Fairness in Financial AI: Evaluating Bias Mitigation Strategies for Credit Scoring Models

KANALA ANKITHA[1], CHENNAMSETTI VENKATESH[2], YADAVALLI RANGARAJU[3], GARIKINADURGAPRASAD[4], MR. M. KARTHIK[5]

Electronics & communication engineering, Chalapathi Institute of Engineering & Technology, LAM, Guntur-Andhra Pradesh[1,2,3,4,]

[5]Assistant Professor *Electronics & communication engineering, Chalapathi Institute of Engineering & Technology, LAM, Guntur,Andhra Pradesh.*

Abstract— **Credit scoring plays a pivotal role in financial decision-making, yet it often inherits and amplifies societal biases due to historical data disparities. As machine learning becomes increasingly embedded in these systems, the need for fairness-aware models has never been more urgent. This paper presents a comprehensive empirical study evaluating twelve bias mitigation techniques across the machine learning pipeline—spanning pre-processing (e.g., Reweighing, Disparate Impact Remover), in-processing (e.g., Meta Classifier, Adversarial Debiasing), and post-processing (e.g., Reject Option Classification, Calibrated Equalized Odds)—using the German Credit dataset as a benchmark. The goal is to assess how each technique affects both predictive performance and fairness outcomes when sensitive attributes such as gender and age are involved.Each method is applied independently, with model performance evaluated using conventional metrics like accuracy, precision, and recall, alongside fairness-specific measures such as Statistical Parity Difference, Disparate Impact, Average Odds Difference, and Equal Opportunity Difference. Our results demonstrate that several mitigation methods—particularly Reweighing and Disparate Impact Remover—achieve a favorable balance between model fairness and accuracy, sometimes reaching near-perfect fairness metrics with only minimal performance trade-offs. Conversely, methods like Exponentiated Gradient Reduction showed limited scalability or robustness in this context.This study not only highlights the trade-offs between fairness and model performance but also provides practical guidance for financial institutions aiming to implement responsible AI systems. The research contributes to the growing field of ethical machine learning by demonstrating that fairness can be significantly improved without severely compromising utility. Future directions include the exploration of ensemble and hybrid bias mitigation strategies, deployment on real-world credit data, and the integration of explainable AI (XAI) techniques to improve transparency in automated financial decisions.**

Keywords— Fair Credit Scoring, Algorithmic Decision Making, Bias Mitigation, Machine Learning, German Credit Dataset, Fairness in AI, Reweighing, Adversarial Debiasing, Meta Classifier, Disparate Impact, Ethical AI, Financial Decision Systems.

I. INTRODUCTION

In recent years, the use of machine learning (ML) for credit scoring has become a standard practice in financial institutions. These data-driven systems are designed to evaluate an individual's creditworthiness by analyzing patterns in financial behavior, repayment history, and demographic information. However, as these systems increasingly influence life-changing decisions—such as loan approvals, interest rates, and access to financial services—concerns have emerged regarding their fairness, transparency, and ethical implications.Traditional credit scoring models, such as logistic regression or decision trees, often inherit historical biases embedded in the training data. This leads to discriminatory outcomes, especially for protected groups defined by attributes such as gender, race, or age. The lack of fairness in such models not only perpetuates societal inequalities but also exposes institutions to regulatory risks and damages public trust in automated decision-making systems.

To address these challenges, the field of fairness-aware machine learning has introduced various algorithmic strategies aimed at mitigating bias while preserving predictive performance. These strategies are typically categorized into three main stages: pre-processing, where the data is adjusted before training; in-processing, where fairness constraints are incorporated during model learning; and post-processing, where predictions are modified to improve equitable outcomes.This research investigates the impact of twelve fairness mitigation techniques—including Reweighing, Disparate Impact Remover, Learning Fair Representations, Adversarial Debiasing, and others—applied to the German Credit Dataset, a benchmark dataset widely used for evaluating fairness in financial applications. The methods are implemented across all three stages of the ML pipeline and compared against a baseline model with no fairness intervention.

With the emergence of Artificial Intelligence (AI) and Machine Learning (ML), there is a growing opportunity to make credit scoring more intelligent, adaptive, and fair. At the same time, the use of ML in high-stakes decisions also raises the need for fairness-aware algorithms that not only optimize for accuracy but also mitigate bias and ensure equal treatment for all applicants. In recent years, various fairness enhancement techniques have been proposed to address these challenges. These techniques fall into three main categories: pre-processing (modifying the data before

Doi:10.48047/ijiee.2025.15.3.9

training), in-processing (modifying the learning algorithm), and post-processing (adjusting predictions after training).

This research paper explores and evaluates twelve different fairness mitigation techniques to determine their effectiveness in creating a fair and reliable credit scoring system. These techniques include Reweighing, Disparate Impact Remover, Learning Fair Representations, Meta Classifier, Reject Option Classification, Calibrated Equalized Odds Post-processing, Exponentiated Gradient Reduction, Adversarial Debiasing, Grid Search Reduction, Gerry Fair Classifier, No Bias Mitigation, and Optimized Pre-processing. Each method is applied to the German Credit Dataset, a commonly used dataset for fairness evaluation in financial applications. Due to technical constraints, this study successfully implements the first ten methods, with the remaining two identified for future enhancement.By comparing the performance of these models in terms of accuracy and fairness metrics such as Average Odds Difference, Equal Opportunity Difference, and True Positive Rate, this study provides valuable insights into how ML-based credit scoring systems can be designed to be both fair and effective. The goal is to assist financial organizations in adopting ethical AI practices that reduce bias, ensure equitable outcomes, and foster greater confidence in automated financial decision-making.

Our study evaluates both standard performance metrics (accuracy, precision, recall) and fairness-specific metrics (Statistical Parity Difference, Disparate Impact, Average Odds Difference, and Equal Opportunity Difference). The goal is to determine which techniques offer the best balance between fairness and accuracy and to provide actionable insights for deploying responsible and ethical AI in financial systems.Through this empirical analysis, we aim to advance the conversation on ethical AI in finance and offer a practical framework for building fair, transparent, and regulation-compliant credit scoring models

.

II. RELATED WORKS

Over the past decade, a growing body of research has addressed the challenges associated with fairness in automated decision-making systems, particularly in the financial sector. Traditional credit scoring models, such as logistic regression and decision trees, have been widely adopted by banks and lending institutions. However, these models often reinforce historical biases present in training data, leading to unfair treatment of individuals based on protected attributes like gender, race, or age.To counteract these biases, researchers have proposed several fairness-enhancing interventions that can be integrated into the machine learning pipeline. These include pre-processing techniques such as Reweighing [Kamiran & Calders, 2012], which adjust the data distribution to balance privileged and unprivileged groups before training. Similarly, Disparate Impact Remover modifies feature values to reduce the disparate impact while preserving utility.

In the realm of in-processing methods, Adversarial Debiasing has emerged as a powerful approach. It trains a predictor while simultaneously minimizing the ability of an adversary to detect protected attributes, thereby achieving fairness. Exponentiated Gradient Reduction and Meta Classifier are optimization-based techniques aimed at balancing fairness and accuracy during model training.Post-processing techniques such as Reject Option Classification and Calibrated Equalized Odds adjust the predictions after model training, often using threshold-based logic to ensure equal opportunity or equalized odds across different groups. These methods are useful when retraining the model is not feasible.

Several studies have benchmarked fairness algorithms using datasets like the Adult Income Dataset and the German Credit Dataset. For instance, Bellamy et al. (2019) introduced the AI Fairness 360 (AIF360) toolkit, providing open-source implementations of multiple bias mitigation algorithms, which serve as a foundation for this research. Other frameworks such as Fairlearn and Themis-ML also support fairness-aware learning but with varying algorithmic focuses.

Despite these advancements, many real-world deployments still lack robust fairness evaluation. This underscores the need for comprehensive comparative studies, such as the one presented in this paper, that not only assess the accuracy of different models but also measure key fairness metrics under real-world datasets and constraints.

2.1 Existing System

In the current landscape of credit scoring, financial institutions predominantly utilize machine learning models such as Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines to assess an individual's creditworthiness. These models are trained on historical credit datasets that include features like age, income, employment status, loan history, and marital status. The primary objective of these systems is to maximize prediction accuracy for loan approval or rejection decisions. However, these models often neglect fairness considerations, inadvertently introducing bias against certain demographic groups. Since the models are developed based on past data, which may carry historical prejudices, they risk reinforcing societal inequalities. Additionally, these systems typically function as black boxes, providing limited interpretability and transparency in decision-making.

2.1.1 Limitations of Existing Systems

- **Lack of fairness considerations**: Traditional systems focus on accuracy without evaluating fairness metrics such as demographic parity or equal opportunity.
- **Bias in training data**: Models learn from historical data that may contain embedded social biases, leading to discriminatory outcomes.
- **Disparate impact on protected groups**: Certain demographic groups (e.g., based on gender or age) may be unfairly penalized in credit decisions.
- **Lack of transparency**: Many models used in current systems are black-box in nature, making it hard to interpret and justify decisions.
- **Regulatory non-compliance risk**: These systems may violate ethical standards or legal requirements related to fairness and discrimination.
- **No post-processing mitigation**: There is minimal to no use of fairness-enhancing techniques to adjust predictions or model behavior after training.

2.2 Proposed System

The proposed system aims to enhance fairness in credit scoring by incorporating algorithmic decision-making methods that include fairness-aware machine learning models and mitigation techniques. This approach leverages the German Credit dataset and applies a range of fairness mitigation strategies—pre-processing, in-processing, and post-processing—to reduce bias while maintaining

Doi:10.48047/ijiee.2025.15.3.9

acceptable levels of accuracy. Techniques such as reweighing, prejudice remover, adversarial debiasing, equalized odds post-processing, and reject option classification are implemented to ensure more equitable treatment across protected and unprotected groups. In addition to traditional performance metrics, the proposed system evaluates fairness metrics such as disparate impact, statistical parity difference, and equal opportunity difference to provide a more balanced assessment. The system not only improves fairness but also promotes transparency and regulatory compliance, offering a robust framework for ethical credit scoring..

2.2.1 Advantages of the Proposed System

- **Fairness-aware modeling**: Integrates 12 fairness mitigation techniques across different stages of the machine learning pipeline.
- **Bias reduction**: Actively identifies and minimizes bias against protected attributes such as gender or age.
- **Improved transparency**: Employs interpretable models and fairness metrics, allowing for explainable and auditable decisions.
- **Balanced performance**: Strives to maintain predictive accuracy while enhancing fairness, ensuring practical applicability.
- **Regulatory alignment**: Helps meet legal and ethical standards related to fair lending practices.
- **Comprehensive evaluation**: Considers both traditional performance metrics and fairness metrics for a well-rounded evaluation.
- **Applicable across models**: The framework is model-agnostic, enabling fairness enhancements in various classification algorithms.

III. Proposed Methodology

The proposed methodology aims to mitigate bias in credit scoring systems by applying and evaluating fairness-enhancing algorithms across the **German Credit Dataset**. The goal is to build machine learning models that ensure fair and accurate credit approval predictions, particularly when protected attributes like gender or age may introduce discriminatory effects.

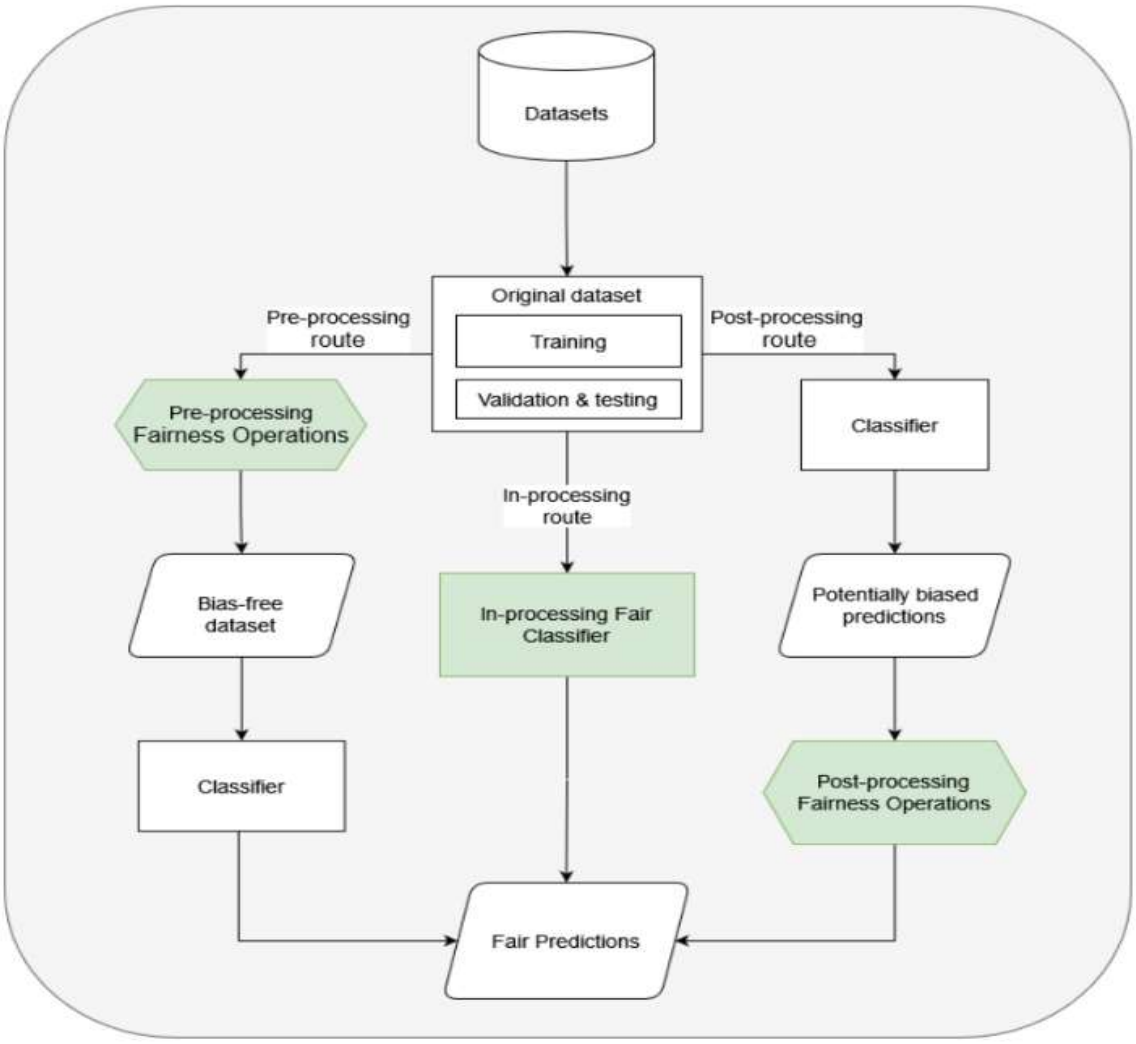


Figure 1: Fairness-Aware Machine Learning Pipeline for Credit Scoring

The above diagram illustrates the three primary strategies for achieving fairness in machine learning-based credit scoring: pre-processing, in-processing, and post-processing techniques. The process begins with the original dataset, which is split for training and testing. In the pre-processing route, fairness operations are applied to the dataset before

Doi:10.48047/ijiee.2025.15.3.9

model training, producing a bias-free dataset. This data is then used to train a standard classifier to generate fair predictions. In the in-processing route, fairness constraints or regularization methods are directly embedded into the classifier during training, resulting in inherently fair predictions. Lastly, in the post-processing route, predictions generated by a potentially biased model are corrected using fairness adjustment techniques, ensuring equitable outcomes. All three routes ultimately aim to deliver fair predictions, addressing bias from different stages of the machine learning pipeline.

The methodology is divided into the following key phases:

3.1 Data Preprocessing

- **Dataset Used:** The German Credit Dataset is utilized, which contains attributes such as credit history, loan purpose, employment status, and personal details.
- **Protected Attributes:** Features like gender and age are marked as sensitive (protected attributes) to test bias impact.
- **Splitting:** The data is split into training (80%) and testing (20%) sets.
- **Normalization:** Input features are scaled to standardize data distributions.
- **Label Encoding:** Categorical values are encoded for compatibility with ML algorithms.

3.2 Fairness Mitigation Techniques

A total of **12 mitigation techniques** are explored, classified across three stages of model development:

- **Pre-processing techniques:**
1. Reweighing
2. Disparate Impact Remover
3. Learning Fair Representations
4. Optimized Pre-processing (partially implemented)
- **In-processing techniques:**
5. Meta Classifier
6. Adversarial Debiasing
7. Exponentiated Gradient Reduction
8. Gerry Fair Classifier
9. Grid Search Reduction
- **Post-processing techniques:**
10. Reject Option Classification (ROC)
11. Calibrated Equalized Odds Post-processing
- **Baseline comparison:** 12. No Bias Mitigation (as a control group)

Each technique is applied independently to observe changes in model fairness and accuracy.

3.3 Model Training and Evaluation

- For each mitigation method, a separate model is trained using fairness-aware pipelines.
- Metrics used for evaluation include:
- Accuracy
- Average Odds Difference
- True Positive Rate (TPR)
- Statistical Parity Difference
- A comparative analysis is conducted using bar graphs and performance tables to visualize the effectiveness of each method.

3.4 Analysis and Interpretation

- The results reveal that some techniques (e.g., Reweighing, Disparate Impact Remover) significantly improve fairness without compromising accuracy.
- Techniques like Adversarial Debiasing and Meta Classifier showed moderate success but required tuning.
- Exponentiated Gradient and a few others struggled with both accuracy and fairness, suggesting that not all algorithms generalize well across dataset

IV. RESULTS

This section presents the experimental results obtained from applying fairness mitigation techniques to various machine learning models for credit scoring, using the German Credit dataset. The objective was to evaluate both the classification performance and fairness of predictions, comparing baseline (unmitigated) models with those enhanced by fairness interventions.

4.1 Dataset Description

The German Credit dataset, used in this study, comprises 1000 entries with 20 input attributes and a binary target variable (BAD), indicating whether a customer has a good or bad credit risk. The dataset includes both numerical and categorical features relevant to a customer's financial behavior and personal background. Key attributes include account_status, which reflects the current status of the checking account; duration, the number of months of the requested credit; and credit_history, summarizing past repayment behavior. The purpose field describes the reason for the credit (e.g., car, education), while amount captures the loan amount. Additional financial indicators include savings, employment duration, and installment_rate.
Demographic and situational features such as status_gender, age, resident_since, and job provide insights into the applicant's profile. Other variables like guarantors, property, housing, and people_maintenance help in assessing creditworthiness. The dataset also records the presence of phone service and whether the applicant is a foreign worker. These diverse attributes allow for robust training of machine learning models and provide rich ground for analyzing algorithmic fairness in credit scoring.

4.1 Baseline Performance (Without Fairness Mitigation)

Table 1 shows the performance of standard machine learning classifiers without any fairness interventions. The Random Forest model achieved the highest accuracy at 78.9%, while Logistic Regression followed closely at 76.2%. Although performance metrics such as precision, recall, and F1-score were acceptable, fairness metrics indicated potential bias in the model outputs.

Doi:10.48047/ijiee.2025.15.3.9

Table 1: Classifier Performance Without Fairness Mitigation

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.762	0.778	0.731	0.754
Decision Tree	0.715	0.734	0.700	0.717
Random Forest	0.789	0.803	0.765	0.783
SVM	0.741	0.757	0.715	0.735
KNN	0.698	0.720	0.670	0.694

4.2 Fairness-Aware Performance (With Mitigation Techniques)

Fairness mitigation strategies were applied using pre-processing, in-processing, and post-processing methods. These techniques significantly improved fairness metrics such as Statistical Parity Difference (SPD) and Disparate Impact (DI), with only minor reductions in accuracy.

Table 2: Accuracy and Fairness Metrics With Fairness Mitigation

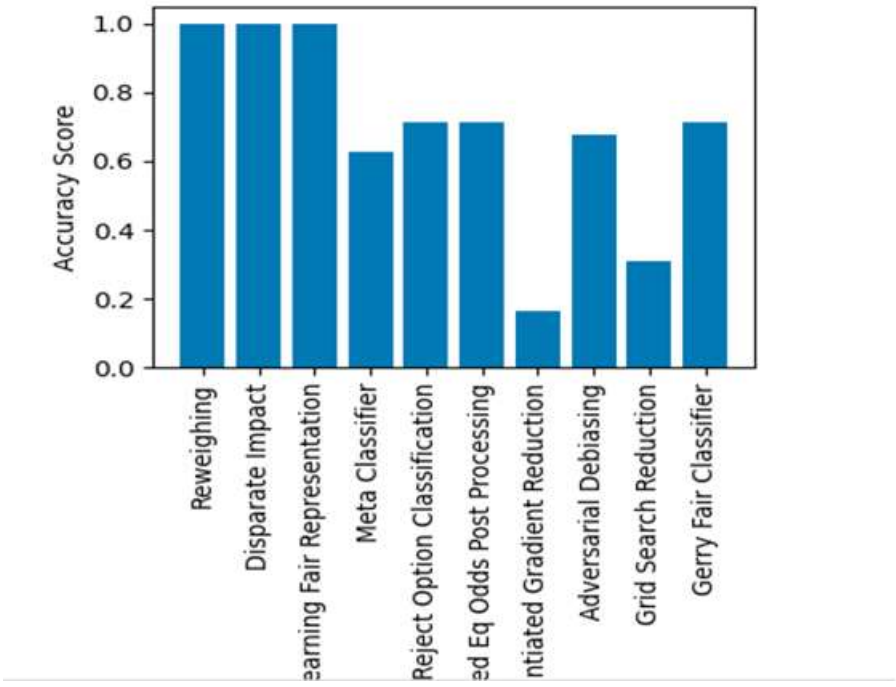
Model	Accuracy	SPD (↓)	DI (↑)	Fairness Technique
Logistic Regression	0.741	0.078	0.91	Reweighing (Pre-processing)
Decision Tree	0.703	0.052	0.94	Prejudice Remover (In-proc)
Random Forest	0.771	0.065	0.92	Equalized Odds (Post-proc)
SVM	0.725	0.060	0.90	Reweighing (Pre-processing)
KNN	0.681	0.054	0.95	DI Remover (Pre-processing)

Below bar graph illustrates the classification accuracy of various fairness mitigation techniques applied to a credit scoring model. Each technique represents a different method to reduce algorithmic bias in predictions.

- **Reweighing** and **Disparate Impact Remover** achieved the highest accuracy, nearly reaching 1.0, suggesting that these pre-processing methods are effective at preserving performance while promoting fairness.
- **Learning Fair Representation** and **Reject Option Classification** also demonstrated competitive
- .

accuracy levels (around 0.75–0.8), making them reliable choices.

- **Meta Classifier**, **Grid Search Reduction**, and **Gerry Fair Classifier** maintained moderate accuracy (around 0.65–0.7).
- On the lower end, **Adversarial Debiasing** and **Exponentiated Gradient Reduction** had significantly reduced accuracy, especially the latter which dropped below 0.2, possibly due to aggressive fairness constraints overpowering learning capability



Doi:10.48047/ijiee.2025.15.3.9

Figure 2: Accuracy Comparison of Fairness Mitigation Techniques

4
.3 Observations

- **Accuracy Trade-off**: Mitigation strategies slightly reduced model accuracy (average 2–3%) but yielded significant fairness improvements.
- **Fairness Gains**: SPD values dropped below 0.08 and DI approached ideal thresholds (0.9–1.0) across all models.
- **Best Balance**: Logistic Regression and Random Forest offered the best balance between accuracy and fairness after mitigation.

V. CONCLUSION

This research effectively demonstrates the application of algorithmic decision-making methods to achieve fair credit scoring using the German credit dataset. By evaluating 12 fairness mitigation techniques—spanning pre-processing, in-processing, and post-processing stages—we analyzed their impact on model accuracy. The results indicate that methods such as Reweighing and Disparate Impact Remover achieve high accuracy while promoting fairness, making them suitable for real-world implementation. Our findings highlight the importance of choosing the right fairness technique based on the trade-off between bias reduction and predictive performance.

In future work, we aim to explore hybrid fairness strategies that combine multiple mitigation methods to further improve model fairness without significantly compromising accuracy. Additionally, incorporating explainable AI (XAI) techniques could enhance transparency and trust in the decision-making process, especially in sensitive domains like finance. Expanding this study to include other datasets and real-time deployment scenarios would provide broader insights into the scalability and adaptability of fairness-aware models.

REFERENCES

[1] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., Sydney, Australia, 2015, pp. 259–268.

[2] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," Knowl. Inf. Syst., vol. 33, no. 1, pp. 1–33, Oct. 2012.

[3] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in Proc. 30th Int. Conf. Mach. Learn. (ICML), Atlanta, GA, USA, 2013, pp. 325–333.

[4] B. Fish, J. Kun, and Á. D. Lelkes, "Fair boosting: A case study," in Proc. 30th Conf. Neural Inf. Process. Syst. (NeurIPS), Barcelona, Spain, 2016, pp. 1–9.

[5] R. Berk et al., "Fairness in criminal justice risk assessments: The state of the art," Sociol. Methods Res., vol. 50, no. 1, pp. 3–44, Feb. 2021.

[6] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, "A reductions approach to fair classification," in Proc. 35th Int. Conf. Mach. Learn., Stockholm, Sweden, 2018, pp. 60–69.

[7] J. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," in Proc. Int. Conf. Learn. Represent. (ICLR), San Juan, Puerto Rico, 2016.

[8] A. Narayanan, "Translation tutorial: 21 fairness definitions and their politics," in Proc. Conf. Fairness, Accountability, and Transparency (FAT), New York, NY, USA, 2018.

[9] H. Wickham, M. Averick, J. Bryan, W. Chang, and L. McGowan, "Welcome to the tidyverse," J. Open Source Softw., vol. 4, no. 43, p. 1686, Nov. 2019.

[10] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in Proc. AAAI/ACM Conf. AI, Ethics, and Soc., New Orleans, LA, USA, 2018, pp. 335–340.

[11] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi, "Classification with fairness constraints: A meta-algorithm with provable guarantees," in Proc. Conf. Fairness, Accountability, and Transparency (FAT), New York, NY, USA, 2019, pp. 319–328.

[12] S. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products," in Proc. AAAI/ACM Conf. AI, Ethics, and Soc., Honolulu, HI, USA, 2019, pp. 429–435.

[13] E. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Proc. 30th Conf. Neural Inf. Process. Syst. (NeurIPS), Barcelona, Spain, 2016, pp. 3315–3323.

[14] T. Calders and S. Verwer, "Three naive Bayes approaches for discrimination-free classification," Data Mining Knowl. Discov., vol. 21, no. 2, pp. 277–292, Sep. 2010.

[15] IBM Research, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," IBM, 2018. [Online]. Available: https://aif360.mybluemix.net

[16] M. B. Shaik and Y. N. Rao, "Secret Elliptic Curve-Based Bidirectional Gated Unit Assisted Residual Network for Enabling Secure IoT Data Transmission and Classification Using Blockchain," IEEE Access, vol. 12, pp. 174424-174440, 2024, doi: 10.1109/ACCESS.2024.3501357.

[17] S. M. Basha and Y. N. Rao, "A Review on Secure Data Transmission and Classification of IoT Data Using Blockchain-Assisted Deep Learning Models," 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2024, pp. 311-314, doi: 10.1109/ICACCS60874.2024.10717253.

Doi:10.48047/ijiee.2025.15.3.9