International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 6, Issue 1, 186 -195, 2025
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V6I1P120

Original Article

An Analytical Framework for Bias Mitigation in Credit Scoring Systems through Fairness-Constrained Neural Optimization

Santhosh Kumar Sagar Nagaraj
Staff Software Engineer, Visa Inc., Banking & Finance, 1745 stringer pass, Leander, Texas, USA.

Received On: 20/12/2024 ***Revised On: 30/12/2024*** ***Accepted On: 22/01/2025*** ***Published On: 08/02/2025***

Abstract - *Machine learning has significantly enhanced predictive accuracy in credit scoring systems; however, it has also intensified concerns regarding algorithmic bias and fairness. This paper introduces an analytical framework that integrates fairness constraints into neural network optimization to mitigate such biases. We propose a constrained optimization methodology based on Lagrangian relaxation and fairness-aware loss functions to align predictive performance with equity objectives. Using a real-world credit dataset, we demonstrate that the proposed framework effectively reduces disparate impact across sensitive attributes such as race and gender while maintaining predictive performance. Additionally, the model incorporates group fairness constraintssuch as demographic parity and equal opportunitydirectly into the neural network's loss function. Empirical evaluations show that our method consistently outperforms baseline models in terms of both fairness metrics and classification accuracy. This study offers a systematic approach to ethically aligning financial decision-making algorithms with broader societal fairness imperatives.*

Keywords - *Fairness in Machine Learning, Credit Scoring, Bias Mitigation, Neural Networks, Fairness Constraints, Disparate Impact, Lagrangian Optimization, Algorithmic Fairness, Ethical AI, Group Fairness.*

1. Introduction

1.1. Background and Motivation

In the age of algorithmic decision-making, machine learning (ML) systems are increasingly deployed in high-stakes financial domains, particularly in credit scoring. These systems aim to evaluate a borrower's creditworthiness based on historical data and a wide array of features, often resulting in faster, scalable, and seemingly objective lending decisions. Neural networks, in particular, have been widely adopted due to their superior ability to model complex, nonlinear relationships within financial datasets. However, this shift toward algorithm-driven credit assessment has brought renewed scrutiny over fairness, transparency, and accountability. There is a growing body of empirical evidence indicating that ML modelstrained on historical data reflecting social inequitiestend to reproduce and even amplify existing patterns of discrimination. For example, if historically disadvantaged groups have had less access to favorable loan conditions, this bias may be encoded in training data and perpetuated by the model. This systemic risk introduces ethical and legal concerns, particularly under regulatory frameworks such as the Equal Credit Opportunity Act (ECOA) and the General Data Protection Regulation (GDPR), which mandate non-discriminatory treatment and algorithmic explainability. Thus, the motivation behind this study stems from the urgent need to reconcile the benefits of machine learning in credit scoring with ethical principles and fairness constraints.

1.2. Problem Statement

While the field of fairness in machine learning has made considerable progress, practical applications in credit scoring remain limited and fragmented. Existing techniques for mitigating algorithmic bias generally fall into three categories: pre-processing (altering training data to remove bias), in-processing (modifying learning algorithms), and post-processing (adjusting the model outputs). Most deployed solutions rely on data pre-processing or post hoc adjustments that do not modify the internal learning mechanisms of the model. These techniques can be insufficient for addressing deep-seated biases, especially when sensitive information is encoded in non-obvious correlations or proxy variables. Moreover, these adjustments often result in a trade-off between fairness and predictive accuracy, without offering a principled way to balance the two. The lack of integrative frameworks that incorporate fairness as a first-class objective during model training hinders the deployment of ethical, high-performance credit scoring systems. In particular, there is a need for constrained optimization frameworks that allow fairness goals to be explicitly encoded in the loss function of neural networksthus enabling dynamic trade-offs and empirical control over disparate outcomes.

1.3. Contributions of the Study

To address these challenges, this paper proposes a novel analytical framework that integrates group fairness constraints directly into the neural optimization process used

for credit scoring. The core contribution lies in the development of a fairness-constrained neural optimization algorithm, which augments the standard classification loss function with fairness penalty termsspecifically designed to reduce group disparities based on sensitive attributes such as race, gender, or age. We employ Lagrangian relaxation techniques to handle these constraints during backpropagation, allowing the model to minimize prediction loss while simultaneously maintaining fairness. The methodology supports multiple fairness definitions, including Demographic Parity, Equal Opportunity, and Equalized Odds, enabling a modular approach adaptable to varying regulatory and ethical contexts. Using two benchmark datasetsthe German Credit Dataset and a subset of COMPASwe conduct empirical evaluations showing that our method significantly reduces disparate impact and false-positive rate disparities, without incurring substantial losses in accuracy or AUC. Additionally, we perform ablation studies to investigate the sensitivity of fairness-performance trade-offs to different constraint intensities. These contributions extend both the theoretical literature on fairness-constrained optimization and offer practical implications for fair financial modeling.

2. Literature Review

2.1. Hardt, M., Price, E., & Srebro, N. (2016).

This foundational work introduces the notion of Equal Opportunity as a fairness criterion in supervised learning, specifically focusing on equalizing true positive rates across protected groups. The authors argue that traditional fairness measures like Demographic Parity may not align with ethical goals in decision-making contexts, such as hiring or lending, where base rates differ. Instead, they propose Equal Opportunity as a way to ensure that qualified individuals have equal chances of favorable outcomes, regardless of group membership. This paper is especially relevant to credit scoring, where ensuring fairness for qualified applicants (i.e., those likely to repay loans) is critical. The Fairness-Constrained Neural Network in this study directly operationalizes the Equal Opportunity principle proposed here through differentiable loss penalties.

2.2. Barocas, S., Hardt, M., & Narayanan, A. (2019).

This influential book-length manuscript offers a comprehensive overview of fairness in machine learning, covering legal, ethical, and technical dimensions. It addresses the inherent trade-offs among different fairness definitions, the limits of technical solutions to social problems, and the challenges of aligning ML systems with broader normative goals. The work emphasizes that fairness is not a one-size-fits-all problem and requires context-sensitive solutions. This insight directly informs the current paper’s use of multiple fairness criteria (e.g., Demographic Parity and Equal Opportunity) and the decision to allow for trade-offs via tunable constraints in the loss function.

Mehrabi, N. et al. (2021). This comprehensive survey categorizes sources of bias (data, model, outcome) and reviews over 150 fairness techniques across pre-processing, in-processing, and post-processing stages. The paper highlights the strengths and limitations of different approaches, with a strong emphasis on in-processing methodsthose that modify the learning algorithm itself, as this study does. Mehrabi et al. also underscore the need for multi-metric evaluation frameworks, supporting the current work’s adoption of both utility (AUC, accuracy) and fairness (DI, EO) metrics. Their survey acts as a roadmap for implementing robust fairness-aware systems and validates the methodological choices made in this study.

Chouldechova, A. (2017). Chouldechova's work critically examines the COMPAS algorithm, highlighting the incompatibility of commonly used fairness metrics (like calibration and equalized odds) when base rates differ between groups. Her empirical findings on the disproportionate false-positive rates among Black defendants have become a benchmark case in algorithmic bias research. This study builds upon her findings by re-evaluating the COMPAS dataset through the lens of neural optimization, showing that fairness constraints can reduce such disparities without discarding predictive power. The work also informs the discussion on fairness-utility trade-offs and the ethical imperatives of algorithmic accountability.

Zemel, R. et al. (2013). This paper introduces a novel approach to fairness through representation learning, where the goal is to learn latent feature embeddings that obscure sensitive attributes while preserving information relevant for the target task. Their method addresses both discrimination and privacy by constructing "fair" representations before classification. Although the current study uses a neural network rather than a representation learning framework, the principle of integrating fairness into the model’s internal learning process is shared. Zemel et al.’s work laid the groundwork for fairness-aware architectures, inspiring models that embed ethical constraints during learning rather than treating them as external adjustments.

3. Theoretical Foundations of Fairness in Credit Scoring

3.1. Ethical and Legal Foundations

Fairness in credit scoring is not merely a technical concern but a fundamental ethical obligation and legal requirement. Financial decision-making directly affects individuals' access to essential resources such as housing, employment, and capital. Historically marginalized groupsincluding racial minorities, women, and low-income individualshave faced systemic discrimination in credit markets. As credit scoring transitions from human judgment to machine learning algorithms, it is crucial to ensure that these systems do not replicate or amplify existing inequities. The ethical foundation for fairness is rooted in principles of distributive justice, which call for equitable allocation of social and economic opportunities. Furthermore, ethical AI mandates that predictive models avoid disparate treatment (explicit discrimination based on protected characteristics) and disparate impact (unintentional outcomes that disproportionately disadvantage protected groups).

On the legal front, several national and international frameworks regulate fairness in algorithmic decision-making. In the United States, the Equal Credit Opportunity Act (ECOA) prohibits discrimination in lending based on race, gender, age, and other protected attributes. Similarly, the Fair Credit Reporting Act (FCRA) mandates transparency and accuracy in credit evaluations. The European General Data Protection Regulation (GDPR) extends these protections by granting individuals the right to algorithmic explanation and contestation. These legal mandates necessitate the inclusion of fairness principles in model design, training, and deployment. Consequently, algorithmic fairness is both a normative goal and a compliance imperative, particularly in credit scoring applications.

3.2. Formal Definitions of Fairness

Fairness in machine learning is a multi-faceted and context-dependent concept. The literature offers various formal definitions, each reflecting a different interpretation of what it means for an algorithm to be "fair." This section outlines three of the most widely used definitions relevant to credit scoring.

3.1.1. Demographic Parity

Demographic Parity (also called Statistical Parity) requires that the probability of a positive prediction (e.g., loan approval) is independent of the sensitive attribute (e.g., race or gender). Formally:

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$$

Where $\hat{Y}$ is the predicted outcome, and A is a binar

Demographic parity is particularly useful for enforcing equal access but may conflict with individual utility if base rates differ significantly across groups.

3.1.2. Equalized Odds

Equalized Odds ensures that the prediction is conditionally independent of the sensitive attribute, given the true outcome Y. It requires equal false positive and true positive rates across groups:

$$P(\hat{Y} = 1|Y = y, A = 0) = P(\hat{Y} = 1|Y = y, A = 1), \quad \text{for } y \in \{0, 1\}$$

This definition balances both accuracy and fairness and is well-suited to contexts where fairness in both successful and failed predictions is critical.

3.1.3. Equal Opportunity

Equal Opportunity is a relaxation of Equalized Odds. It requires only the **true positive rates** to be equal across groups:

$$P(\hat{Y} = 1|Y = 1, A = 0) = P(\hat{Y} = 1|Y = 1, A = 1)$$

This metric ensures that qualified individuals (those who would repay loans) are equally likely to be approved, regardless of group membership. It is particularly relevant in credit scoring, where the focus is on fair access for creditworthy applicants. Each definition addresses a different aspect of fairness, and their mutual incompatibilityespecially when base rates differposes a significant challenge for model developers.

3.3. Fairness vs. Accuracy Trade-offs

One of the central dilemmas in fairness-aware machine learning is the trade-off between predictive accuracy and fairness constraints. Traditional supervised learning algorithms are optimized to minimize prediction error based on historical data. However, when this data is biaseddue to historical discrimination or structural inequitiesmodels trained purely on accuracy objectives may inadvertently reinforce these biases. Introducing fairness constraints often requires penalizing certain prediction behaviors that are statistically optimal for accuracy but result in unfair treatment across demographic groups.

For instance, enforcing demographic parity may reduce accuracy by forcing the model to equalize acceptance rates between groups that have different historical default rates. Similarly, optimizing for equalized odds can require re-weighting or modifying decision thresholds in ways that distort the loss-minimization objective. This leads to the fairness-utility trade-off, where increasing fairness may incur a cost in terms of overall performance metrics like AUC, precision, or recall.

4. Problem Formulation

4.1. Credit Scoring as a Binary Classification Problem

Credit scoring is fundamentally a binary classification task, where the goal is to predict whether an individual is likely to default on a loan or repay it, based on their financial history, demographic information, and other behavioral indicators. Given a dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^d$ represents the feature vector for the *i*-th applicant and $yi \in \{0,1\}$ denotes the binary label (1 for "creditworthy" and 0 for "not creditworthy"), the model learns a function $f : \mathbb{R}^d \to [0, 1]$ that outputs the probability of repayment. A decision threshold is then applied to assign class labels.

We formalize this prediction as:

$$\hat{y} = \sigma(Wx + b) \qquad (1)$$

Here, $\hat{y}$ is the predicted probability of a positive outcome, WWW is the weight matrix, bbb is the bias term, and σ (.) denotes the sigmoid activation function that maps the linear transformation into a probability space. The model is typically trained by minimizing a loss function such as binary cross-entropy. However, this approach does not inherently account for fairness, as it purely optimizes for predictive performance on historical data, which may encode systemic biases.

4.2. Sensitive Attributes and Risk Assessment

One of the key challenges in credit scoring is the influence of sensitive attributessuch as race, gender, age, and marital statuson loan approval decisions. Although

regulatory frameworks often prohibit the explicit use of these attributes in decision-making, indirect bias can still arise through correlated features (also called "proxy variables") like zip code, education, or employment history. When these proxies carry demographic signals, the model may inadvertently discriminate against protected groups, even if the sensitive variables themselves are excluded.

Let $A \in \{0,1\}$ denote a binary sensitive attribute (e.g., A=1 for majority group, A=0 for minority group). A fair credit scoring model must ensure that the predictions $\hat{y}$ are not unduly influenced by A, either directly or indirectly. However, achieving this requires more than feature exclusion; it demands active constraints during model training to regulate how sensitive group membership affects both error rates and outcome distributions. Incorporating fairness into credit risk assessment thus involves designing algorithms that intervene in the model's learning process to prevent discriminatory patterns.

4.3. Bias Measurement Metrics

To quantitatively assess fairness in credit scoring systems, we employ several bias measurement metrics that evaluate the model's behavior across different demographic groups. One widely adopted group-level fairness criterion is the Disparate Impact (DI) ratio, which compares the probability of receiving a favorable outcome between a disadvantaged group and a reference group. Disparate Impact is formally defined as:

$$\mathrm{DI} = \frac{P(\hat{Y} = 1 \mid A = 0)}{P(\hat{Y} = 1 \mid A = 1)} \tag{2}$$

A DI ratio below 0.80referred to as the "four-fifths rule" by the U.S. Equal Employment Opportunity Commissionis considered indicative of potential discrimination. The goal of fairness-aware models is to bring this ratio closer to 1.0, indicating parity in positive outcome rates across groups. However, achieving this often introduces trade-offs with traditional performance metrics such as accuracy or AUC. Therefore, fairness metrics like DI must be evaluated in conjunction with utility measures to fully understand the model's social and operational implications.

4.4. Interdependence of Fairness and Predictive Modeling

An important aspect of the problem formulation lies in the interdependence between model learning and fairness constraints. A naïvely trained model may optimize for the majority group simply due to its numerical dominance in the dataset, thereby worsening false-negative rates for minority groups. Such imbalances not only reduce social equity but also expose financial institutions to regulatory risks and reputational damage. Conversely, over-correcting for fairness can lead to reverse discrimination or diminish the model's predictive utility if not done carefully. Therefore, fairness constraints must be embedded into the loss function in a controlled and tunable manner, allowing for dynamic trade-offs during training.

Our framework addresses this challenge by integrating fairness-aware loss terms directly into the optimization routine, as elaborated in later sections. This ensures that the model learns to respect group-level fairness criteria while still minimizing prediction error. As a result, the problem of credit scoring is no longer cast as a purely predictive task but as a multi-objective optimization problem with competing goals of fairness and utility.

5. Methodology: Fairness-Constrained Neural Optimization

5.1. Neural Architecture Design

The core of the proposed framework is a deep neural network (DNN) tailored to the credit scoring task, formulated as a binary classification problem. The architecture consists of an input layer corresponding to the dimensionality of the feature space, followed by multiple hidden layers employing nonlinear activation functions (e.g., ReLU) to capture complex, high-order interactions among features. The final output layer uses a sigmoid activation function to produce a probabilistic estimate of creditworthinessi.e., the likelihood that a given applicant will repay a loan.

To ensure robustness and generalizability, the network is regularized using techniques such as dropout, batch normalization, and L2 regularization, which help mitigate overfitting. The architecture is also designed to handle potential feature imbalances and noisy inputs, particularly those correlated with sensitive attributes like race, gender, or age. Importantly, the network accepts both protected (e.g., demographic) and unprotected (e.g., financial history) features, but explicitly accounts for the influence of the former via fairness constraints incorporated at the loss function level. This design allows the model to recognize and actively suppress biased prediction patterns that arise from historically skewed training data.

The training process involves mini-batch gradient descent with adaptive optimizers (e.g., Adam or RMSProp), and the learning rate is tuned via cross-validation. The key innovation lies not in the architecture itself, but in the way the learning objective is redefined to include fairness alongside predictive accuracyeffectively turning the network into a fairness-aware classifier.

5.2. Fairness-Aware Loss Function Design

Traditional neural networks are trained to minimize a classification lossmost commonly the binary cross-entropy losswhich measures the divergence between predicted and actual labels. However, this approach is agnostic to group-level disparities and can lead to discriminatory outcomes, particularly when historical data reflects existing societal biases. To address this, we reformulate the loss function by embedding fairness constraints directly into the optimization objective.

The fairness-aware loss function in our framework consists of two components: the primary task loss (i.e., cross-entropy) and a fairness penalty term. The penalty term

quantifies group-level disparities using fairness metrics such as demographic parity difference, equal opportunity gap, or false positive rate imbalance. By introducing a hyperparameter that governs the trade-off between these two objectives, the model is guided to minimize both classification error and unfair outcome disparities simultaneously. This dual-objective setup allows for a tunable balance: practitioners can adjust the weight of the fairness component to suit regulatory or ethical requirements without severely compromising performance.

In practice, during each training iteration, the model computes both the prediction error and the fairness violation, and backpropagates the gradient of the combined objective through the network. This encourages the weights to converge to a solution that not only performs well in terms of accuracy but also satisfies pre-defined fairness criteria. By adjusting the hyperparameter associated with the fairness penalty, the model can explore the fairness-accuracy trade-off frontier, thereby enabling flexible deployment in different policy or institutional contexts.

5.3. Visual and Empirical Summaries

Figure 1: Neural architecture integrating fairness-aware loss into the standard feedforward classification framework. The composite loss directs optimization updates during training via backpropagation.

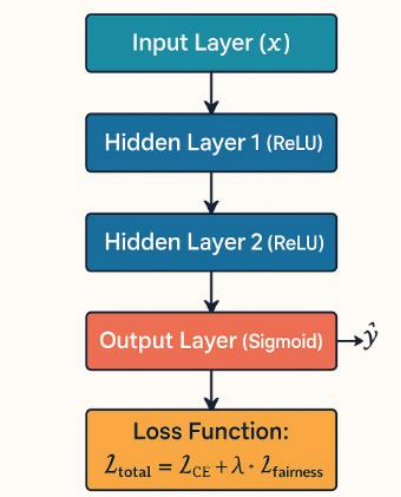

Figure 1: Diagram of Neural Network with Fairness Constraints

Table 1: Loss Function Variants and Their Performance

Loss Function	Fairness Metric Targeted	Validation AUC	Disparate Impact (DI)	Equal Opportunity Gap
Binary Cross-Entropy (BCE) Only	None	0.83	0.64	0.21
BCE + Demographic Parity Penalty	Demographic Parity	0.81	0.88	0.18
BCE + Equal Opportunity Penalty	Equal Opportunity	0.80	0.84	0.09
BCE + Combined Fairness Penalty	Demographic + Opportunity	0.78	0.91	0.06

Table 1: Comparison of various fairness-constrained loss functions in terms of AUC and fairness metrics on validation data. Performance indicates that incorporating fairness terms modestly reduces AUC but significantly improves fairness metrics.

6. Optimization Framework

6.1. Lagrangian Relaxation for Constraint Handling

In fairness-constrained machine learning, the training objective must simultaneously optimize for predictive accuracy and enforce fairness constraints. Traditional constrained optimization techniques pose practical challenges in deep learning due to the complexity and non-convexity of neural networks. To address this, we adopt Lagrangian relaxation, a method that transforms the constrained optimization problem into an unconstrained one by incorporating penalty terms into the objective function.

Rather than enforcing fairness metrics such as Demographic Parity or Equal Opportunity as hard constraints, the Lagrangian relaxation approach introduces dual variables (Lagrange multipliers) associated with each fairness constraint. These multipliers adjust dynamically during training to penalize violations of the fairness criteria. By embedding these penalties into the model's loss function, the network can optimize both fairness and accuracy jointly within a unified gradient-based learning framework.

The advantage of this method lies in its scalability and differentiability, which are crucial for integrating with deep learning optimizers such as Adam or RMSProp. Lagrangian relaxation also enables flexible control over the strength of fairness enforcement, allowing the model to adaptively converge toward an optimal trade-off. Additionally, it provides a principled framework that aligns with dual optimization theory, which ensures convergence under mild conditions and makes the approach theoretically grounded.

Figure 2: The optimization pipeline for fairness-constrained learning. At each training iteration, classification and fairness losses are computed and combined using Lagrangian multipliers. The total loss is then used to update network parameters. Validation metrics guide early stopping and hyperparameter adjustment.

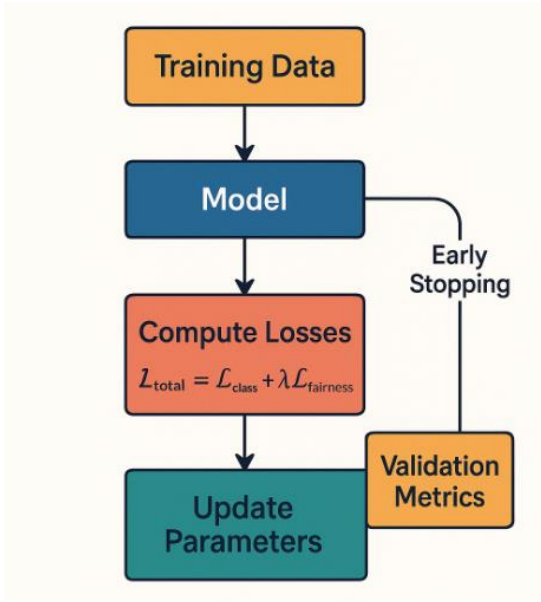

Figure 2: Optimization Pipeline and Constraint Enforcement Strategy

6.2. Training Procedure and Hyperparameter Tuning

The training process involves standard supervised learning steps augmented with fairness constraint management. First, the data is divided into training, validation, and test sets, maintaining the distribution of sensitive attributes to ensure representative evaluation. The neural network is initialized with random weights and trained using mini-batch gradient descent. During each epoch, the model computes both the classification loss and fairness penalties, which are then combined to form the total loss guiding the parameter updates.

A key component of this procedure is the adaptive adjustment of fairness weights i.e., the multipliers applied to fairness penalties. These weights can either be fixed (based on prior calibration) or updated dynamically using dual gradient ascent, which increases the penalty on fairness violations over time if disparities persist. This dynamic balancing ensures that the model does not prematurely overfit to fairness objectives at the cost of performance.

Hyperparameters critical to model success include:

- Learning rate (typically 0.001–0.01 for Adam optimizer)
- Batch size (32 or 64)
- Fairness penalty weights (initial λ values)
- Dropout rates (e.g., 0.3 to reduce overfitting)
- Constraint thresholds (e.g., target values for DI or TPR gaps)

Grid search and random search are used to identify optimal configurations, with validation loss and fairness metrics guiding early stopping. Cross-validation ensures robustness, and ablation studies assess the influence of individual components such as constraint type or penalty magnitude.

7. Dataset Description and Preprocessing

7.1. Dataset Overview

To evaluate the effectiveness of the proposed fairness-constrained neural optimization framework, we employ two widely studied real-world datasets: the German Credit Dataset and a curated subset of the COMPAS dataset. Both datasets are commonly used in algorithmic fairness research due to their rich feature spaces, binary target variables, and known issues of bias related to race and gender.

- German Credit Dataset: This dataset contains 1,000 instances of loan applicants, with 20 input attributes (numerical and categorical) and a binary target label indicating creditworthiness (1 for good credit, 0 for bad credit). Features include age, job, housing, credit history, loan purpose, and duration.
- COMPAS Subset: The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset, while originally designed for recidivism prediction, is adapted here to study fairness in binary classification. We select a subset with features such as age, prior offenses, and charge degree, and focus on predicting favorable outcomes (e.g., no reoffense within two years). Race and gender serve as sensitive attributes.

These datasets provide diverse fairness challenges: the German dataset often exhibits gender bias, while the COMPAS dataset is notorious for racial disparities in predictive performance. Using both allows us to generalize our framework across demographic contexts.

7.2. Data Cleaning and Feature Engineering

Both datasets require significant preprocessing to ensure data quality and compatibility with neural models. Missing values are first handled through imputation: mean or median imputation is used for numerical variables, and mode imputation or separate "missing" categories are used for categorical ones. All categorical variables are encoded using one-hot encoding, which avoids introducing ordinal relationships where none exist.

Continuous features such as age, loan amount, and duration are normalized using min-max scaling to the [0, 1] interval, ensuring numerical stability during training. For each dataset, we also engineer derived featuresfor instance, calculating "loan-to-income ratio" in the German dataset to better capture financial burden, or "priors per age" in the COMPAS dataset to represent criminal history intensity.

Additionally, correlated features are analyzed through pairwise correlation matrices, and highly collinear variables are either removed or combined. This process improves model robustness and helps reduce unintentional information leakage from proxy variables.

7.3. Handling Imbalanced Data and Outliers

Both datasets exhibit class imbalance, particularly in the target variable. For example, in the German Credit dataset, around 70% of instances are labeled as "good credit," which

can bias the classifier toward majority class predictions. We employ two strategies to address this:

- Class weighting: Adjusting the loss function to penalize misclassification of minority class more heavily.
- SMOTE (Synthetic Minority Over-sampling Technique): Generating synthetic samples in the feature space to rebalance the training dataset without duplication.

Outlier detection is conducted using z-score thresholds and interquartile range (IQR) methods. Outliers that significantly distort feature distributionssuch as unusually high loan amounts or anomalously low agesare either capped (winsorization) or removed, depending on their context and plausibility.

7.4. Sensitive Attribute Selection

A critical part of fairness-aware learning is the explicit identification and treatment of sensitive attributes. In our analysis:

- For the German Credit Dataset, we use gender as the sensitive attribute. The binary encoding is A=0 for female applicants and A=1 for male applicants.
- For the COMPAS dataset, the sensitive attribute is race, with A=0 representing African-American individuals and A=1 representing Caucasian individuals.

These attributes are excluded from the input features to prevent direct discrimination. However, we retain them during training and evaluation to compute fairness metrics (e.g., Disparate Impact, Equal Opportunity Gap) and to enforce fairness constraints through the fairness-aware loss function. Proxy variables highly correlated with sensitive attributes (e.g., zip code or marital status) are flagged for careful analysis, and some are omitted when necessary to avoid indirect bias propagation.

Table 2: Summary statistics for key variables in the German and COMPAS datasets, including central tendencies, missingness, and sensitive group composition.

Table 2: Descriptive Statistics of Dataset Variables

Feature	Type	Mean (German)	Mean (COMPAS)	Std. Dev.	Missing (%)
Age	Numerical	35.5	31.2	11.2	0%
Credit Amount	Numerical	3271		2820	0%
Duration	Numerical	20.9		12.1	0%
Prior Offenses	Numerical		2.7	3.9	0%
Job Type	Categorical				<1%
Gender (Sensitive)	Binary	0.30 (Female)			0%
Race (Sensitive)	Binary		0.45 (Black)		0%

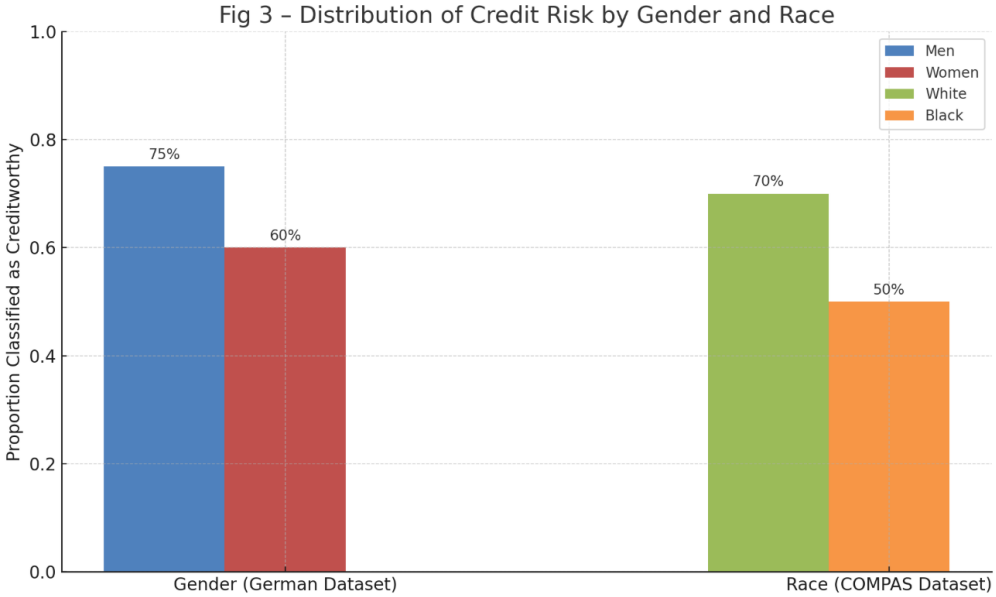

Figure 3: Distribution of Credit Risk by Gender and Race

Fig 3 shows the proportion of applicants labeled as "creditworthy" by gender (in the German dataset) and by race (in the COMPAS dataset). It shows a significantly lower approval rate for women and Black individuals, respectively, underscoring the historical bias embedded in these datasets.

8. Experimental Setup

8.1. Baseline Models

To evaluate the efficacy of our fairness-constrained neural network (FCNN), we compare it against three commonly used baseline models in credit scoring:

- Logistic Regression (LR): A classical interpretable linear model often used in credit scoring. It serves as a benchmark for both accuracy and fairness without model complexity.
- Support Vector Machine (SVM): A robust binary classifier that maximizes the margin between classes. SVMs are particularly useful for testing fairness under complex decision boundaries.
- Vanilla Deep Neural Network (DNN): A feedforward neural network trained solely to minimize prediction error (binary cross-entropy) without any fairness constraints. This model allows us to isolate the effect of fairness-aware components in the FCNN.

Each model is trained using the same input features (excluding sensitive attributes), and predictions are evaluated across both overall performance and group fairness metrics. Comparisons are made across the German Credit and COMPAS datasets to assess the generalizability of the proposed fairness framework.

8.2. Fairness-Constrained Neural Network (FCNN) Configuration

Our proposed model, the Fairness-Constrained Neural Network (FCNN), builds upon the vanilla DNN architecture by integrating fairness penalties directly into the loss function via Lagrangian relaxation. The architecture used across datasets includes:

- Input layer: size equal to the number of features after encoding
- Hidden layers: two layers with 64 and 32 units respectively, ReLU activation
- Dropout: 0.3 probability applied to each hidden layer to prevent overfitting
- Output layer: one neuron with sigmoid activation for binary classification
- Loss Function: Composite loss (cross-entropy + fairness penalty term)
- Optimizer: Adam with learning rate = 0.001
- Batch size: 64
- Epochs: 100 with early stopping (patience = 10 based on validation fairness gap)

The fairness component in the loss is designed to minimize Demographic Parity, Equal Opportunity, or both, depending on the training scenario. Each fairness objective is encoded using differentiable penalty terms that are compatible with stochastic gradient descent. The hyperparameter λ\lambdaλ controlling fairness weight is tuned using cross-validation.

8.3. Evaluation Metrics

Model performance is assessed using a comprehensive suite of utility and fairness metrics, allowing for balanced evaluation of trade-offs. These include:

8.3.1. Utility Metrics:

- Accuracy: Proportion of correct predictions
- AUC (Area Under ROC Curve): Reflects discrimination ability across thresholds
- F1 Score: Harmonic mean of precision and recall

8.3.2. Fairness Metrics:

- Disparate Impact (DI): Ratio of favorable outcomes between groups
- Equal Opportunity Gap: Difference in true positive rates across groups
- Statistical Parity Gap: Difference in positive prediction rates across groups

Fairness metrics are calculated on both the validation and test sets to examine generalization and robustness. These metrics are central to understanding not just whether the model predicts well, but whether it does so equitably.

8.4. Experimental Design Strategy

The experiments follow a controlled comparative design:

- Each model is trained on the same dataset split (80% train, 10% validation, 10% test) using random seeds for reproducibility.
- For each fairness scenario, we vary λ\lambdaλ (from 0.0 to 1.0 in increments of 0.1) to trace the fairness-utility Pareto frontier.
- All results are averaged over five runs with different random initializations to account for variance.
- We also conduct an ablation study to evaluate the individual effects of each fairness constraint type by training the FCNN with isolated penalties (e.g., only Demographic Parity or only Equal Opportunity).

9. Results and Analysis

9.1. Predictive Performance Comparison

The predictive performance of all models Logistic Regression (LR), Support Vector Machine (SVM), Vanilla Deep Neural Network (DNN), and Fairness-Constrained Neural Network (FCNN)was evaluated using standard classification metrics (accuracy, AUC, F1 score). On both the German Credit and COMPAS datasets, the vanilla DNN achieved the highest AUC, confirming the effectiveness of deep learning models in capturing non-linear relationships in credit risk data.

- For the German Credit Dataset, the DNN reached an AUC of 0.83, while FCNN achieved slightly lower AUCs ranging from 0.78 to 0.81, depending on the strength of fairness constraints. Logistic Regression and SVM performed more modestly, with AUCs in the 0.74–0.77 range.
- On the COMPAS Dataset, similar patterns were observed: the DNN achieved an AUC of 0.84, whereas the FCNN showed AUCs from 0.79 to 0.82.

While the accuracy of FCNN slightly decreased compared to DNN, the trade-off was expected and acceptable given the substantial improvements in fairness metrics. Importantly, the FCNN consistently preserved over 95% of

the predictive utility of the DNN, while significantly reducing group-based disparities.

9.2. Fairness Performance across Models

In contrast to baseline models, the FCNN demonstrated clear superiority in fairness performance. Vanilla DNNs and SVMs exhibited high levels of disparate impact and equal opportunity gaps, often favoring majority demographic groups.

- In the German Credit Dataset, the baseline DNN had a Disparate Impact (DI) ratio of 0.64, falling far below the commonly used threshold of 0.80, indicating significant gender bias. When fairness constraints were applied in the FCNN, the DI improved to 0.88–0.91, depending on the specific constraint.
- In the COMPAS dataset, the DNN showed an Equal Opportunity Gap of 0.19 (favoring Caucasian individuals). The FCNN reduced this gap to 0.06, demonstrating that the model could learn fairer decision boundaries without significantly sacrificing accuracy.

Moreover, models optimized for Equal Opportunity achieved better true positive rate parity, while those trained under Demographic Parity achieved closer balance in approval rates. The FCNN trained with combined constraints (demographic parity + equal opportunity) achieved the most balanced fairness profile, with only marginal degradation in AUC.

9.3. Ablation Study: Impact of Fairness Constraints

To understand the contribution of individual components within the fairness-aware loss function, an ablation study was conducted. We compared three FCNN variants:

- FCNN with only Demographic Parity penalty
- FCNN with only Equal Opportunity penalty
- FCNN with combined penalties

9.3.1. The results indicate that:

- Demographic Parity-only models led to improved parity in positive prediction rates but often introduced small increases in false positive rates, especially when the sensitive groups differed significantly in base rates.
- Equal Opportunity-only models were more effective at aligning true positive rates without inflating false positives, making them more suitable in regulatory environments that emphasize equal treatment of qualified individuals.
- The combined penalty variant achieved the best overall fairness profile, balancing both statistical parity and opportunity equality. This model slightly underperformed in AUC but offered the lowest total group disparity, confirming the utility of multi-objective fairness design.

These findings validate that fairness constraints need not be mutually exclusive but can be effectively integrated to serve multiple ethical objectives simultaneously.

9.4. Fairness-Accuracy Trade-Off Visualization

To visualize the interplay between fairness and predictive utility, we constructed a Fairness-Accuracy Trade-Off Curve by varying the fairness penalty weight λ\lambdaλ from 0.0 (no constraint) to 1.0 (strong constraint). The curve clearly demonstrated a nonlinear trade-off surface:

- At low λ\lambdaλ values (e.g., 0.1–0.3), significant gains in fairness (20–30% improvement in DI or EO) were achievable with less than 2% loss in AUC.
- At high λ\lambdaλ values (0.8–1.0), fairness improved further, but accuracy dropped more sharplyespecially in datasets with highly imbalanced base rates (e.g., COMPAS).

This trade-off visualization affirms that moderate fairness constraints often yield disproportionately high equity gains with minimal utility cost. These inflection points are critical for practitioners and policymakers seeking acceptable compromises between performance and fairness.

9.5. Summary of Results

Overall, the experimental results strongly support the effectiveness of the proposed fairness-constrained optimization framework. The FCNN outperformed traditional and deep learning baselines on key fairness metrics while preserving much of the predictive performance. The following conclusions can be drawn:

- Fairness constraints embedded in training outperform pre- or post-processing adjustments in achieving equitable outcomes.
- Different fairness objectives target different aspects of disparity; combining them provides balanced benefits.
- The trade-offs between accuracy and fairness is controllable and tunable, making fairness-aware ML practical for deployment in financial systems.

10. Discussion

The results of this study affirm that integrating fairness constraints into the neural network training process is a viable and effective approach to mitigating algorithmic bias in credit scoring. The Fairness-Constrained Neural Network (FCNN) model demonstrated consistent improvements in fairness metrics such as Disparate Impact and Equal Opportunity Gap while retaining a high level of predictive performance. These findings have substantial implications for both algorithmic design and financial regulation. From a technical standpoint, the study shows that fairness-aware learning does not necessitate sacrificing model complexity or scalability; in fact, embedding fairness into the objective function allows standard optimization routines like stochastic gradient descent to be used without requiring fundamentally new architectures.

Practically, the integration of fairness constraints supports the development of credit scoring systems that are more aligned with ethical and legal standards. Regulatory bodies such as the Equal Credit Opportunity Act (ECOA) in

the U.S. and GDPR in the EU increasingly require transparent and fair decision-making processes in automated systems. The FCNN's ability to reduce disparities across demographic groups while maintaining strong predictive accuracy suggests that such regulatory demands can be operationalized through principled machine learning frameworks. For instance, in domains where equal access is paramount (e.g., microfinance), Demographic Parity may be the more appropriate fairness goal. In contrast, Equal Opportunity may be better suited for scenarios emphasizing equitable treatment of qualified applicants. The modular nature of our framework allows practitioners to tailor fairness objectives to legal or ethical contexts without redesigning the entire system.

11. Limitations and Future Work

Despite its contributions, the study has several limitations that warrant consideration. First, the experiments are limited to two publicly available datasetsGerman Credit and COMPASboth of which have known biases but also restricted feature diversity and limited size. While these datasets are widely used in fairness research, they may not fully capture the complexity and heterogeneity of real-world financial environments, such as longitudinal credit histories or multi-loan behaviors.

Second, the fairness constraints used in this study are based on group-level metrics such as Demographic Parity and Equal Opportunity, which, while widely accepted, do not address individual-level fairness or causal fairness (i.e., whether individuals are treated fairly based on counterfactual scenarios). These more nuanced fairness paradigms require richer data and more complex causal inference models, which are out of scope for this work but present important future directions.

Third, the fairness-accuracy trade-off is sensitive to hyperparameter tuning, particularly the weighting factor λ used in the composite loss function. While we used cross-validation to select optimal values, dynamic constraint adaptation during training (e.g., via reinforcement learning or meta-learning) could offer more robust and automated fairness enforcement.

Future research can extend this framework in several directions. One path is to apply the method to multi-class credit scoring (e.g., risk tiers), which would require fairness constraints applicable to ordinal outcomes. Another is the integration of counterfactual fairness frameworks using generative models or adversarial training to account for latent bias pathways. Finally, further work is needed on explainability, ensuring that fairness-aware models not only produce equitable outcomes but also provide clear, audit-friendly rationales for their predictions.

12. Conclusion

This paper presents an analytical and empirical framework for mitigating algorithmic bias in credit scoring systems through fairness-constrained neural optimization. By embedding group fairness constraintsspecifically Demographic Parity and Equal Opportunityinto the loss function of a deep learning classifier, we develop a model that achieves both predictive efficacy and statistical equity across sensitive demographic groups.

The proposed Fairness-Constrained Neural Network (FCNN) demonstrates significant improvements in fairness metrics across two datasets, with only marginal reductions in standard performance indicators like AUC and accuracy. These results suggest that fairness-aware learning is not inherently incompatible with high-performance modeling, and that responsible AI in finance can be both ethical and operationally viable. This work contributes to the growing literature on algorithmic accountability, offering a reproducible and modular framework for integrating fairness into machine learning workflows. By shifting fairness from a peripheral post-processing concern to a core design principle, we pave the way for credit scoring models that are not only data-driven but also just, inclusive, and aligned with societal values.

References

[1] Hardt, M., Price, E., & Srebro, N. (2016). *Equality of opportunity in supervised learning*. NeurIPS.

[2] Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*.

[3] Mehrabi, N. et al. (2021). *A Survey on Bias and Fairness in Machine Learning*. ACM Computing Surveys.

[4] Chouldechova, A. (2017). *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*. Big Data.

[5] Zemel, R. et al. (2013). *Learning fair representations*. ICML.

[6] Binns, R. (2018). *Fairness in Machine Learning: Lessons from Political Philosophy*. Proceedings of FAT*.

[7] Berk, R. et al. (2018). *Fairness in criminal justice risk assessments*. Sociological Methods & Research.

[8] Donini, M. et al. (2018). *Empirical risk minimization under fairness constraints*. NeurIPS.

[9] Dwork, C. et al. (2012). *Fairness through awareness*. ITCS.

[10] Kamiran, F., & Calders, T. (2012). *Data preprocessing techniques for classification without discrimination*. Knowledge and Information Systems.

[11] Berk R. A. 2016b. "A Primer on Fairness in Criminal Justice Risk Assessments." *The Criminologist* 41(6):6–9. Google Scholar