Iosifidis and E. Ntoutsi, "Fabboo-online fairness-aware learning under class imbalance," in *Discovery Science*, 2020.

[51] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, "Controlling attribute effect in linear regression," in *2013 IEEE 13th international conference on data mining*. IEEE, 2013, pp. 71–80.

[52] L. Oneto, M. Donini, and M. Pontil, "General fair empirical risk minimization," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[53] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[54] A. Noriega-Campero, M. A. Bakker, B. Garcia-Bulle, and A. Pentland, "Active fairness in algorithmic decision making," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 77–83.

[55] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Learning fair classifiers," *arXiv preprint arXiv:1507.05259*, vol. 1, no. 2, 2015.

[56] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," *arXiv preprint arXiv:1511.00830*, 2015.

[57] B. Jain, M. Huber, R. Elmasri, and L. Fegaras, "Using bias parity score to find feature-rich models with least relative bias," *Technologies*, vol. 8, no. 4, p. 68, 2020.

[58] United States Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, "Criminal recidivism in a large cohort of offenders released from prison in florida, 2004-2008," Available at https://www.icpsr.umich.edu/icpsrweb/NACJD/studies/27781/version/1/variables (2010/07/29), 2010.

[59] A. Bhati and C. G. Roman, "Evaluating and quantifying the specific deterrent effects of dna databases," *Evaluation review*, vol. 38, no. 1, pp. 68–93, 2014.

[60] United States Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, "Recidivism of prisoners released in 1994," Available at https://www.icpsr.umich.edu/icpsrweb/NACJD/studies/3355/variables (2014/12/05), 2014.

[61] R. Kohavi *et al.*, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid." in *Kdd*, vol. 96, 1996, pp. 202–207.

[62] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[63] J. Brownlee, *Deep learning with python: Develop deep learning models on theano and tensorflow using keras*. Machine Learning Mastery, 2016.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[65] A. K. Jain, J. Mao, and K. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, no. 3, pp. 31–44, 1996.

[66] A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[67] R. E. Schapire, "Explaining adaboost," in *Empirical inference*. Springer, 2013, pp. 37–52.

[68] K. L. Priddy and P. E. Keller, *Artificial neural networks: an introduction*. SPIE press, 2005, vol. 68.

[69] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller, "The case for process fairness in learning: Feature selection for fair decision making," in *NIPS Symposium on Machine Learning and the Law*, vol. 1, 2016, p. 2.

[70] Q.-S. Xu and Y.-Z. Liang, "Monte carlo cross validation," *Chemometrics and Intelligent Laboratory Systems*, vol. 56, no. 1, pp. 1–11, 2001.

[71] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 2017, pp. 797–806.

[72] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[73] I. Zliobaite, "On the relation between accuracy and fairness in binary classification," *arXiv preprint arXiv:1505.05723*, 2015.

[74] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, p. 0049124118782533, 2018.

[75] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[76] A. Dal Pozzolo, O. Caelen, Y.-A. Le Borgne, S. Waterschoot, and G. Bontempi, "Learned lessons in credit card fraud detection from a practitioner perspective," *Expert systems with applications*, vol. 41, no. 10, pp. 4915–4928, 2014.

[77] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk, "Potential biases in machine learning algorithms using electronic health record data," *JAMA internal medicine*, vol. 178, no. 11, pp. 1544–1547, 2018.

[78] C.-L. Huang, M.-C. Chen, and C.-J. Wang, "Credit scoring with a data mining approach based on support vector machines," *Expert systems with applications*, vol. 33, no. 4, pp. 847–856, 2007.

[79] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 702–712.

[80] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[81] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" in *Advances in neural information processing systems*, 2018, pp. 2483–2493.

[82] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, "Data decisions and theoretical implications when adversarially learning fair representations," *arXiv preprint arXiv:1707.00075*, 2017.

[83] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 35–50.

[84] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

BIOGRAPHICAL STATEMENT

Bhanu Jain lives in the United States of America. She received her B.S. degree from I.E.T, Lucknow, India and her M.S. and Ph.D. degrees from The University of Texas at Arlington. Her current research interests are in Machine Learning, Deep Learning, Bias in prediction results and Accuracy of predictions. Some domains of her current and past interests as well as work are fairness in machine learning based predictions, trash management(reduce-reuse-recycle), education (expediting learning and increasing retention), and achieving human potential through kindness.

.