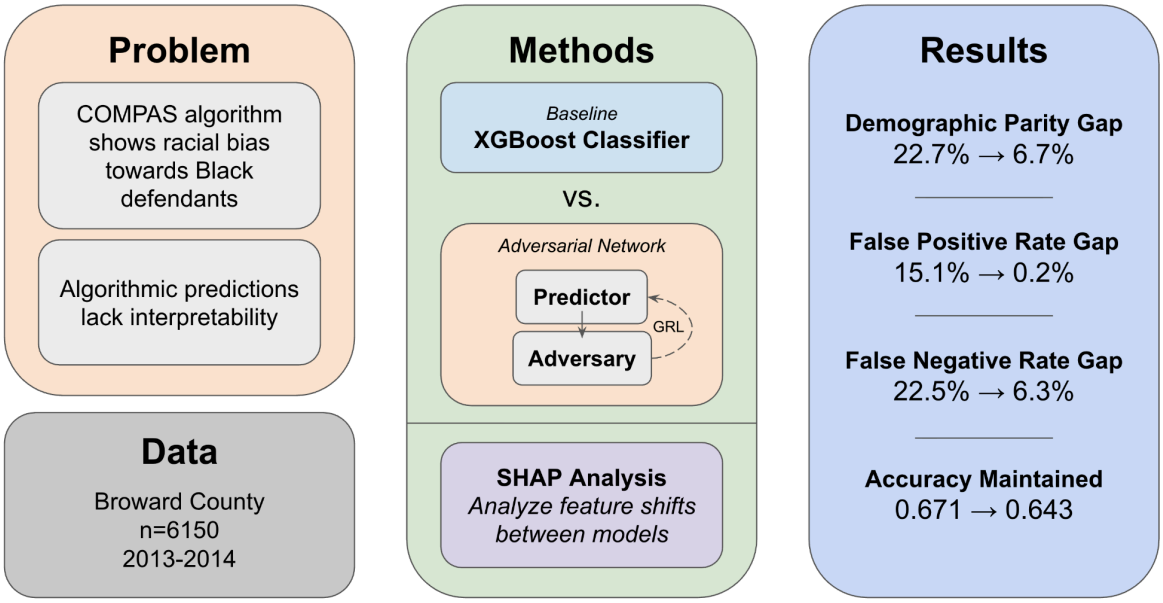
OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Exploring Bias in Recidivism Prediction via Adversarial Learning and Shapley values

Rishab Jain
rishabjain10x@gmail.com

ABSTRACT



Machine learning models that predict criminal recidivism raise important ethical concerns because they can reinforce systemic racial biases. These algorithms often lack interpretability and have been shown to disproportionately misclassify African-American defendants. In this study, we developed an adversarially-trained predictor to minimize the influence of race on recidivism classification. We hypothesized that adversarial training would reduce racial disparities in prediction rates while maintaining accuracy. To understand feature importance, we leveraged the SHAP framework to compare feature contributions between a baseline classifier and the adversarially-trained predictor. The adversarial model demonstrated a small decrease in overall accuracy but significantly reduced gaps in demographic parity, false positive rates, and false negative rates between Caucasian and African-American defendants. Additionally, it indicated reduced reliance on race-related input features for the model output. These

findings suggest that adversarial training can be used to efficiently identify and reduce racial bias in predictive algorithms, while maintaining high model performance.

INTRODUCTION

The desire for consistency in machine learning predictive tools has led to new conversations about the role of fairness. In recent years, the machine learning community has developed an increased interest in including transparency in evaluation frameworks [1]. The coalition of research centered around shifting away from "black-box" algorithms, deemed EXplainable Artificial Intelligence (XAI), offers a path forward towards transparency, interpretability, and collaboration between humans and machine-driven algorithms [2]. However, most widely used predictive algorithms lack clear insight into how input features influence decisions.

An example of applying fairness methods to machine learning is criminal recidivism prediction, which leverages supervised learning to predict the likelihood of a person to reoffend [3]. Pretrial detention facilities in urban areas and regions with high rates of crime are limited in holding capacity, posing a problem for institutions and the taxpayers that fund them. In the United States, local jails spend $14 billion annually on housing pretrial detainees, who make up over two-thirds of inmates [4], [5]. Jails across the country have sought to leverage predictive tools to advise judges in determining "high-risk" defendants that should be detained, significantly lowering the number of people held pre-trial. These algorithms typically produce a "risk assessment" score representing the probability of a defendant recidivating, advising judges whether to detain the offender or release them on bail.

Northpointe, Inc.'s Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software is the most widely used risk assessment algorithm in the United States, prominently in California, New York, and Florida's Broward County [6]. Following their arrest, pretrial detainees complete a detailed questionnaire to inquire further into their past criminal history (if applicable) and gauge social and emotional welfare [7]. This data is supplemented with official documentation listing past and current charges, and then inputted into the COMPAS algorithm to produce a risk assessment score. This score ranges from 1 to 10, with a higher score indicating a greater probability of recidivating during a two-year period following release. Scores are categorized into three levels of risk: low (1-4), medium (5-7), and high (8-10).

In 2016, ProPublica published an analysis of the COMPAS software that claimed to find large racial disparities in the algorithm's score distributions. By comparing the produced scores to criminal arrest history, they found that African-American defendants were more than two times as likely as Caucasian defendants to be misclassified as high-risk by the algorithm, while Caucasian defendants were consistently classified at lower risk levels [7]. When considering overall model performance, COMPAS correctly predicted recidivism only 61% of the time. This number is alarmingly lower than expected considering its widespread application; one study found that the average person was more likely to correctly predict recidivism given a defendant's profile than the COMPAS algorithm [8].

False positives and false negatives can have detrimental effects on those impacted by algorithms, which is why it is crucial not only to mitigate these numbers but also ensure parity across groups. ProPublica's report found a 21.4% False positive gap between African-American and Caucasian defendants (African-American defendants were more likely to be falsely classified as high-risk), and a 19.73% False negative gap (Caucasian defendants were more likely to be mislabeled low-risk) [7]. While some researchers (including Northpointe) dispute the accuracy of ProPublica's statistical analysis, it remains evident that it is important to better understand what drives these algorithms to produce their results [9], [10]. However, the COMPAS algorithm's proprietary nature prevents researchers from understanding the specifics behind it.

Increasing transparency in algorithms provides critical insight into how machine learning can perpetuate institutional biases. By understanding the roles of various input features, researchers can examine sensitive attributes and their interactions with model predictions. In some instances, significant disparities that parallel historic biases (such as racial or demographic trends) prompt additional action to be taken. In this paper, we explored the application of mitigation techniques to improve fairness and equity in recidivism prediction.

We developed a binary classification model to predict the risk of recidivism for a given defendant, trained on annotated criminal data obtained from Florida's Broward County. Unlike the COMPAS algorithm, we intentionally include demographic features (e.g. race, sex, age) to observe their impact on model output [11].

We hypothesized that the baseline classifier would display significant racial biases, measured by demographic parity gaps and imbalanced FPR and FNR. We further predicted that Shapley value approximations would reveal that race-related features were highly influential on the model output. Following adversarial training, we expected the Shapley values for race and correlated input features to decrease significantly, as well as improved fairness with respect to equalized odds and demographic parity.

We found that adversarial training flattened SHAP value distributions across all input features. Although SHAP values for race marginally increased, the model reduced its overall dependence on correlated features such as prior charges and age, which had strongly contributed to disparities in the baseline classifier. The adversarial predictor also reduced gaps in demographic parity, false positive rates, and false negative rates across African-American and Caucasian defendants, while only slightly lowering overall accuracy. However, training curves suggest that the adversary had a weaker influence than anticipated on the predictor's loss. These results partially support our hypothesis that adversarial training can reduce reliance on sensitive attributes such as race and improve fairness in recidivism prediction with limited impact on model performance.

METHODS

This section describes the dataset, preprocessing, model architectures, training procedures, fairness metrics, and Shapley value computations used in the study. All code was implemented in Python 3.12.4, XGBoost 3.0.2, scikit-learn 1.6.1, NumPy 1.26.4, Pandas 2.2.2, SHAP 0.46.0, and PyTorch 2.4.0. Models were trained on an Apple M2 Macbook Air system with an 8-core GPU. A fixed random seed of 42 was used for the baseline model, while the adversarial model did not require a fixed seed because of its repeated training runs.

All code files and datasets used can be found in the project GitHub: github.com/rishabjainX/recidivism/.

Data Collection and Preprocessing

We used the COMPAS dataset published by ProPublica, containing over ten thousand people arrested in Broward County, Florida from 2013 to 2014. The study's task was to predict "future recidivism", defined as an additional arrest within two years of the original case. This definition is consistent with Northpointe's own guidelines for interpreting COMPAS risk assessment scores [10].

The original dataset included 7214 rows and 53 columns. Since African-American and Caucasian defendants accounted for over 85% of the data, we restricted our analysis to these two groups, reducing our dataset to 6150 rows. Administrative features (e.g. arrest date, case number, screening date) were removed, and ten features were selected as input data (Table I). Of the ten, three were demographic (age, sex, race) and seven represented criminal history and behavior. The target variable was a binary label indicating whether the individual was arrested again for a separate offense within two years of the original case [7].

Feature	Description
Age	Age of individual at arrest (numerical)
Sex	Sex of individual (M/F)
Race	Race of individual (African-American/Caucasian)
Number of Prior Charges	Number of prior charges an individual has on record.
Charge Degree	Degree classification for current charge (Felony/Misdemeanor)
Charge Description	Description for current charge
Days Before Screening Arrest	Number of days between individual's arrest and date screened for COMPAS risk assessment (numeric)
Number of Juvenile Felony Charges	Number of juvenile felony charges on individual's record (prior to current arrest)
Number of Juvenile Misdemeanor Charges	Number of juvenile misdemeanor charges on individual's record (prior to current arrest)
Number of Other Juvenile Offenses	Number of other juvenile offenses on individual's record (prior to current arrest)
Recidivated Within Two Years of Current Arrest	Indication of whether individual was arrested for a separate offense within two years (binary)

Table I. Features included in the baseline model. List of the ten input features included in baseline XGBoost model, following preprocessing. Features include demographic attributes (age, sex, race) and criminal history (priors count, charge degree, charge description, juvenile records, and time from arrest to COMPAS screening).

We used a 70-30 train-test split after testing several split sizes. We preprocessed the training data by scaling numeric features and one-hot encoding categorical features. Scaling is necessary to ensure fair contributions regardless of magnitude; one-hot encoding converts categorical features into binary format [12], [13].

We observed that the charge description feature had high cardinality in the dataset (410 unique values). Applying one-hot encoding caused overfitting, so we opted for target encoding instead, which assigned numerical values based on the average recidivism rate for each label. This helped improve model stability and reduce overfitting.

Baseline Model

We used the XGBoost library to build the baseline binary classification model. We built a scikit-learn pipeline to ensure consistent preprocessing across training and testing steps. Model parameters (n_estimators, max_depth, learning_rate, subsample, colsample_bytree, reg_alpha, and reg_lambda) were selected using Grid Search to identify the best performing setup.

To ensure that model performance was not dependent on data splits, we used 5-fold cross-validation. We recorded accuracy and AUROC across splits as a stable reference for comparison with the adversarial model.

Adversarial Learning

Our adversarial setup followed the framework described by Wadsworth et al., which trains an adversary and predictor side-by-side to maximize model accuracy and minimize racial disparities [14]. We took inspiration from their work to build our own adversarial network while focusing on the insights that Shapley values provide in regards to model feature contributions.

The adversarial network consisted of two networks: a predictor *N* and an adversary *A* (Fig. 1). The predictor learned to classify two-year recidivism, while the adversary attempted to predict the race label from the predictor's output logit. A gradient-reversal layer was used to invert the adversary's gradients during backpropagation, teaching the predictor to maximize the adversary's loss. Through the training process, *N* learned to improve at predicting the target variable while using feature representations independent of the sensitive attribute to minimize overall loss. We optimized the loss $L = L_y - \alpha L_d$, where L_y represents the predictor's loss and L_d represents the loss of the adversary [14]. The hyperparameter α controlled the penalty for the adversary's accuracy.

The predictor class contained three fully connected layers, including two 256-unit ReLu layers. The adversary contains two fully connected layers, including a 100-unit ReLu layer, with a final two-unit output for binary race classification. Both networks used an Adam optimizer. To prevent bias due to class imbalance, we applied a weighted sampler to balance racial groups within each batch. The final adversarial predictor used $\alpha = 1.2$, a learning rate of e^{-4}, and batch size of 32.

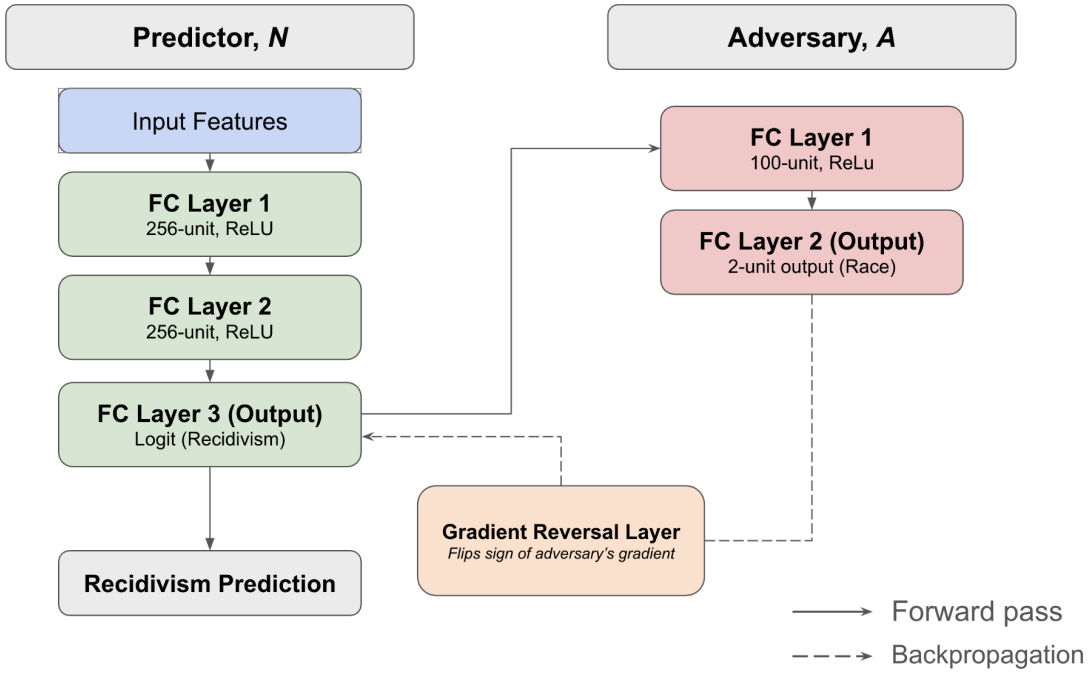


Fig. 1. Adversarial Network Architecture. Diagram depicting architecture for adversarial network. Predictor network attempts to classify two-year recidivism, while the goal of the adversary is to predict race. A Gradient Reversal Layer (GRL) is used to flip the sign of the adversary's gradient during backpropagation and teach the predictor network to maximize the loss of the adversary.

The model was trained for 300 epochs. A ramp function was added, which gradually increased α across the first 25 epochs to stabilize early training by controlling pressure. To achieve stability, the model was trained ten times independently and results were averaged across the training runs.

Fairness Metrics

We evaluated fairness using demographic parity and equalized odds. All metrics were computed on the same test set. These metrics allowed for direct comparison of fairness between both models.

Demographic parity was defined as the equality of positive prediction rates across racial groups. We quantified disparities using the demographic parity gap: the absolute difference in positive prediction rates between African-American and Caucasian defendants [14].

Equalized odds was defined as equal false positive rates (FPR) and false negative rates (FNR) across groups. A false negative prediction is a person determined not to be at risk of recidivism pre-trial, but proceeded to recidivate; a false positive, conversely, is someone predicted to recidivate, but who did not within two years of the prediction [7]. We computed FPR and FNR for each rate, and defined the FPR gap and FNR gap as the absolute difference in these values across groups.

Shapley Value Computation

Shapley values were originally developed as a tool to allocate payouts in cooperative games [15]. Lloyd Shapley formalized the assignment of values to players based on their contributions within a group. The Shapley value for player *i* is defined as:

$$\varphi_i(v) = \frac{1}{n} \sum_{S \subseteq N \setminus \{i\}} \binom{n-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S))$$

where the Shapley value for player *i* (denoted by φ_i) represents the weighted average of its marginal contribution across all possible subsets *S* of *N* total players.

Štrumbelj and Kononenko first adapted Shapley values for machine learning, by showing that it could serve as a model-agnostic tool for explainability [16]. By replacing players with input features, Shapley values quantify each feature's contributions to the model's predictions. However, exact calculation is computationally expensive because the number of subsets that can be formed for a set of *n* features grows exponentially [17]. Researchers have developed approximation methods that capture general trends while maintaining efficiency.

The SHAP library implements efficient algorithms to approximate Shapley values [18]. For tree-based models such as the baseline XGBoost classifier, we used SHAP's TreeExplainer, which computes precise Shapley values with reduced time complexity. For models with custom architecture, such as the adversarial network, SHAP's KernelExplainer was used instead. KernelExplainer randomly samples feature subsets and fits a linear model to approximate Shapley values based on model weights. We used 200 background samples when applying KernelExplainer to the adversarial model.

Mean absolute SHAP values (MASV) were used to summarize each feature's average contribution to the predicted probability of future recidivism. We used SHAP's beeswarm plots to visualize the distribution of SHAP values for each feature and observe how feature values influenced contributions [18].

SHAP values were calculated for all samples in the test set. For each prediction, the model's output logit served as the value function. The baseline value of the logit represents the expected prediction across all samples, and SHAP values adjust this baseline to produce the final output logit of the model. This logit was then passed to the sigmoid function, which compresses the value to a probability. We used a threshold of 0.5 to classify samples as positive or negative, which is standard for binary classification tasks [19].

RESULTS

We first evaluated the baseline classifier to establish a reference for model performance and understand which features most strongly contribute to predictions. The baseline XGBoost Classifier model achieved an accuracy of 0.671 and an AUROC of 0.666. We performed 5-fold cross-validation to assess model stability across different data splits, achieving a mean accuracy of 0.673 ± 0.010, and a mean AUROC of 0.722 ± 0.011. The single-test AUROC (0.666) was recorded separately for direct comparison to the

adversarial model, which did not undergo cross-validation. SHAP feature analysis indicated that priors count, age, and charge description were the most influential features in predicting two-year recidivism (Fig. 2). The race features showed slightly higher mean SHAP values for the African-American feature compared to the Caucasian feature, and the beeswarm plot reveals variance in SHAP contributions across both groups (Fig. 3).

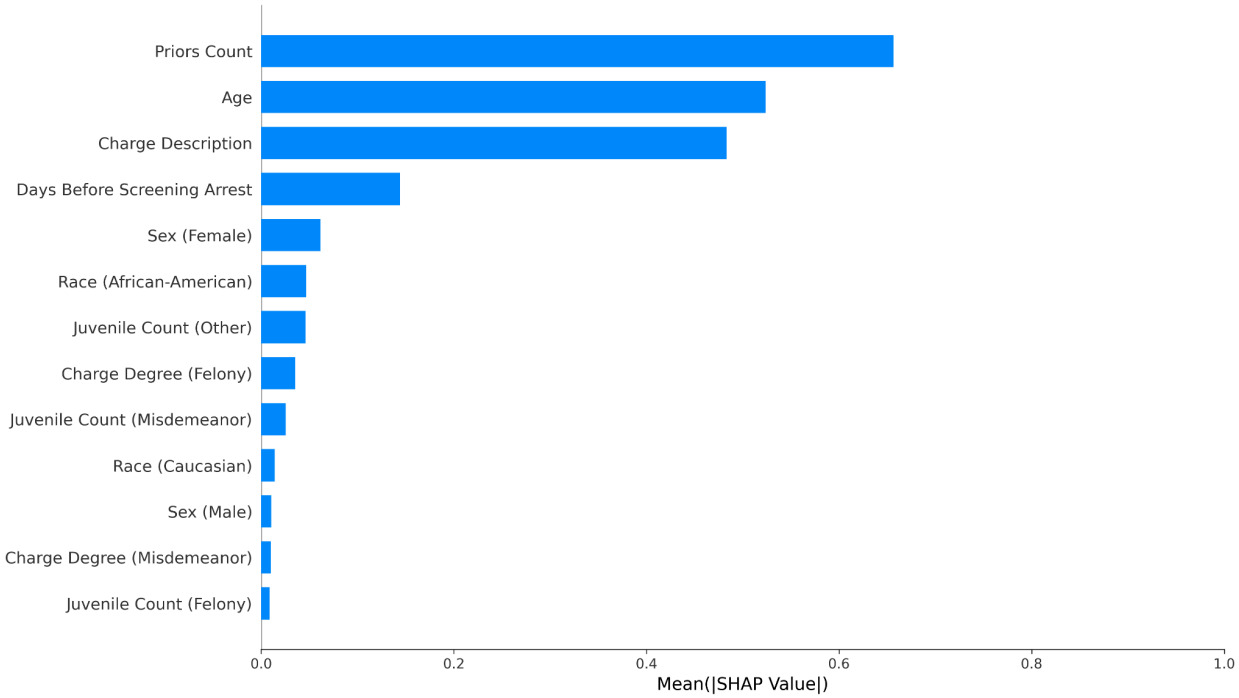


Fig. 2. Feature importance in baseline XGBoost model measured by mean absolute SHAP values (MASV). Mean absolute SHAP values (MASV) for input features in the baseline XGBoost classifier predicting two-year recidivism. SHAP values were calculated using TreeExplainer. Higher values indicate stronger average contribution to the model's output.

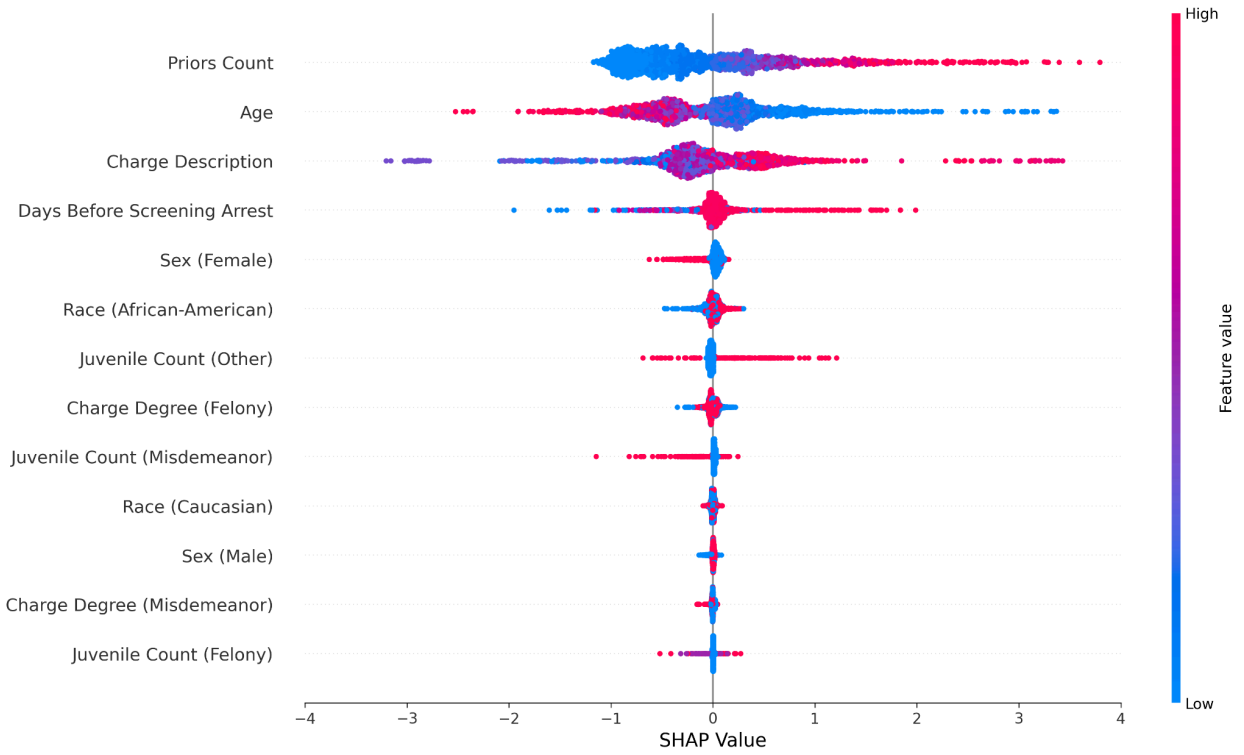


Fig. 3. SHAP value distributions for baseline XGBoost model features. Beeswarm plot displaying SHAP values for each feature in the baseline XGBoost classifier (range=7). Each point represents one defendant, colored by the feature value. SHAP values contain both the magnitude and direction of each feature's influence on the model's confidence in classifying recidivism.

Next, we evaluated whether the adversarial network successfully removed racial data from the predictor's output. Training curves show the predictor loss decreasing steadily across epochs, while the adversary's binary cross-entropy loss remained steady around 0.69, which corresponds to random guessing in a balanced binary classification task (Fig. 4). This indicates that the adversary was unable to infer racial information from the predictor's output.

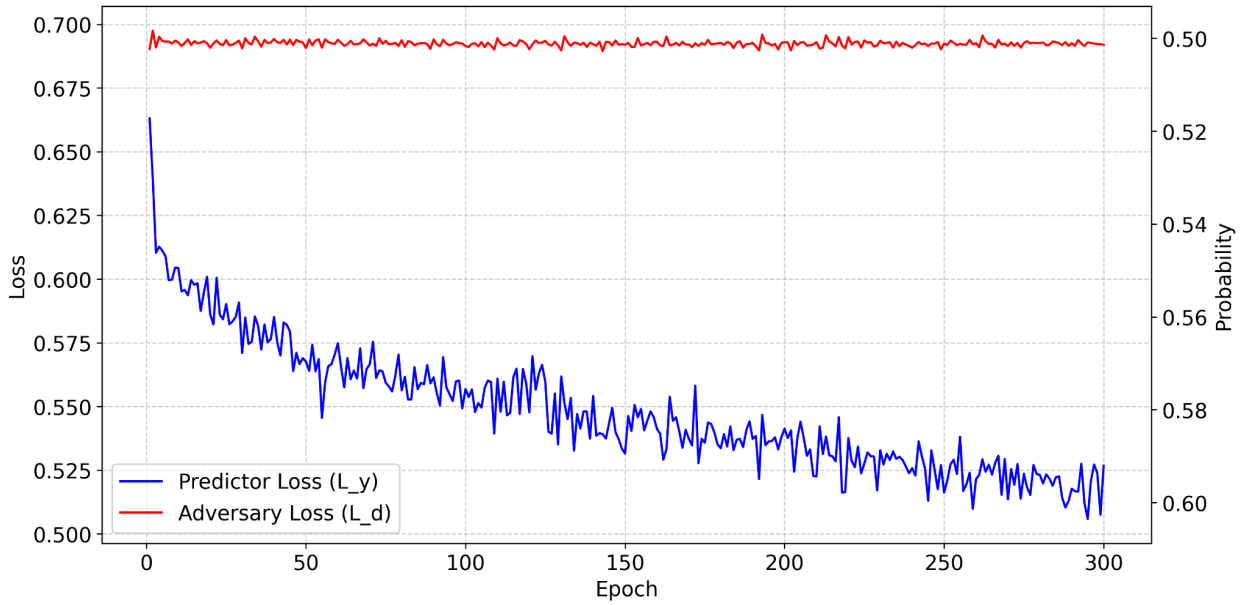


Fig. 4. Training curves for predictor and adversary networks for a sample training run. Loss functions for the training of the adversarially-trained model across 300 epochs. Right vertical axis maps losses to probability. Predictor loss decreases steadily, while adversary binary cross-entropy loss converges at approximately 0.69, equivalent to random guessing in a balanced binary classification task.

We then compared classification performance between the baseline and adversarial models. On the same test set, the adversarial predictor achieved an accuracy of 0.643 and an AUROC of 0.673 (Table II). Cross-validation was not applied to the adversarial model because it was retrained ten times to evaluate stability across different training runs. Both the baseline and adversarial models produced more false positives than false negatives, with the adversarial model displaying slightly higher error rates overall.

	Baseline Model (XGBoost)	**Adversarial Model**
True Positive	**27.26%**	26.12%
True Negative	**39.84%**	38.16%
False Positive	**19.95%**	21.08%
False Negative	**12.95%**	14.63%
Accuracy	**0.671**	0.643
AUROC	0.666	**0.673**

Table II. Performance metrics for baseline and adversarial models. Classification results on test set (30% of total dataset). Metrics reported include accuracy, AUROC, and rates of true positives, true negatives, false positives, and false negatives. Adversarial information displayed is the mean of the ten training runs.

We further examined fairness metrics to test our hypothesis that adversarial training would reduce racial bias. The adversarial predictor showcased considerable improvements across all fairness metrics. While the baseline model showed a demographic parity gap of 22.7%, the adversarial model reduced this to 6.7% (Fig. 5A). The FPR gap decreased from 15.1% to 0.2%, effectively achieving parity across both groups (Fig. 5B). The FNR gap decreased from 22.5% to 6.3%, although the adversarial model remained slightly favorable towards Caucasian defendants (Fig. 5C).

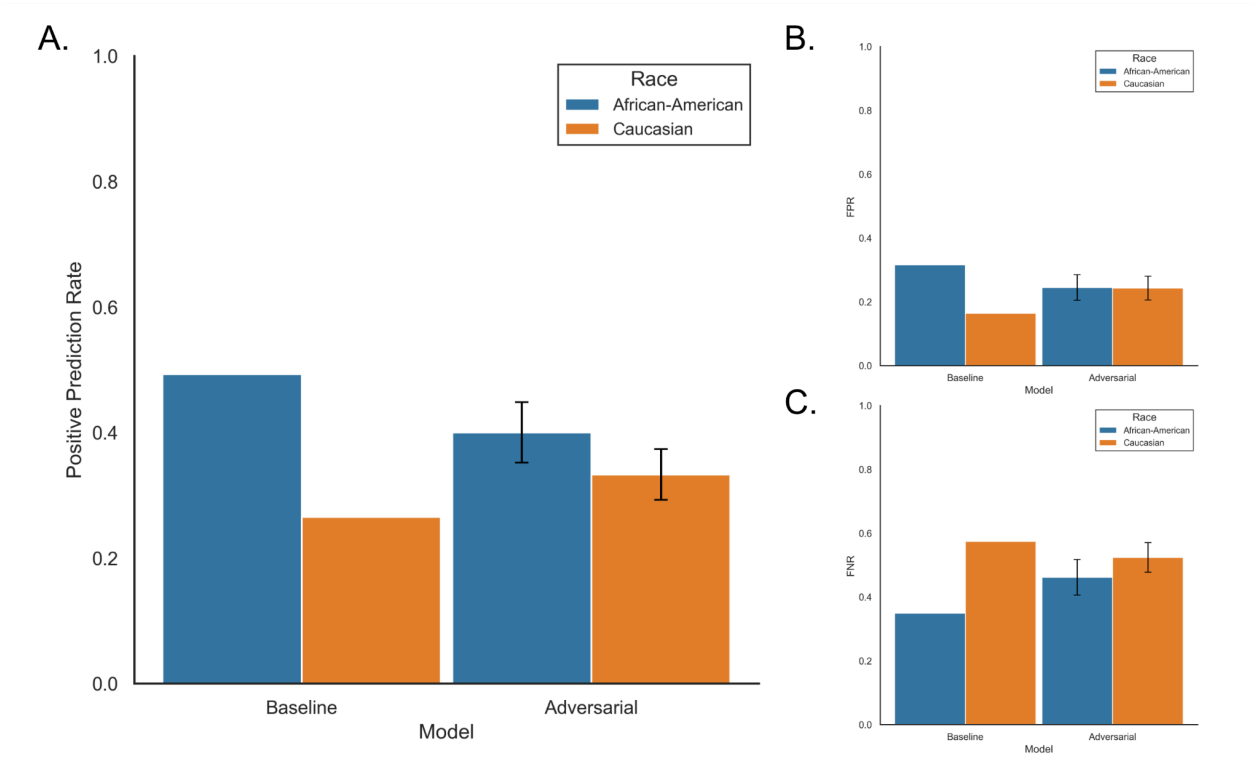


Fig. 5. Positive prediction rates, false positive rates, and false negative rates across models. PPR (A), FPR (B), and FNR (C) for African-American and Caucasian defendants compared between baseline classifier and the adversarial predictor, displayed with error bars across multiple adversarial training runs (n=10). The adversarial predictor reduced the demographic parity gap from 22.7% to 6.7%, reduced the FNR gap from 22.5% to 6.3%, and reduced the FPR gap from 15.1% to 0.2%.

Finally, we investigated feature contributions following adversarial training to understand how the model's reliance on input features had changed. SHAP analysis of the adversarial model highlighted reduced dependence on all input features, with a range of 1.2 compared to 7.0 (Figs. 3 and 6). Mean SHAP value comparisons confirmed this trend: priors count decreased from 0.66 to 0.13, age decreased

from 0.52 to 0.08, and charge description decreased from 0.48 to 0.07 (Fig. 7). Surprisingly, mean absolute SHAP values (MASV) for both race features increased by approximately 0.05 each (African-American: 0.05 to 0.10, Caucasian: 0.01 to 0.06). Although these changes were small in magnitude, the Caucasian feature exhibited a proportionally large increase due to its low value in the baseline model. This reflects the redistribution of feature contributions after reducing reliance on dominating features through adversarial training.

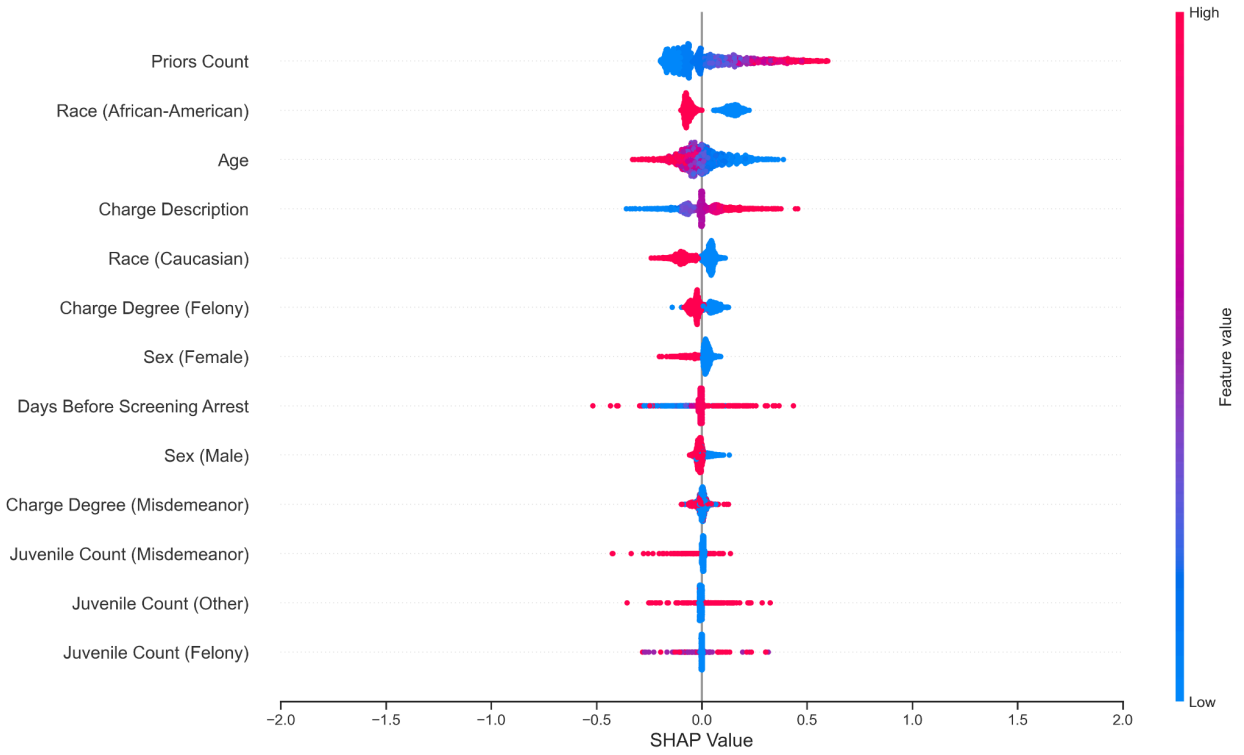


Fig. 6. SHAP value distributions for adversarial model features for a sample training run. Beeswarm plot of SHAP values from the adversarial predictor model using KernelExplainer (range=1.2). SHAP values were flattened across all features in comparison to the baseline model, indicating reduced reliance on any single attribute.

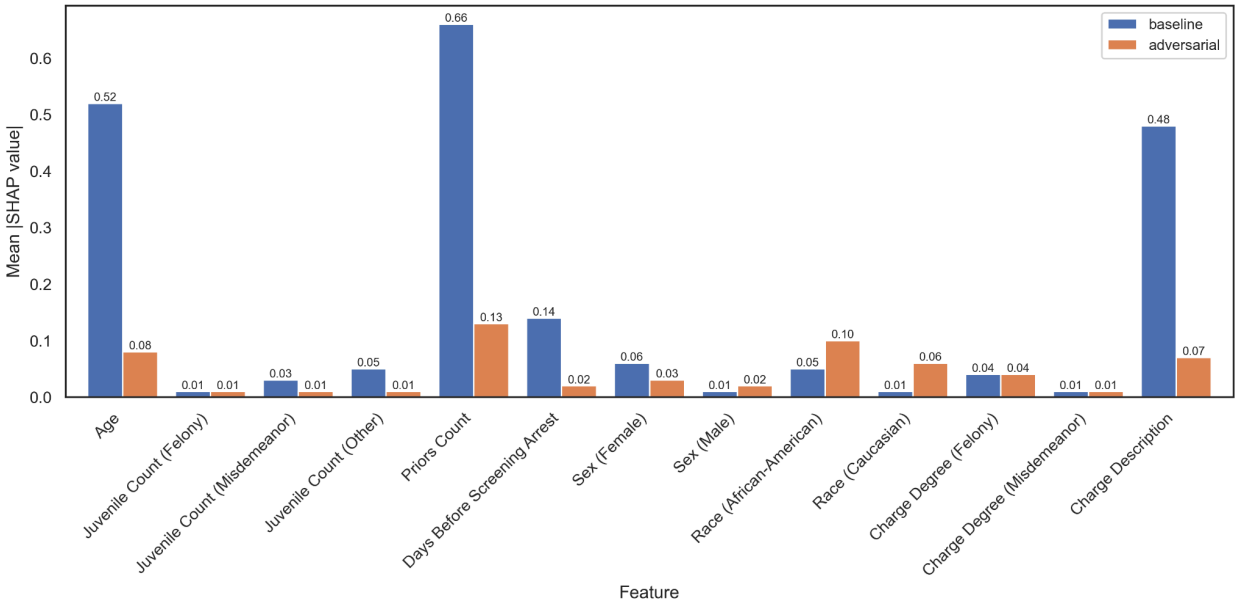


Fig. 7. Mean absolute SHAP values (MASV) across models. Comparison of mean absolute SHAP values between baseline and adversarial predictors across all features. Adversarial training decreased MSV for major proxy features such as prior charges and charge description, while marginally increasing MSV for race features.

DISCUSSION

In this study, we implemented an adversarially-trained predictor network to improve fairness in recidivism prediction while maintaining reasonable predictive performance. Although the adversarial model displayed a small decrease in accuracy (-0.028), its performance remained higher than the widely-used COMPAS algorithm (0.643 compared to 0.610). The adversarial model recorded a slightly higher AUROC than the baseline model on the single-test split (+0.007), but it remained lower than the baseline model's cross-validation AUROC (0.722 ± 0.011). During training, the adversary's loss converged to a value indicative of random guessing, suggesting that it was unable to infer racial information from the predictor's output [20]. However, its stability across training runs implies that the adversary did not learn to predict race well from the predictor's output logit in the first place. This weakens the role of the adversary in the training setup, as the predictor may have learned to correct for racial biases independently of the adversary.

Maintaining predictive accuracy while improving fairness is the primary challenge in algorithmic fairness research. In our study, the reduction in accuracy appears to be a consequence of the adversarial training objective, which specifically prioritizes satisfying fairness metrics over accuracy. Despite this trade-off, the predictor model significantly improved demographic parity by reducing the difference in positive prediction rate between Caucasian and African-American defendants by 16%, from 22.7% to 6.7%. It also

moved towards equalized odds, reducing the FPR gap to 0.2% and the FNR gap to 6.3%. Although neither metric is fully met, the model demonstrated clear progress towards parity across both racial groups. These findings reflect the trade-off commonly observed in similar studies, where enforcing fairness constraints leads to slight reductions in accuracy [14]. Together, these results helped us understand how adversarial learning affected model performance, fairness conditions, and feature contributions.

SHAP analysis indicated that the adversarially-trained predictor redistributed importance more evenly across the input features. Leading features in the baseline model such as priors count, age, and charge description showed substantial decreases in MASV following adversarial training, which produced a flatter overall SHAP distribution. This reflected a reduced reliance on a small set of dominating features and a shift towards a greater spread of feature contributions.

The marginal increases in SHAP values for both race-related features were unexpected given that the goal of adversarial training was to minimize the role of race in the predictor network. Based on the training setup, we expected a decrease in SHAP values for race, not an increase. One possible explanation involves the redistribution of SHAP values following a reduced dependence on influential features. If such features had captured information correlated with race, decreasing their importance could proportionally increase the contributions of other features, including race. Since our study did not test for these correlations directly, further analysis is needed to determine whether these features served as proxies for race. Another explanation may have been linked to the role of the adversary during training, which could have caused variation in how the predictor learned patterns related to race.

These findings partially support our hypothesis that adversarial training can reduce racial disparities in recidivism prediction. The baseline classifier displayed significant gaps in demographic parity, FPR, and FNR across racial groups, along with heavy dependence on features potentially connected with race. After adversarial training, the predictor model demonstrated reduced gaps across all fairness metrics and a more evenly distributed dependence on input features. However, it remains unclear how much of a role the adversary had on the predictor network's reduced reliance on race-related features. Despite a small increase in mean SHAP values for race features, the overall shift in feature importance and progression towards satisfying fairness metrics suggest that adversarial learning, combined with Shapley values, offers a promising approach in identifying and mitigating model biases.

Limitations

This study has several limitations that may influence how the results can be generalized. For example, this analysis was restricted to African-American and Caucasian defendants due to sample sizes, which does not reflect the true diversity of real-world pretrial detainees. As a result, the improvements in fairness recorded may not be generalized to other racial groups not tested. Since there were a comparably small number of defendants from other demographics present in the COMPAS dataset, the approach presented would need to be modified to avoid overfitting on the training data. Increasing the total number of racial groups also challenges the adversary's ability to infer the race of the defendant, which poses a risk for the overall learning of the adversarial network.

Another limitation was the dataset's diversity, which solely featured criminal history from Florida's Broward County, and thus is not representative of pre-trial populations as a whole. Offense types and demographic splits may have affected how the models learned patterns, as well as the way that adversarial training corrected disparities.

Additionally, while SHAP values brought transparency to feature contributions, our study did not directly test if highly influential features, such as prior counts and age, served as proxies for race. This meant that we could not directly attribute the racial disparities to imbalances found in these features.

Finally, the fairness metrics leveraged are only a subset of the many possible statistical definitions for fairness. While the adversarial network made significant gains on the presented fairness metrics (demographic parity and equalized odds), it may not meet other criteria for racial equality. Different applications of predictive tools in the justice system may prioritize other fairness criteria, which changes the adversarial training objective.

Future Work

Future work can build on the findings of this study to continue exploring the efficacy of adversarial learning and Shapley values. A more diverse dataset from various regions and broader demographic groups with balanced sample sizes would be needed to reflect real-world conditions. This could enable the use of the proposed adversarial setup by providing ample data to support model training and learning.

As discussed earlier, the adversary's influence on the predictor during training appeared to be minimal. Future work could attempt to strengthen the role of the adversary in the training process by tuning hyperparameters or modifying the network architecture. Once again, this requires a careful balance between preserving predictive performance while striving to satisfy fairness metrics. The study presented provides important context for researchers looking to adjust adversarial strength and establish the bounds for testing.

Further experiments could analyze if different influential features functioned as proxies for sensitive attributes such as race. Trials with additional fairness metrics, such as calibration, predictive parity, or equal opportunity can test the versatility of adversarial learning in achieving different metrics for equality. This could be used towards establishing a framework for algorithmic fairness in the criminal justice system that accounts for various fairness metrics and describes which are the most important to optimize.

While Shapley values were used in this study to bring transparency to model predictions, future work could explore alternative methods for interpretability in recidivism prediction. Similarly, investigating the use of techniques beyond adversarial learning can create comparable patterns to better evaluate explainability tools. Applying the joint framework presented in this study to other domains would provide more insight on whether adversarial training paired with Shapley values can be scaled beyond recidivism prediction.

The field of XAI continues to grow and present new methods for analyzing previously opaque predictive systems. Researchers must remain vigilant in keeping interpretability and fairness priorities in automated decision-making.

REFERENCES

[1] N. Balasubramaniam, M. Kauppinen, A. Rannisto, K. Hiekkanen, and S. Kujala, "Transparency and explainability of AI systems: From ethical guidelines to requirements," *Information and Software Technology*, vol. 159, 2023, Art. no. 107197, doi: 10.1016/j.infsof.2023.107197.

[2] S. Ali et al., "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," Information Fusion, vol. 99, 2023, Art. no. 101805, doi: 10.1016/j.inffus.2023.101805.

[3] J. Zhang, "Research on the criminal recidivism prediction based on machine learning algorithm," in Proc. 2022 2nd Int. Conf. Business Administration and Data Science (BADS), Atlantis Press, 2022.

[4] P. Wagner and B. Rabuy, Following the Money of Mass Incarceration. Prison Policy Initiative, 2017. [Online]. Available: https://www.prisonpolicy.org/reports/money.html.

[5] M. T. Stevenson and S. G. Mayson, "Pretrial detention and bail," in Academy for Justice: A Report on Scholarship and Criminal Justice Reform, E. Luna, Ed. 2017.

[6] K. Kirkpatrick, "It's not the algorithm, it's the data," Communications of the ACM, vol. 60, no. 2, pp. 21–23, 2017, doi: 10.1145/3022181.

[7] 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, May 23, 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[8] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," Science Advances, vol. 4, no. 1, 2018, Art. no. eaao5580.

[9] C. Rudin, C. Wang, and B. Coker, "The age of secrecy and unfairness in recidivism prediction," Harvard Data Science Review, vol. 2, no. 1, 2020, doi: 10.1162/99608f92.6ed64b30.

[10] W. Dieterich, C. Mendoza, and T. Brennan, "COMPAS risk scales: Demonstrating accuracy, equity, and predictive parity," Northpointe Inc., 2016.

[11] S. Corbett-Davies, E. Pierson, A. Feller, and S. Goel, "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear," The Washington Post: Monkey Cage, Oct. 17, 2016. [Online]. Available: https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

[12] "StandardScaler," Scikit-learn Documentation. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[13] "OneHotEncoder," Scikit-learn Documentation. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html.

[14] C. Wadsworth, F. Vera, and C. Piech, "Achieving fairness through adversarial learning: An application to recidivism prediction," arXiv:1807.00199, 2018.

[15] L. Shapley, "A value for n-person games," in Contributions to the Theory of Games II, H. W. Kuhn and A. W. Tucker, Eds. Princeton, NJ: Princeton Univ. Press, 1953, pp. 307–317.
[16] E. Štrumbelj and I. Kononenko, "An efficient explanation of individual classifications using game theory," Journal of Machine Learning Research, vol. 11, pp. 1–18, 2010.
[17] J. Castro, D. Gómez, and J. Tejada, "Polynomial calculation of the Shapley value based on sampling," Computers & Operations Research, vol. 36, no. 5, pp. 1726–1730, 2009, doi: 10.1016/j.cor.2008.04.004.
[18] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.
[19] "Classification threshold," Scikit-learn Documentation. [Online]. Available: https://scikit-learn.org/stable/modules/classification_threshold.html.
[20] D. Godoy, "Understanding binary cross-entropy/log loss: A visual explanation," Towards Data Science, Nov. 21, 2018. [Online]. Available: https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a.