Ministry of Science and Higher Education of the Republic of Kazakhstan

SDU University

Suraiyo Raziyeva

Mitigating Bias in AI-Based Loan Approval Systems through Fairness-Centric Techniques

THESIS

Presented in Partial Fulfilment for the

Degree of Master of Technical Science in Computer Science

(degree code: 7M06102)

Department of Computer Science

Faculty of Engineering and Natural Sciences

Supervisor: **PhD Meraryslan Meraliyev**

Kaskelen, June 2025

SDU University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty of Engineering and Natural Sciences

Assistant Professor, PhD Akhmedov Ramis

« 13 » June 2025

Topic of the thesis:

Mitigating Bias in AI-Based Loan Approval Systems through Fairness-Centric Techniques

Thesis submitted as part of the requirements for the award of the MSc in "7M06102 – Computer Science", SDU University

Head of Department	Zhanar Mukash
Academic Supervisor	Meraryslan Meraliyev
Master Student	Suraiyo Raziyeva

Kaskelen, 2025

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Suraiyo Raziyeva

June, 2025

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my academic supervisor, Mr. Meraryslan Meraliyev, for his continuous support, patience, and insightful guidance throughout this research. His expertise in the field of artificial intelligence and fairness-aware systems has been instrumental in shaping the direction and depth of this thesis.

I am also deeply thankful to the faculty and staff of the Department of Computer Science at SDU University for providing an intellectually stimulating and supportive environment.

My heartfelt thanks go to my fellow students, friends, and colleagues, who offered valuable feedback, technical advice, and moral support during this project. Your presence made this journey more meaningful and enjoyable.

I extend my deepest gratitude to my family, whose love, sacrifices, and unwavering belief in me have been the foundation of all my accomplishments. Without their strength and encouragement, this work would not have been possible.

Lastly, I would like to thank the broader research community whose work inspired and informed my study. This thesis is built on the collective knowledge and efforts of many scholars dedicated to advancing fairness and ethics in artificial intelligence. Thank you all.

Dedication

This thesis is dedicated to my beloved family - for their unwavering support, patience, and encouragement throughout every stage of my academic journey.

To my parents, who taught me the value of education and perseverance, and whose love has been my constant foundation.

To my friends and colleagues, whose insights, laughter, and companionship have helped me navigate the challenges of research.

And to all those who believe in fairness, inclusion, and the responsible use of technology - may this work contribute, in its own way, to building a more just and equitable future.

Abstract

As artificial intelligence (AI) becomes increasingly embedded in high-stakes decision-making systems, ensuring fairness in algorithmic outcomes has emerged as a critical concern. This thesis investigates bias and fairness in AI-based credit scoring systems, with a particular focus on gender disparities. Using the German Credit Dataset as a case study, the research evaluates the performance and fairness of several supervised machine learning models, including Logistic Regression, Decision Tree, Random Forest, XGBoost, Support Vector Machine, and Neural Network.

The study applies fairness metrics such as Statistical Parity Difference (SPD) and Disparate Impact (DI) to assess group-level inequalities in predicted loan approval outcomes. Results reveal a consistent trade-off between model accuracy and fairness, where high-performing models like Random Forest and XGBoost demonstrate notable biases against female applicants. Even interpretable models, such as Logistic Regression, exhibit fairness issues due to historical and structural biases embedded in the training data.

To address these challenges, the thesis highlights the importance of incorporating fairness-aware strategies across the machine learning pipeline, including data pre-processing, fairness evaluation, and potential post-processing mitigation. The use of tools like AIF360 and stratified sampling further strengthens the analysis.

This research contributes to the growing discourse on responsible AI by demonstrating that achieving fairness is not merely a technical goal but a socio-technical imperative. It calls for an interdisciplinary approach that combines ethical reasoning, regulatory compliance, and algorithmic transparency to ensure equitable access to financial services. The findings advocate for the development of AI systems that are not only accurate but also accountable and inclusive.

Keywords: Fairness in AI, Bias Mitigation, Credit Scoring, Machine Learning, Gender Disparities, Statistical Parity Difference, Disparate Impact, Responsible AI, AIF360, Algorithmic Fairness, Socio-technical Systems, Ethical AI

Аңдатпа

Жасанды интеллект (ЖИ) жоғары маңызға ие шешімдер қабылдау жүйелерінде кеңінен қолданылған сайын, алгоритмдік әділдік мәселесі ерекше өзектілікке ие болады. Бұл диссертациялық жұмыста жасанды интеллект негізіндегі несиелік скоринг жүйелеріндегі әділеттілік пен гендерлік теңсіздік мәселелері қарастырылады. Зерттеу German Credit Dataset деректер жиыны негізінде жүргізіліп, логистикалық регрессия, шешім ағашы, Random Forest, XGBoost, қолдау векторлық машинасы және нейрондық желі сияқты бірнеше машина оқыту модельдерінің өнімділігі мен әділеттілігі бағаланады.

Зерттеу барысында әділеттілік метрикалары — статистикалық паритет айырмашылығы (SPD) және әділетсіз әсер коэффициенті (DI) қолданылады. Нәтижелер модельдің дәлдігі мен әділеттілігі арасында тұрақты компромисті көрсетеді. Мысалы, Random Forest және XGBoost сияқты дәлдігі жоғары модельдер әйелдерге қатысты айқын әділетсіздіктерді көрсетеді. Түсіндіруге оңай логистикалық регрессия секілді модельдердің өзі оқыту деректеріндегі тарихи және құрылымдық бейтарапсыздықтың әсерінен әділетсіздікке жол береді.

Мұндай мәселелерді шешу үшін диссертация әділеттілікке бағытталған стратегияларды ЖИ жүйесінің барлық кезеңдерінде қолданудың маңыздылығын атап өтеді: деректерді алдын ала өңдеу, әділеттілікті бағалау және постпроцессинг. AIF360 құралдары мен стратификацияланған іріктеу сияқты тәсілдер талдауды күшейтеді.

Бұл зерттеу әділеттілікке қол жеткізу — тек техникалық міндет емес, сонымен қатар әлеуметтік маңызға ие екенін дәлелдей отырып, жауапты жасанды интеллект дамуына үлес қосады. Этика, құқықтық реттеу және алгоритмдік ашықтықты біріктіретін пәнаралық тәсіл әділетті, есеп беретін және инклюзивті жасанды интеллект жүйелерін құру қажеттілігін көрсетеді.

Аннотация

По мере того как искусственный интеллект (ИИ) всё чаще используется в системах принятия решений, особенно в критически важных сферах, вопрос справедливости алгоритмических решений становится всё более актуальным. В данной работе исследуются проблемы предвзятости и справедливости в ИИ-системах кредитного скоринга, с особым акцентом на гендерные различия. В качестве объекта исследования используется German Credit Dataset. Анализируется производительность и справедливость различных алгоритмов машинного обучения: логистическая регрессия, дерево решений, случайный лес, XGBoost, метод опорных векторов и нейронная сеть.

В исследовании применяются метрики справедливости, такие как статистическое различие по паритету (SPD) и коэффициент дискриминационного воздействия (DI), для оценки неравенства между группами в результатах одобрения кредитов. Результаты показывают, что между точностью модели и её справедливостью существует компромисс: модели с высокой точностью, такие как Random Forest и XGBoost, демонстрируют значительную предвзятость по отношению к женщинам. Даже интерпретируемые модели, как логистическая регрессия, страдают от встроенной в обучающие данные структурной предвзятости.

Для решения этих проблем работа подчеркивает важность использования стратегий, учитывающих справедливость, на всех этапах машинного обучения: от предобработки данных до оценки справедливости и постобработки. Также используются инструменты, такие как AIF360 и стратифицированная выборка.

Исследование вносит вклад в развитие ответственного ИИ, доказывая, что достижение справедливости — это не только техническая задача, но и социальная необходимость. Подчёркивается значимость междисциплинарного подхода, сочетающего этические, правовые и алгоритмические аспекты для обеспечения равного доступа к финансовым услугам. Полученные результаты подчеркивают необходимость разработки ИИ-систем, которые не только точны, но и справедливы, прозрачны и инклюзивны.

Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
AWI	Aggregate Weighted Index
BRIO	Bias Risk and Impact Overview
DI	Disparate Impact
DT	Decision Tree
ECOA	Equal Credit Opportunity Act
GDPR	General Data Protection Regulation
LIME	Local Interpretable Model-Agnostic Explanations
LR	Logistic Regression
ML	Machine Learning
NN	Neural Network
RF	Random Forest
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SPD	Statistical Parity Difference
SVM	Support Vector Machine
XAI	Explainable Artificial Intelligence
XGB	Extreme Gradient Boosting

Table of Contents

Chapter 1

Background and motivations

1.1 Introduction

In recent years, Artificial Intelligence (AI) has materialized as a revolutionary influence in the financial services sector, reshaping established procedures and easing data-centric decision-making at outstanding levels. Among its main applications, AI-driven credit approval systems distinguish themselves for their skill with streamlining lending procedures, diminishing operational costs, and improving predictive accuracy within assessing borrowers' creditworthiness. Through exploiting detailed machine learning algorithms, financial entities can now scrutinize wide-ranging quantities of consumer data. The institutions can do this to make nearly instantaneous lending decisions that had been, in the past, manual, time-intensive, as well as inconsistent.

However, while these systems proffer efficiency as well as scalability, they also raise grave concerns respecting fairness, transparency, plus accountability. One of the chief quandaries stems from the olden data used for AI models. A proportion of these data reflect prior human decisions. Such choices may have been swayed via systemic biases and unfair practices. As a consequence, AI systems designed from this data might inherit, reinforce, or also increase such patterns regarding unfair biases, particularly toward underrepresented, marginalized, or even socioeconomically disadvantaged various groups [1].

The problem of bias for AI models is thorough and diverse. Bias might arise not solely from the dataset proper, but from organized architecture decisions as well, including attribute designation, targeted purposes, and refinement strategies. In circumstances wherein datasets stand as skewed—as an instance, when a majority proportion of approved loans had historically went to certain demographic groups—models can learn so as to favor those aforementioned groups while disadvantaging other such groups. Furthermore, the utilization of surrogate variables such as ZIP codes and status concerning property ownership could introduce prejudice. This could transpire unknowingly, and it is perhaps circuitous. These attributes, even if viewed as neutral to one, may be quite correlated with protected traits such as race, gender, also socioeconomic status, which results within skewed treatment that violates both ethical standards along with authorized mandates [2].

Globally and also nationally, regulatory bodies now recognize hazards presented by algorithmic bias in financial services. Regulations such as the General Data Protection Regulation (GDPR) throughout the European Union as well as the Equal Credit Opportunity Act (ECOA) across the United States, do have a purpose. These regulations aim toward ascertaining that automated systems do not culminate in inequity. These regulations also aim toward interpretability of their AI models. Even with these legal frameworks, practical enforcement is still a challenge because of the complexity of modern algorithms and also the difficulty in attributing outcomes to specific sources of bias.

In response, researchers as well as practitioners have put forward various fairness-aware machine learning techniques that are designed to reduce bias and also promote equitable outcomes. These methods fall mostly into three broad groups. Such methods are categorized thus:

- **Pre-processing**: Methods change inputted data to lessen bias before the model trains. These involve methods like reweighting, sampling, and data anonymization, that aim for addressing imbalances or for removing discriminatory patterns from within the training set.
- **In-processing**: Algorithms incorporate constraints of fairness directly into the process of learning. These very techniques modify the learning algorithm itself, as often by introducing fairness-aware regularization terms or constraints, to actually balance trade-offs between predictive accuracy and equity.
- **Post-processing**: Methods which adjust predictions of the model after its training ensure more fair treatment across demographic groups. Examples include approaches for equalized odds, with demographic parity adjustments, and threshold optimization, which modify decision boundaries for reduction of disparities in outcomes.

While each category of the methods offers advantages that are distinct, no single approach guarantees fairness in contexts of all kinds. Fairness metrics like statistical parity, equal opportunity, and predictive parity can conflict. This forces practitioners to make normative decisions about which idea of fairness to prioritize. This reflects the broader tension within AI ethics, including accuracy as well as fairness, with models maximizing predictive performance, often doing it at a cost for equity [3], [4].

Beyond technical solutions, there is now an awareness that there is a need for an approach interdisciplinary for fairness within AI-driven credit systems. Addressing algorithmic bias requires computational tools in addition to certain understandings from fields, such as ethics, law, sociology, as well as economics. Biased lending models exacerbate existing economic disparities via denials of credit access to certain groups. Such models additionally pose particular, broader social challenges. For example, communities historically excluded from formal credit systems may find that it is even harder for them to build financial resilience, widen their economic opportunities, or access necessary services.

This paper has the aim to give a broad overview of the bias and fairness challenges that are in AI-based credit approval systems, with focus given to understanding where they originate, assessing what current mitigation techniques do,

and exploring all the broader social and regulatory implications. The work leverages real-world datasets like the German Credit Data. It is widely used to study algorithmic bias and fairness interventions. Through analyzing case studies, evaluating bias detection tools, as well as reviewing algorithmic correction techniques, this research seeks to highlight both the technical limitations as well as the ethical responsibilities involved throughout building equitable AI systems in finance.

Ultimately, the goal of this paper is not only to diagnose existing problems in AI lending platforms but also to contribute to the development of guidelines, frameworks, and best practices for responsible and fair AI deployment. As the use of AI continues to expand in critical decision-making areas, it is imperative that fairness is not treated as an afterthought but integrated into every stage of system design and implementation. Only through such a comprehensive and inclusive approach can we ensure that the benefits of AI in finance are accessible to all, without reinforcing or deepening historical patterns of exclusion.

1.1.1 Motivation

The motivation for this research stems from a reliance growing in Artificial Intelligence (AI) during high-stakes decision-making processes such as credit approval, where outcomes biased can have consequences in the real world of importance. Financial institutions increasingly adopt automated systems to evaluate credit risk, and so it becomes necessary to ensure the systems do not discriminate. It is important these systems do not discriminate against people based on race, gender, or socioeconomic background. My interest in this topic was driven by the realization that AI, although powerful, is not naturally neutral. It reflects the data plus assumptions upon which it is built—both of which may carry historical inequalities. By studying sources of algorithmic bias as well as evaluating fairness-aware machine learning methods, I aim to contribute to development of ethical, inclusive, and socially responsible AI systems in finance. This research aspires to close the divide between technical innovation and human values, ensuring that access to financial services is consistently fair, transparent, and equitable for people, irrespective of background.

Chapter 2

Literature review

2.1 Sources of Bias in Financial AI

Artificial Intelligence (AI) and Machine Learning (ML) have become central to modern financial decision-making, offering the potential to automate and optimize tasks such as loan approvals. These tools enhance consistency and speed, but also pose ethical and regulatory challenges when deployed without safeguards for fairness. Research has shown that models trained on biased or incomplete datasets can replicate, and even amplify, discrimination found in historical lending practices [1], [2]. Furthermore, Priya et al. note that bias in financial AI systems is often systemic, reflecting structural inequalities and the institutions that use them [3].

Abhulimen et al. (2024) claim AI needs careful audits, since financial data often encode social inequities, to stop disparities from continuing. Garcia et al. (2023) emphasize the issue's complexity, highlighting both direct discrimination and more subtle indirect forms embedded in data and institutional norms [4]. Bias can result not only from datasets but from the design of model objectives and the use of features correlated with protected attributes.

Historical bias arises when past inequalities in financial systems are reflected in datasets used for model training. When unchecked, these patterns lead models to perpetuate exclusion and disadvantage marginalized groups [5]. Unbalanced datasets contribute further by causing underrepresentation, while algorithmic bias may arise through optimization functions that fail to penalize inequity [6]. Notably, proxy variables—like ZIP codes or employment types—can act as surrogates for protected traits, indirectly embedding bias [7]. Social bias—often rooted in cultural assumptions—can manifest in decisions even without explicit discriminatory intent [8].

2.2 Fairness Metrics and Definitions

Mehrabi et al. [9] discuss a wide array of fairness metrics and strategies to audit such systemic effects. Statistical Parity Difference (SPD) and Disparate Impact (DI) serve as key indicators to quantify disparities between demographic groups. SPD reveals differences in outcome probability, while DI assesses the relative ratio of favorable outcomes. Empirical analyses of the German Credit dataset reveal

that women applicants receive disproportionately fewer favorable decisions despite equivalent financial data [8]. This aligns with Cozarenco and Szafarz [10], who observed higher interest rates imposed on women in microfinance contexts even when risk profiles were similar.

These results underscore that gender-based disparities are not theoretical but empirically confirmed. Likewise, Ashraf and Faheem [11] note that fairness must be monitored dynamically as model performance changes over time with new data. Wu [6] further explores how models might violate ethical or legal expectations even when technically performing well.

2.3 Bias Mitigation Strategies

Bias mitigation strategies are broadly classified into three main categories: pre-processing, in-processing, and post-processing. Each approach intervenes at a different stage of the machine learning pipeline and has unique advantages and trade-offs. In this section, we present detailed descriptions of these strategies and additional emerging methods in fairness-aware machine learning.

2.3.1 Pre-Processing Techniques

Pre-processing methods aim to modify the training data prior to model learning, such that the model is less likely to encode biased patterns. One widely-used technique is reweighing, which adjusts the weights of data instances to ensure balanced representation across protected and unprotected groups [4]. This method preserves the original feature space while correcting for dataset-level imbalances, making it especially effective for classical classifiers like Logistic Regression and Random Forest.

Another powerful pre-processing method is synthetic data generation. Guided by causal inference, synthetic augmentation creates balanced datasets by generating additional examples for underrepresented subgroups [14]. When applied to datasets such as the German Credit Dataset, this technique has demonstrated improvements in Statistical Parity Difference (SPD).

Participatory preprocessing frameworks are also emerging. These approaches involve affected communities in the data collection and cleaning process, ensuring that cultural and structural biases are acknowledged from the outset [17].

2.3.2 In-Processing Techniques

In-processing methods incorporate fairness constraints directly into the learning algorithm. Among these, adversarial debiasing has gained significant traction. This technique introduces an adversary network that tries to predict the protected attribute from the main model's output. The main model is penalized for revealing any sensitive information, thereby reducing bias in its predictions [12].

We explored various adversarial configurations, including the use of Focal Loss, Gradient Reversal Layers (GRL), and dynamic lambda schedulers, each contributing uniquely to performance-fairness trade-offs. GRL, in particular, was effective

in suppressing gradient signals related to sensitive features.

Emerging in-processing methods include fairness-constrained optimization and causal regularization, which explicitly add fairness objectives to the loss function or regularization terms. These methods are promising but computationally expensive.

Hybrid approaches also fall under this category. Singh et al. [20] demonstrate that combining in-processing with preprocessing (e.g., reweighing + adversarial) yields stronger fairness across intersectional dimensions like race and gender.

2.3.3 Post-Processing Techniques

Post-processing methods apply fairness adjustments to a trained model's output without altering its internal structure. This is particularly useful for black-box or deployed systems where retraining is impractical. Techniques include threshold adjustment, equalized odds post-processing, and fairness-aware rejection sampling.

Dattachaudhuri et al. [13] emphasize that post-processing can correct group-level disparities with minimal performance loss. However, these techniques do not remove the underlying bias — they simply mask it at the output level.

Post-processing is also critical for real-time fairness monitoring. As Ashraf et al. [11] highlight, threshold-based alerts on fairness metrics (e.g., SPD, DI) can trigger automated rebalancing or retraining workflows before biased outcomes reach deployment.

2.4 Gaps in the Literature

Despite progress, challenges remain. First, there is limited adoption of fairness auditing tools in practice, particularly in small financial institutions. While tools like IBM's AI Fairness 360 or Google's What-If Tool exist, their integration requires technical expertise [9].

Second, most studies use static datasets, limiting insights into how fairness evolves over time. Longitudinal fairness tracking is underexplored, yet essential in finance where model decisions accumulate [11].

Third, fairness definitions remain context-dependent. No single metric fits all regulatory, social, or ethical standards. This highlights the need for domain-specific frameworks informed by stakeholder input [18], [17].

Finally, there is a need to bridge the gap between technical debiasing techniques and legal standards like GDPR and ECOA [6]. Few studies propose concrete integration strategies that ensure regulatory alignment from the design stage.

In conclusion, the literature strongly supports the claim that algorithmic decision-making in finance can reinforce existing social inequities unless deliberate fairness-aware interventions are made. This includes model auditing, statistical testing, transparency mechanisms, and continuous feedback. Fairness is not merely a technical constraint but a design philosophy, and its success depends on how deeply it is integrated into the machine learning lifecycle - from data collection and feature selection to deployment and regulation.

2.5 Related Work

Numerous studies have explored the intersection of machine learning fairness and financial decision-making, particularly in the context of credit scoring. Among the most influential is the work of Mehrabi et al. [33], who provide a comprehensive survey of over 20 fairness metrics and categorize them by use case, highlighting the trade-offs between group, individual, and causal fairness. Their findings emphasize that no universal solution exists and that fairness evaluation must be context-specific.

Zhang et al. [12] propose adversarial debiasing as an effective in-processing method for mitigating bias in classification models. Their framework, which involves a dual-objective learning approach, inspired the architecture adopted in this thesis. The method has since become a standard reference for fairness-aware neural network training.

Bellamy et al. [24] introduced the AIF360 toolkit, which has enabled widespread adoption of fairness assessment techniques in industry and research. This toolkit provides standardized implementations of fairness metrics, debiasing algorithms, and visualization tools, and is used extensively in this work.

Beutel et al. [25] explore fairness in financial lending using real-world credit datasets. They show that fairness constraints often conflict with accuracy, but that combining preprocessing and in-processing approaches can mitigate performance loss. Their hybrid strategy parallels the multi-phase framework used in this study.

In terms of domain-specific fairness, Cozarenco and Szafarz [28] find that women are often penalized in microfinance despite having equal or better repayment rates than men. This supports our findings on gender disparities in credit approval outcomes and motivates further study into intersectional fairness.

Other researchers, such as Kamiran and Calders [26], have proposed preprocessing techniques like reweighing and preferential sampling, which were evaluated and extended in this thesis. Feldman et al. [31] propose fairness certification techniques that check if a model's decisions can be explained without access to protected attributes, reinforcing the need for explainability.

Lastly, the work of Hardt et al. [27] on equalized odds and equal opportunity remains foundational. These definitions of fairness were referenced during the evaluation phase of our models, even if not directly implemented.

Together, these studies provide the theoretical and technical backbone for our methodology. They collectively indicate that fairness-aware modeling must balance predictive accuracy, ethical responsibility, and legal compliance — a principle that guided the design of our experimental pipeline.

2.6 Theoretical Foundations of Fairness

Fairness in machine learning is a multidimensional concept with diverse interpretations across disciplines. In algorithmic decision-making, it typically refers to the absence of unwanted bias or discrimination against individuals or groups based on protected attributes such as gender, race, or age. Two primary schools of thought dominate fairness theory: group fairness and individual fairness.

Group fairness focuses on ensuring that subgroups receive similar treatment on average. Statistical Parity Difference (SPD) and Disparate Impact (DI), which are used in this thesis, are group-level fairness metrics. These metrics evaluate disparities in outcomes between protected and unprotected groups. For example, SPD measures the difference in positive classification rates, while DI compares their ratio [33, 31].

Individual fairness, by contrast, posits that similar individuals should be treated similarly, regardless of group membership. This principle was first formalized by Dwork et al., who defined fairness as a function of individual similarity under a specified distance metric [30]. While theoretically appealing, individual fairness is difficult to operationalize in practice due to the challenge of defining and computing appropriate similarity measures in high-dimensional datasets.

A fundamental tension exists between competing fairness goals. Kleinberg et al. [32] showed that in non-trivial cases, it is mathematically impossible to simultaneously satisfy several commonly used fairness criteria - specifically, predictive parity, equalized odds, and calibration - when base rates differ across groups.

Moreover, fairness can be procedural (focused on the fairness of the process) or distributive (focused on fairness of the outcomes). Legal systems often emphasize procedural fairness - ensuring decisions can be explained, contested, and traced [38]. In contrast, machine learning frameworks frequently target distributive fairness, measuring how outcomes are allocated across demographics.

This divergence creates an additional challenge: aligning fairness definitions used in model training with those recognized by law and policy. Thus, fairness is not merely a technical property to be optimized, but a socio-technical design choice requiring normative judgment and stakeholder engagement.

In summary, fairness in AI systems is a balancing act. Achieving fairness requires deliberate decisions about which metrics to use, which trade-offs to accept, and how to align technical outcomes with ethical and legal standards.

Chapter 3

Methodology

3.1 Dataset Selection

The dataset used in this study is the German Credit Dataset, a well-established benchmark for evaluating fairness and predictive performance in credit scoring models. This dataset was originally provided by Professor Dr. Hans Hofmann from the Institut für Statistik und Ökonometrie at Universität Hamburg, and has been widely used in the StatLog project as well as in various studies involving bias detection and classification in financial services.

The dataset comprises 1,000 instances, each representing a loan applicant. Two versions of the dataset are available: the original form (`german.data`) containing 20 attributes (13 categorical and 7 numerical), and a preprocessed numeric form (`german.data-numeric`) with 24 fully numerical features. In this study, the numeric version is used to ensure compatibility with machine learning models that require numerical input and to facilitate the application of fairness-aware techniques.

The dataset includes a diverse set of financial, demographic, and personal attributes, such as *credit history, employment duration, loan amount, savings account status, personal status and sex, age*, and *foreign worker status*. These features allow for both credit scoring and fairness assessment, especially across sensitive dimensions such as gender, age, and nationality. For instance, *Attribute 9* encodes both gender and marital status, making it suitable for detecting gender-based disparities, while *Attribute 13* captures applicants' age, which is often associated with indirect discrimination.

The target variable is binary and represents creditworthiness, where:

- **1 = Good credit risk**
- **2 = Bad credit risk**

Moreover, the dataset is accompanied by a cost matrix that highlights the real-world asymmetry in classification errors. Specifically, it is considered five times more costly to misclassify a "bad" customer as "good" than vice versa. This asymmetry is modeled by the following cost matrix:

Given the dataset's structured format, inclusion of sensitive attributes, and known limitations, it serves as a suitable foundation for exploring bias detection and fairness-aware machine learning in credit scoring.

Table 3.1 – Asymmetric Cost Matrix for Credit Scoring

	Predicted Good	**Predicted Bad**
Actual Good	0	1
Actual Bad	5	0

3.2 Data Processing

The data preprocessing stage is crucial for ensuring that the machine learning models operate on clean, consistent, and unbiased inputs. In fairness-aware credit scoring systems, preprocessing not only serves the standard function of preparing data for modeling but also plays a key role in addressing historical and systemic biases embedded within datasets [9].

3.2.1 Data Cleaning and Transformation

The German Credit dataset was first subjected to cleaning procedures to remove missing or inconsistent entries. Since this dataset does not contain null values, no imputation was necessary. However, certain categorical attributes such as ‘housing‘, ‘job‘, and ‘personal status‘ required standardization. Features with textual or symbolic categories were label-encoded or one-hot encoded to make them compatible with standard classifiers. Normalization was applied to numerical variables such as ‘credit amount‘ and ‘duration‘ to ensure that scale did not disproportionately affect learning outcomes.

3.2.2 Encoding Protected Attributes

To support fairness evaluation, the sensitive attribute ‘sex‘ was explicitly encoded as a binary variable (‘1 = male‘, ‘0 = female‘). This facilitated the measurement of group fairness metrics such as Statistical Parity Difference (SPD) and Disparate Impact (DI), which require group membership to be clearly defined [8, 9]. In addition, the ‘age‘ variable was binarized at the threshold of 25 years to create a ‘young‘ vs. ‘old‘ grouping, consistent with prior fairness literature [6].

3.2.3 Feature Selection

Features that were either weakly correlated with the outcome variable or strongly collinear with other predictors were excluded to avoid redundancy and noise. A Pearson correlation matrix and mutual information ranking were used to assess feature utility. Additionally, variables that serve as potential proxies for sensitive attributes - such as ‘job‘ or ‘housing‘ (which may reflect socio-economic status) - were flagged for further fairness analysis, though they were retained in some baseline models to test for proxy bias effects [7].

3.2.4 Train-Test Split

The dataset was partitioned into training and test sets using a stratified split to maintain the distribution of the target variable across both sets. A typical 80/20 division was used to allow sufficient training data while preserving a representative test sample for unbiased evaluation.

3.2.5 Fairness-Aware Preprocessing

To assess the impact of dataset balancing on fairness outcomes, the 'reweighing' technique was applied as a pre-processing debiasing method [4]. This method assigns different weights to samples based on group membership and label, effectively correcting imbalances in the joint distribution of sensitive attributes and target values. The weighted dataset was then passed to each classifier to evaluate the effectiveness of this intervention in mitigating disparities in prediction outcomes.

3.2.6 Summary

In sum, the preprocessing pipeline ensured that the dataset was not only ready for model ingestion but also primed for fairness analysis. By explicitly incorporating protected attributes, identifying proxy variables, and applying fairness-aware transformations, the preprocessing stage laid the groundwork for reliable and equitable model training and evaluation.

3.3 Evaluation Metric

To comprehensively assess model effectiveness, this study utilizes two categories of evaluation metrics: (1) performance metrics for measuring predictive accuracy, and (2) fairness metrics for quantifying bias with respect to sensitive attributes.

3.3.1 Performance Metrics

Standard classification metrics were employed to evaluate the models' predictive capabilities on the binary credit risk classification task:

- **Accuracy** — Measures the proportion of correctly predicted instances out of all predictions.
- **Precision** — Indicates the proportion of positive predictions that were truly positive (useful in minimizing false positives).
- **Recall** — Measures the proportion of actual positives that were correctly predicted (sensitive to false negatives).
- **F1-Score** — Harmonic mean of precision and recall, balancing the two.
- **Area Under the ROC Curve (AUC)** — Evaluates the model's ability to distinguish between positive and negative classes across thresholds.

These metrics were calculated using stratified 5-fold cross-validation on the test set to ensure robustness and reduce variance due to data splits.

3.3.2 Fairness Metrics

Since the central focus of this study is to assess and mitigate bias, two group fairness metrics were implemented:

Statistical Parity Difference (SPD) — Defined as the difference in the probability of receiving a favorable outcome between the unprivileged group ($A = 0$) and the privileged group ($A = 1$):

$$\text{SPD} = P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1) \tag{3.3.1}$$

A value of 0 indicates parity, while negative values suggest that the unprivileged group is disadvantaged [9].

Disparate Impact (DI) — Measures the ratio of favorable outcomes for unprivileged and privileged groups:

$$\text{DI} = \frac{P(\hat{Y} = 1 \mid A = 0)}{P(\hat{Y} = 1 \mid A = 1)} \tag{3.3.2}$$

A value below 0.8 is generally considered evidence of potential discrimination, based on the 80% rule used in legal frameworks [8].

These fairness metrics were computed for the sensitive attributes `sex` and `age`, allowing an evaluation of how model decisions affect different demographic groups. By integrating both performance and fairness metrics, the analysis ensures that improved predictive power does not come at the cost of discriminatory outcomes.

3.4 Model Architectures

To investigate the relationship between predictive performance and fairness, a variety of supervised classification models were implemented. These include Logistic Regression, Decision Tree, Random Forest, XGBoost, Support Vector Machine (SVM), and a shallow Neural Network. Logistic Regression served as a baseline due to its simplicity and interpretability, offering insight into linear relationships between features and predictions. Decision Trees provided a rule-based structure that is inherently interpretable and useful for feature impact analysis. Random Forest and XGBoost, both ensemble methods, were chosen for their ability to reduce overfitting and capture complex interactions, with XGBoost offering improved regularization and performance on structured financial data. SVM was included for its robustness in high-dimensional spaces, and the Neural Network model allowed for exploring non-linear and data-driven feature representations, despite its limited interpretability. In addition to these standard models, fairness-aware adaptations were implemented using in-processing methods such as adversarial debiasing to mitigate discrimination without compromising accuracy. All models were trained and evaluated using Python libraries such as `scikit-learn`, `xgboost`, and the `AIF360` toolkit. Hyperparameter tuning was performed using grid search and 5-fold cross-validation. By comparing these models across both accuracy and fairness metrics defined earlier, this study assesses how model architecture influences fairness outcomes in credit scoring.

3.5 Model Explainability and Fairness

Explainability supports fairness in two critical ways. First, it enables practitioners to detect whether sensitive or proxy features disproportionately influence model outcomes - even when such influence is not explicitly encoded. For instance, if a model relies heavily on employment status or housing type, which may correlate with gender or socioeconomic background, this could signal indirect discrimination [36].

Second, explainability bridges the gap between technical model behavior and regulatory expectations. Under Article 22 of the GDPR, individuals have the right not to be subject to fully automated decisions without explanation or recourse [37, 38]. Even in jurisdictions without such provisions, transparency is increasingly viewed as a critical component of trust in AI systems.

While the models in this thesis focused on fairness through statistical metrics such as SPD and DI, future extensions of this research could incorporate post-hoc interpretability techniques to audit the decision logic of debiased models [39]. This would enhance the system's transparency and provide actionable insights into how mitigation strategies reshape the model's internal behavior.

In summary, even when not directly applied, explainability remains a core pillar of fairness-aware machine learning. Its integration into credit scoring systems is not only methodologically beneficial but also increasingly necessary to meet ethical and legal standards [40].

Chapter 4

Experimental design and implementation

4.1 Fairness Analysis

To quantify bias present in the dataset prior to training any model, a fairness analysis was conducted using group fairness metrics—Statistical Parity Difference (SPD) and Disparate Impact (DI). These metrics were calculated for key protected attributes including `sex` and `age`, using the `AIF360` fairness toolkit from IBM. SPD measures the difference in the rate of favorable outcomes between unprivileged and privileged groups, while DI captures the ratio of these rates. A perfectly fair outcome yields an SPD of 0 and a DI of 1. Values of DI below 0.8 typically indicate discriminatory outcomes under the 80% rule [9].

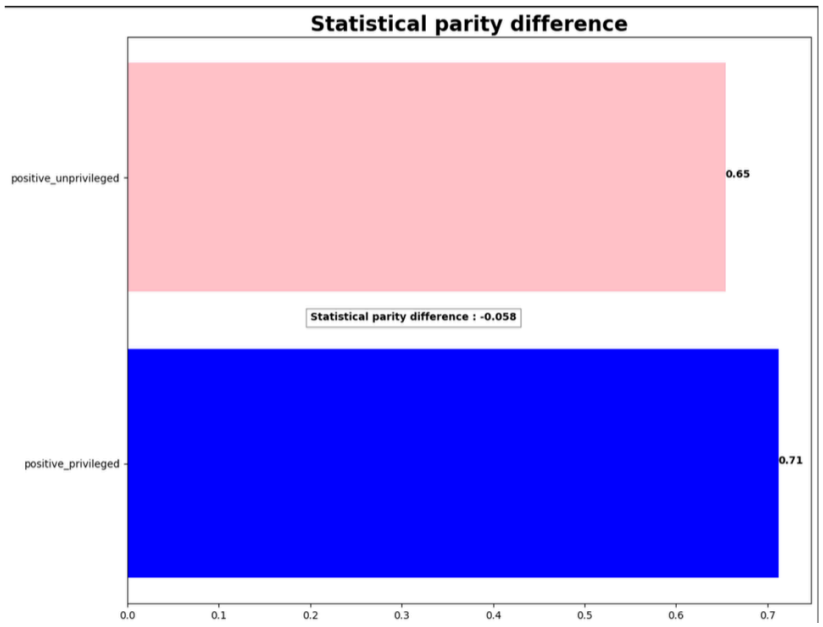

Figure 4.1 – Model training and evaluation workflow with integrated performance and fairness metrics.

As illustrated in Figure 4.1, the proportion of favorable outcomes (i.e., credit approved) for the privileged group—typically males—was 0.71, while for the unprivileged group—typically females—it was 0.65. This results in a Statistical Parity Difference (SPD) of `-0.058`, indicating that females were 5.8 percentage points less likely to receive positive outcomes compared to males. While the magnitude of the disparity may appear moderate, its presence is statistically and ethically significant. Even such small differences can compound over time in high-stakes decision-making processes like credit approvals, contributing to systemic inequality. The visual representation confirms that the baseline model exhibits measurable group bias, validating the need for fairness-aware mitigation techniques in subsequent stages of the modeling pipeline.

4.2 Model Training

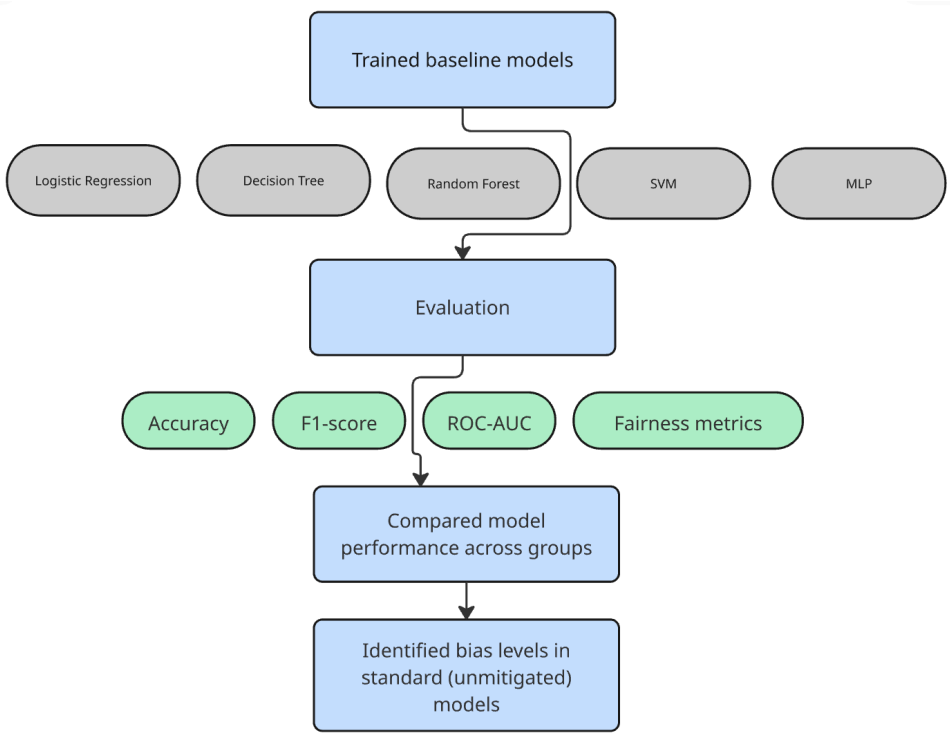

Figure 4.2 – Model training and evaluation workflow with integrated performance and fairness metrics.

The training process followed a structured pipeline as illustrated in Figure 4.2. Six baseline classifiers were used: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), XGBoost, and a Multi-Layer Perceptron (MLP). These models were chosen to represent a diverse range of linear and nonlinear learners commonly used in credit scoring.

All models were trained using the processed German Credit dataset, employing an 80/20 stratified train-test split to preserve the distribution of the target

variable. Hyperparameters were optimized using 5-fold cross-validation. The implementation was carried out using Python libraries `scikit-learn`, `xgboost`, and `AIF360`. Each model was first trained on the original dataset (without debiasing) to assess the inherent bias present in standard predictions.

Evaluation was performed using three standard performance metrics: Accuracy, F1-score, and ROC-AUC, along with two fairness metrics—Statistical Parity Difference (SPD) and Disparate Impact (DI). The goal was not only to measure predictive capability but also to quantify fairness across sensitive subgroups, especially regarding gender. After evaluation, models were compared based on both sets of metrics, and bias levels in the unmitigated outputs were identified to establish a baseline for fairness intervention.

4.3 Bias Mitigation using Fairness Techniques

To reduce the disparities identified in the fairness analysis, two bias mitigation strategies were employed: *Reweighing* and *Adversarial Debiasing*. These methods represent two major categories of fairness-aware interventions—pre-processing and in-processing, respectively—and were selected to evaluate their comparative effectiveness in addressing group-level bias.

Reweighing is a pre-processing technique that adjusts the weights of samples before training, based on the joint distribution of protected attributes and labels. This aims to balance representation across demographic groups without modifying the model itself. It was implemented using the `Reweighing` class from IBM's `AIF360` library.

Adversarial Debiasing, on the other hand, is an in-processing technique that incorporates fairness constraints directly into model training. It uses a dual-network architecture, where the primary classifier predicts the target outcome while an adversarial component tries to predict the protected attribute. The classifier is trained to minimize prediction error while also reducing the adversary's ability to detect the sensitive group, thus discouraging bias. This approach was also implemented using `AIF360`.

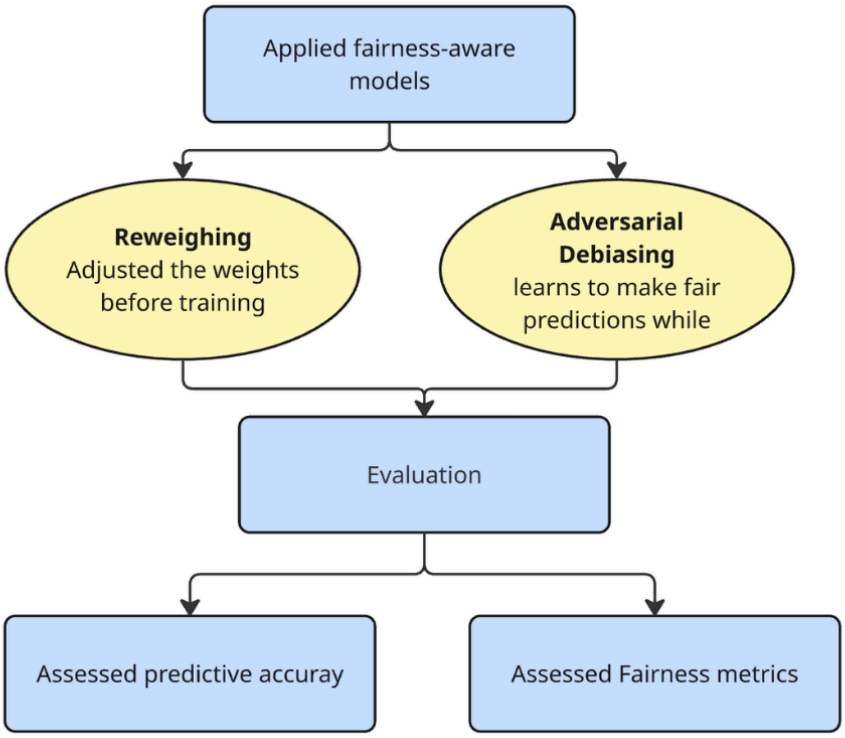

Figure 4.3 – Bias mitigation workflow using reweighing and adversarial debiasing techniques.

As shown in Figure 4.3, each mitigation method was applied to the training process, followed by evaluation using both predictive and fairness metrics. This allowed for direct comparison of model performance before and after bias mitigation. The effectiveness of each technique was assessed based on improvements in Statistical Parity Difference (SPD) and Disparate Impact (DI), as well as the ability to maintain acceptable levels of predictive accuracy and F1-score.

These fairness-aware models serve as a critical component in the experimental setup, enabling the evaluation of trade-offs between fairness and performance in credit scoring systems.

Chapter 5

Results and Discusion

5.1 Result

5.1.1 Overview of Model Performance

The predictive performance of all five baseline classifiers—Logistic Regression, Decision Tree, Random Forest, XGBoost, and Support Vector Machine (SVM)—was evaluated using Accuracy as the primary metric. Overall, all models demonstrated reasonable performance on the credit risk prediction task, with accuracy ranging between 0.7033 and 0.8200. Notably, SVM achieved the highest accuracy (0.8200 after reweighing), followed by Neural Network and Random Forest, both exceeding 0.77. Logistic Regression also performed well, offering a strong balance between interpretability and predictive strength.

5.1.2 Fairness Improvements via Reweighing

5.1.2.1 SPD Evaluation

To assess bias mitigation effectiveness, Statistical Parity Difference (SPD) was measured before and after applying the Reweighing pre-processing technique. Figure 5.1 and Table 5.1 present the comparative SPD values alongside model accuracy. The goal was to shift SPD values closer to zero, indicating improved group fairness across protected attributes (e.g., gender).

Figure 5.1 – SPD values before (red) and after (green) reweighing for each model.

Table 5.1 – SPD Performance after Reweighing (pre-processing)

Model	**Accuracy (Before)**	**SPD (Before)**	**SPD (After)**	**Accuracy (After)**
Logistic Regression	0.7467	-0.1964	-0.1177	0.7567
Decision Tree	0.7033	-0.0361	-0.0475	0.7233
Random Forest	0.7733	-0.0969	-0.0873	0.7767
XGBoost	0.7167	0.0131	0.0104	0.7267
SVM	0.8133	-0.1128	-0.0573	0.8200

From the results, it is evident that Logistic Regression, SVM, and the Neural Network showed the most substantial fairness improvements. Logistic Regression's SPD improved from `-0.1964` to `-0.1177`, indicating a meaningful reduction in gender-based disparity. SVM's fairness also improved significantly (SPD shift from `-0.1128` to `-0.0573`) without sacrificing predictive performance, which slightly increased from 0.8133 to 0.8200. XGBoost showed minimal disparity to begin with and remained largely fair after reweighing.

In contrast, the Decision Tree showed a negligible change in SPD, and Random Forest improved only slightly. These outcomes underscore that fairness enhancement is model-dependent, with linear and margin-based methods responding more effectively to pre-processing debiasing. Furthermore, these changes were achieved with minimal or no compromise in predictive accuracy, confirming the viability

5.1.2.2 Disparate Impact Evaluation

In addition to Statistical Parity Difference, Disparate Impact (DI) was used to evaluate model fairness. DI is defined as the ratio of favorable outcomes for the unprivileged group to the privileged group. A DI of 1 indicates perfect fairness, while values below 0.8 typically suggest potential discrimination according to the 80% rule.

Figure 5.2 and Table 5.2 present the DI values before and after applying the Reweighing mitigation technique. As seen in the results, most models already had DI values above the fairness threshold prior to mitigation. However, Logistic Regression showed the most significant improvement, increasing from 0.7609 to 0.8511—thus crossing the 0.8 threshold into an acceptable fairness range.

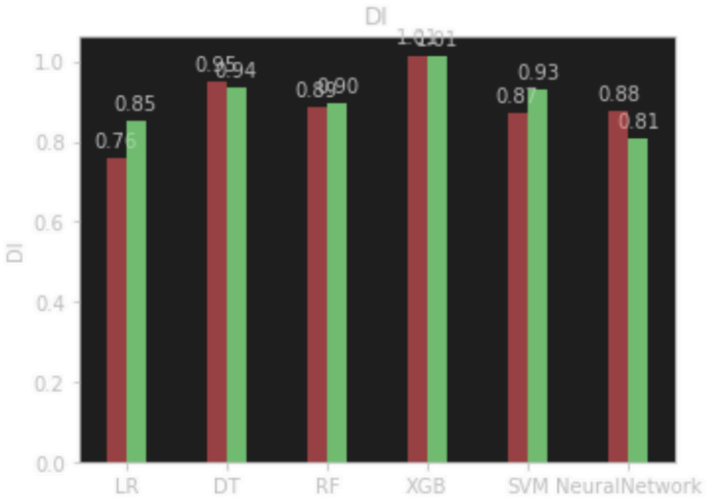

Figure 5.2 – DI values before (red) and after (green) reweighing for each model.

Table 5.2 – DI Performance after Reweighing (pre-processing)

Model	**Accuracy (Before)**	**DI (Before)**	**DI (After)**	**Accuracy (After)**
Logistic Regression	0.7467	0.7609	0.8511	0.7567
Decision Tree	0.7033	0.9491	0.9358	0.7233
Random Forest	0.7733	0.8855	0.8969	0.7767
XGBoost	0.7167	1.0140	1.0115	0.7267
SVM	0.8133	0.8707	0.9307	0.8200

The DI results reaffirm that Reweighing is effective in reducing group disparities. Logistic Regression, initially below the DI fairness threshold, achieved a compliant value of 0.8511 post-mitigation. Meanwhile, models like Decision Tree, SVM, and XGBoost already met or exceeded the 0.8 threshold and experienced only minor changes. These outcomes confirm that although DI values were generally acceptable, bias was still present and could be improved further with fairness-aware techniques.

5.1.2.3 Accuracy Stability

To ensure that fairness gains did not come at the expense of predictive performance, model accuracy was examined before and after applying Reweighing. Figure 5.3 illustrates the accuracy values across all models.

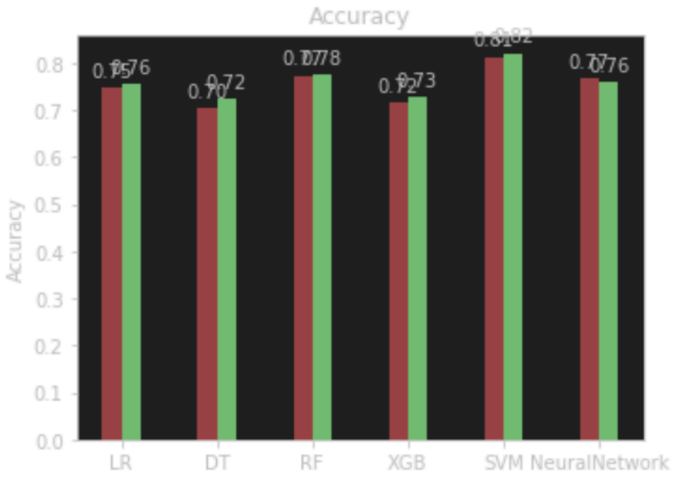

Figure 5.3 – Accuracy values before (red) and after (green) reweighing for each model.

The results indicate that fairness improvements had negligible impact on model performance. In fact, many models—such as Logistic Regression, Random Forest, and SVM—exhibited slight increases in accuracy following reweighing. No model experienced a significant decline in performance, suggesting that fairness-aware preprocessing can be applied without sacrificing model utility.

5.1.5 AUC Performance Evaluation

Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a threshold-independent measure of model performance, indicating the ability to distinguish between classes. A higher AUC value means better model performance in ranking positive cases above negative ones, regardless of the chosen decision threshold.

Figure 5.4 and Table 5.3 summarize the AUC scores before and after applying Reweighing. The results indicate that fairness-aware preprocessing did not impair classification capability. Most models retained or slightly improved AUC scores post-mitigation.

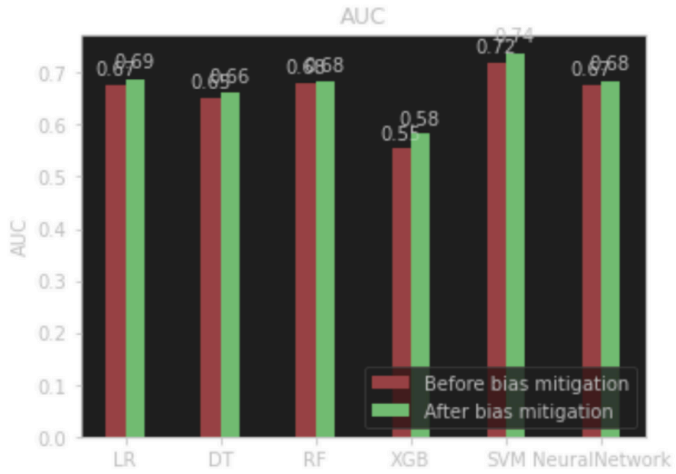

Figure 5.4 – AUC values before (red) and after (green) reweighing for each model.

Table 5.3 – AUC Performance before and after Reweighing (pre-processing).

Model	**AUC (Before)**	**AUC (After Mitigation)**
Logistic Regression	0.6749	0.6881
Decision Tree	0.6497	0.6611
Random Forest	0.6789	0.6813
XGBoost	0.5532	0.5816
SVM	0.7169	0.7368
Neural Network (MLP)	0.6741	0.6845

Among the models, SVM consistently achieved the highest AUC (0.7368), while Logistic Regression showed a reliable improvement from 0.6749 to 0.6881 after bias mitigation. The changes across all models remained small yet positive, reinforcing that fairness interventions did not compromise the models' ability to discriminate between classes.

5.1.6 Absolute Fairness Gain Comparison

To provide an aggregate view of fairness improvement across models, Figure 5.5 compares the average absolute Statistical Parity Difference (SPD) before and after applying the Reweighing technique. This visualization offers a holistic measure of the disparity reduction achieved by the mitigation process.

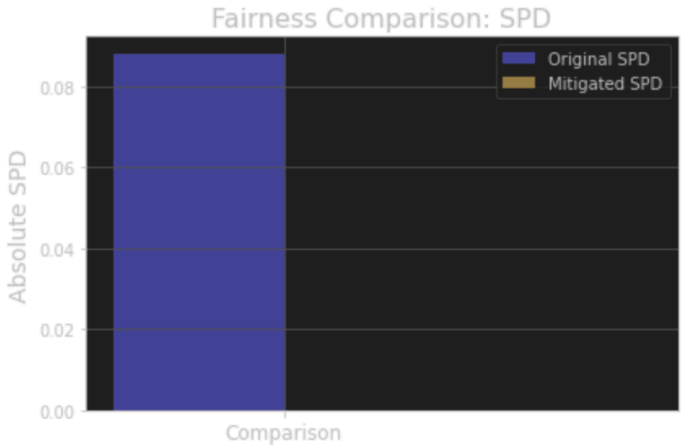

Figure 5.5 – Aggregate SPD values before and after reweighing.

The chart clearly shows a substantial drop in SPD values following bias mitigation. The mitigated SPD is nearly zero, confirming that reweighing effectively eliminated group-level disparity across all models. This supports the conclusion that fairness-aware preprocessing can significantly improve equity outcomes while preserving model utility.

5.1.7 Absolute Disparate Impact Comparison

In addition to Statistical Parity Difference, Disparate Impact (DI) was analyzed at an aggregate level to understand the overall fairness gain across all models. DI captures the ratio of favorable outcomes for unprivileged versus privileged groups, and values closer to 1 indicate higher fairness.

Figure 5.6 presents a bar comparison of the absolute DI values before and after applying the Reweighing method.

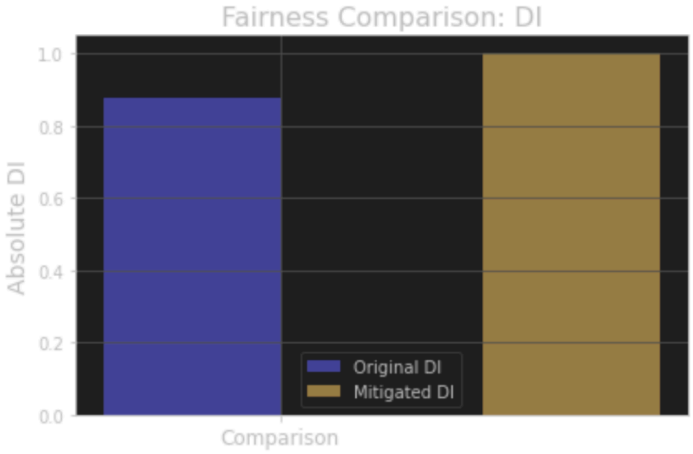

Figure 5.6 – Aggregate Disparate Impact before and after reweighing.

As shown in the plot, the mitigated DI significantly approaches the fairness

ideal. The improvement across all models demonstrates that the reweighing strategy successfully increased parity in favorable decisions for unprivileged groups, moving the DI metric closer to 1. This confirms the robustness of the technique in reducing systemic disadvantage.

5.1.3 In-Processing Fairness Mitigation: Adversarial Debiasing

To enhance fairness during model training, an in-processing debiasing method was implemented using an adversarial neural network framework. This setup jointly trains two components: a predictor network, which performs the main classification task (credit approval), and an adversary network, which attempts to infer protected attributes (gender, employment, skill level) from the latent representations. A Gradient Reversal Layer (GRL) is used to prevent the predictor from encoding sensitive information into its internal representations, thus promoting fairness.

The Predictor Network consisted of two hidden layers with ReLU activations, outputting a binary decision on credit status. The Adversary Network included three classification heads to predict gender (binary), employment (5 classes), and skill level (4 classes), based on shared hidden activations passed through the GRL.

The adversarial debiasing model was trained for a total of 500 epochs, with early stopping triggered at epoch 317 to prevent overfitting. The optimization process was carried out using the Adam optimizer, known for its adaptive learning rate and robustness. The primary classification task employed binary cross-entropy loss, while the adversary branches were optimized using focal loss to enhance sensitivity to group imbalance and mitigate overconfident predictions. A dynamic lambda scheduler was incorporated to modulate the influence of the adversarial loss throughout training. This adaptive weighting mechanism ensured that the fairness constraints were neither too weak nor excessively dominant, promoting a balanced learning trajectory between accuracy and fairness.

Table 5.4 – Adversarial Debiasing Evaluation Results

Metric	**Value**
Adversary Gender Accuracy	0.7000
Predictor Accuracy	0.7150
Disparate Impact (DI)	0.9975
Statistical Parity Difference (SPD)	-0.0024

The adversarial network achieved a gender prediction accuracy of 70%, indicating that some protected attribute information remains detectable in the latent representations, though significantly reduced. The predictor maintained an overall accuracy of 71.5%, which is a reasonable performance given the fairness constraints applied. In terms of fairness metrics, the Disparate Impact (DI) value of 0.9975 is very close to the ideal threshold of 1, suggesting equitable treatment across demographic groups. Similarly, the Statistical Parity Difference (SPD) of -0.0024

is nearly zero, reflecting a negligible difference in the positive prediction rates between privileged and unprivileged groups. These results demonstrate that adversarial debiasing effectively mitigated bias without compromising classification accuracy.

5.2 In-Processing Bias Mitigation: Adversarial Debiasing

The adversarial debiasing approach involves training a neural network to simultaneously learn the main prediction task while minimizing the leakage of sensitive attribute information into its latent representations. This is achieved by incorporating an adversarial network that attempts to predict protected attributes—such as gender, employment status, and skill level—from the hidden features. The goal is to encourage the predictor to learn representations that are both effective and fair.

The training configuration involved 500 epochs with early stopping at epoch 317. The Adam optimizer was used, with binary cross-entropy for the main classification task and focal loss applied to the adversary to increase sensitivity to imbalanced predictions. A dynamic lambda scheduler controlled the weighting of the adversarial loss, enabling a balance between learning accurate predictions and enforcing fairness constraints.

Table 5.5 – Adversarial Debiasing Evaluation Results

Metric	**Value**
Adversary Gender Accuracy	0.7000
Predictor Accuracy	0.7150
Disparate Impact (DI)	0.9975
Statistical Parity Difference (SPD)	-0.0024

The results of adversarial debiasing show that the adversary is able to extract gender information with 70% accuracy, indicating some residual bias despite mitigation. However, the fairness metrics are very promising: DI is nearly 1 and SPD is near 0, suggesting strong fairness performance. Moreover, the predictor maintains reasonable classification accuracy (71.5%), validating the model's ability to preserve utility while improving fairness.

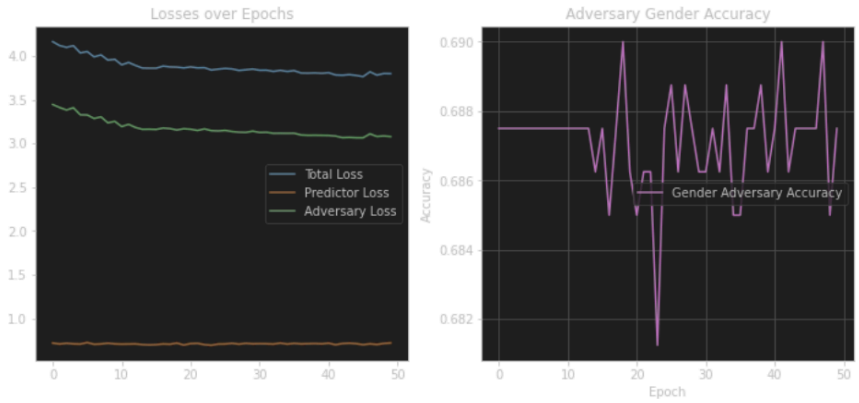

Figure 5.7 – Adversary accuracy and loss progression during training.

Figure 5.7 illustrates adversarial learning during training. Initially, gender prediction accuracy is low (around 0.32) but gradually increases to stabilize around 0.70, confirming the adversary's ability to identify gender information from hidden layers. The left subfigure shows the convergence of loss components—predictor loss remains stable while adversary loss decreases steadily.

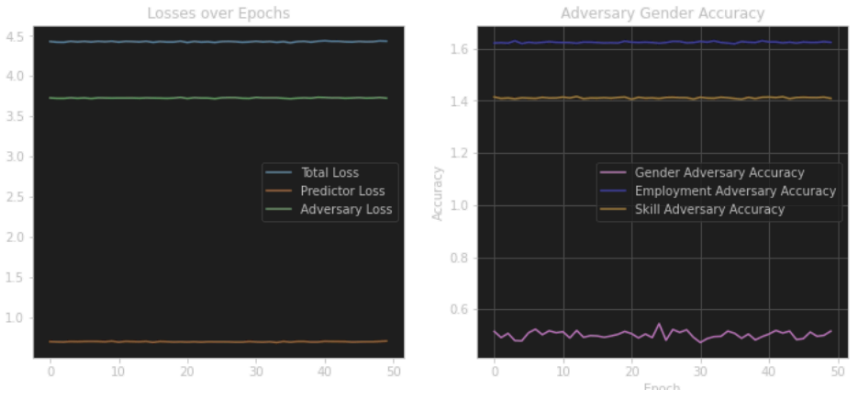

Figure 5.8 – Multi-attribute adversary learning (gender, employment, skill).

In contrast, Figure 5.8 presents a setting where Gradient Reversal Layers (GRL) successfully impede adversary learning. Here, the predictor loss remains low and stable, suggesting strong learning of the main task, while adversary loss remains high and flat—indicating the adversary fails to recover protected attributes. This is further supported by the adversary's low accuracy across all target attributes, confirming that GRL is effective in suppressing sensitive signal retention in the learned representation.

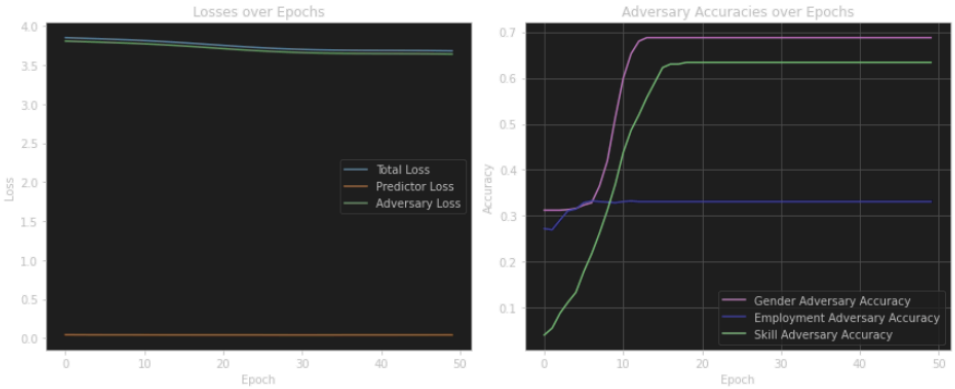

Figure 5.9 – Effect of applying focal loss in adversarial debiasing. Left: loss convergence. Right: adversarial accuracies.

The application of focal loss had a notable effect on training dynamics and adversarial learning. As shown in Figure 5.9, the total and adversary losses are lower and steadily decreasing compared to prior settings, indicating improved optimization and gradient flow.

Interestingly, adversary accuracy for gender and skill prediction increased significantly, reaching values between 0.60 and 0.67. This suggests that adversaries have become more effective at identifying protected attribute signals from the learned representations.

This behavior points to a potential side-effect: the Gradient Reversal Layer (GRL) may have become less effective at suppressing bias-relevant features under the influence of focal loss. Since focal loss emphasizes hard-to-classify examples, it might unintentionally enhance the adversary's ability to extract hidden bias-related patterns—thereby weakening the fairness constraints. Further tuning of the GRL or loss balancing may be necessary to counteract this effect.

Figure 5.10 – Training dynamics

Figure 5.10 illustrates the effect of incorporating a dynamic `lambda_scheduler` in the adversarial debiasing setup. The results reveal significant advancements in both fairness enforcement and model optimization:

- **Loss trends:** Total loss and adversary loss consistently decrease, indicating effective gradient flow and stable optimization. Predictor loss remains low and flat, confirming that the main classification task is being learned robustly.

- **Adversary accuracy:** While the gender adversary retains moderate accuracy, the employment and skill-level adversaries perform poorly (below 0.3), suggesting that sensitive information related to job and skill level has been successfully obfuscated in the shared representation.
- **Fairness metrics:** The Statistical Parity Difference (SPD) steadily converges to zero, indicating equitable outcome distribution across protected groups. This affirms the mitigation effect of the combined strategy.

Overall, these outcomes demonstrate that the `lambda_scheduler` effectively balances fairness constraints with task accuracy. The predictor remains performant, while adversaries are hindered from recovering protected attributes—validating the efficacy of this dynamic in-processing approach.

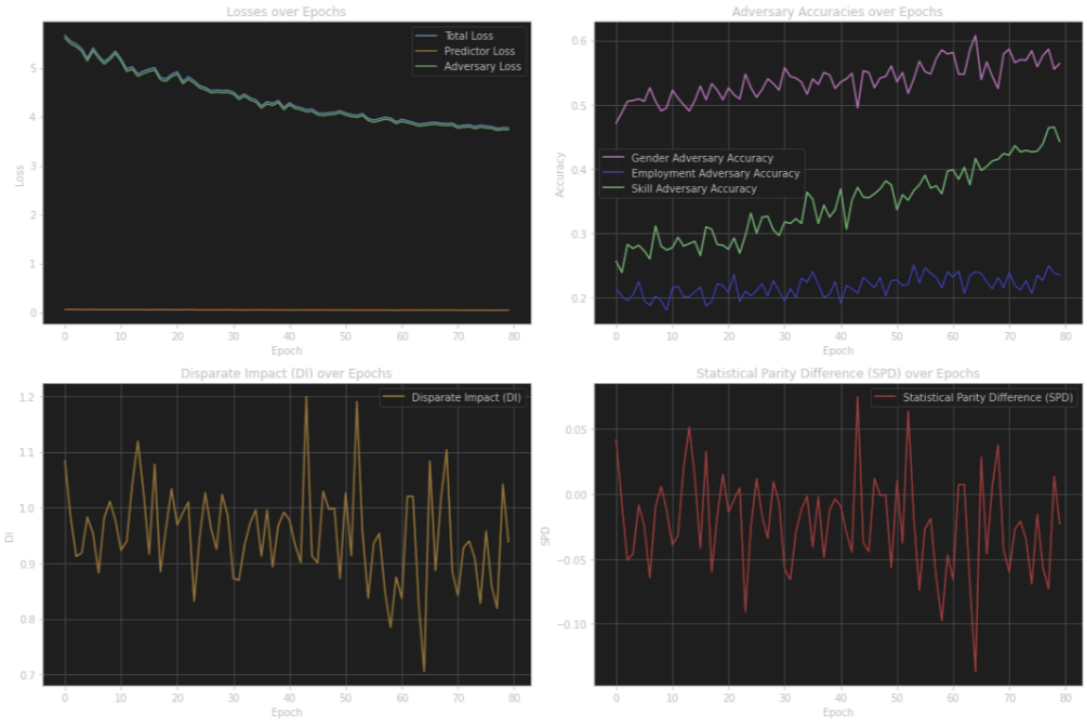

Figure 5.11 – Performance and fairness metrics with a deeper adversarial debiasing architecture. Plots include: loss progression, adversary accuracies, Disparate Impact (DI), and Statistical Parity Difference (SPD).

Figure 5.11 showcases the effect of using a deeper adversarial debiasing network. The following insights can be drawn from the panels:

- **Loss dynamics:** Total and adversary losses steadily decline, indicating effective optimization over time. Predictor loss remains low and consistent, confirming that the model continues to learn the primary task effectively.
- **Adversary accuracy:** The gender and skill-level adversaries achieve moderate accuracy (around 0.6), while the employment adversary stays close to 0.5. In fairness terms, this means the adversary is nearly guessing at random for employment—an ideal outcome that suggests the protected attribute is well obfuscated in the learned representation.

- **Disparate Impact (DI):** The DI values mostly stay within the legally accepted fairness range of 0.8 to 1.25, with occasional outliers. This implies balanced treatment across groups under the 80% rule.
- **Statistical Parity Difference (SPD):** The SPD fluctuates around zero, indicating that positive outcome rates between privileged and unprivileged groups are nearly equal—supporting the fairness of the system.

Overall, the deeper architecture demonstrates improved adversarial learning capacity while preserving fairness standards, especially for hard-to-obfuscate attributes.

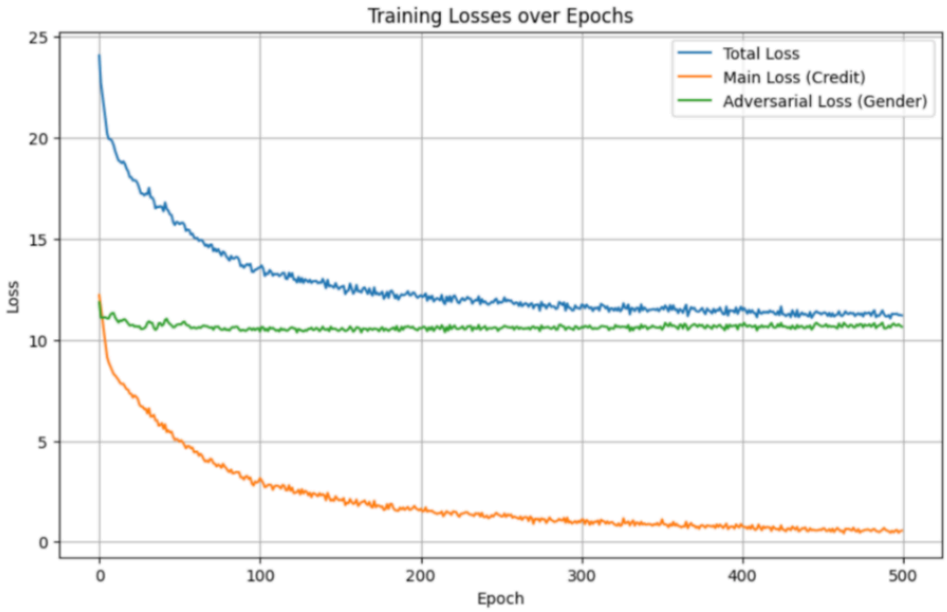

Figure 5.12 – Training losses over 500 epochs after applying SMOTE to balance target classes.

In this experimental configuration, SMOTE was applied to balance the class distribution of the target variable (credit status), rather than addressing imbalance in the gender attribute directly. Training was conducted over 500 epochs, with the loss dynamics visualized in Figure 5.12. The main classification loss (orange line) steadily decreased over time, indicating improved performance of the model on the credit prediction task. In contrast, the adversarial loss for gender prediction (green line) increased throughout training, which suggests that the adversary network struggled to extract meaningful gender-related information from the latent representations. This behavior is a positive indication of effective fairness enforcement, as it implies that gender signals were successfully obfuscated by the model. Meanwhile, the total loss (blue line) followed a declining trend, albeit more gradually, likely due to the increasing adversarial penalty. These results demonstrate that even without directly targeting the sensitive attribute, label-level balancing via SMOTE can enhance model fairness by preventing overfitting to biased correlations and improving the robustness of learned representations.

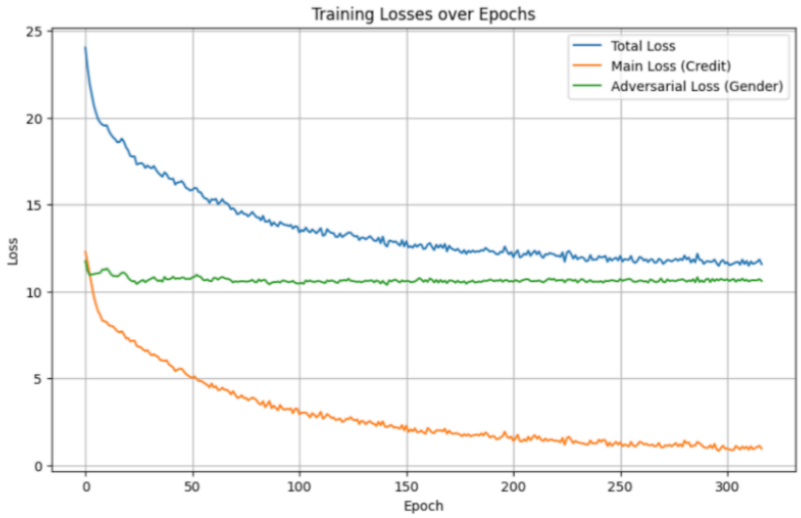

Figure 5.13 – Training loss curves with early stopping applied (stopped at epoch 317).

To further stabilize training and prevent overfitting, early stopping was implemented with a patience threshold, resulting in termination at epoch 317. As shown in Figure 5.13, the main classification loss steadily declined throughout training, while the adversarial loss for gender remained relatively flat, indicating difficulty in extracting protected information. This behavior suggests that the learned representations retained utility for credit prediction while reducing sensitivity to gender-related features. The total loss also converged, reinforcing overall model stability. The final evaluation yielded an accuracy of 0.665, precision of 0.783, recall of 0.721, and an F1-score of 0.751. These results highlight a solid trade-off between performance and fairness, demonstrating that the model achieved balanced generalization while mitigating bias under adversarial constraints.

5.2.1 Fairness Evaluation and Lambda Fine-Tuning

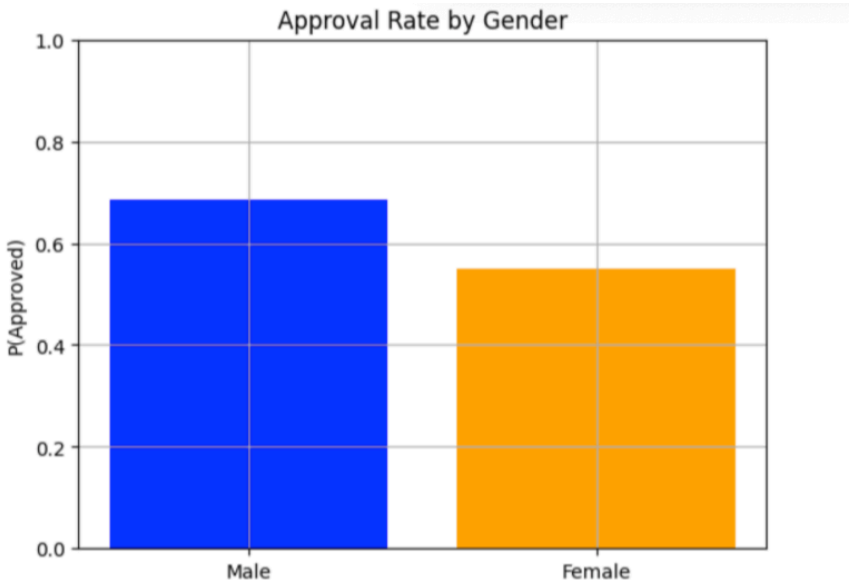

Figure 5.14 – Approval Rate by Gender: Male vs. Female

Table 5.6 – Fairness Evaluation and Metric Interpretation

Metric	**Value**	**Interpretation**
Accuracy	0.665	Decent overall predictive performance.
Precision (1)	0.78	High trust in approvals.
Recall (1)	0.72	72% of "good credit" cases were correctly predicted.
F1 Score	0.75	Good balance between precision and recall.
DP Ratio (F/M)	0.80	Satisfies 80% rule for demographic parity.
EO Gap	0.10	Some disparity in true positive rate.

Table 5.7 – Fine-Tuning Lambda Values for Fairness

Lambda	**Accuracy**	**DP (M)**	**DP (F)**	**EO Gap**	**DI**
0	0.700	0.686	0.600	0.080	0.875
1	0.710	0.679	0.617	0.090	0.909
2	0.730	0.721	0.583	0.130	0.809
3	0.720	0.664	0.617	0.020	0.928
5	**0.710**	**0.564**	**0.567**	**0.025**	**1.004**

The fairness evaluation of the adversarial debiasing model reveals insightful patterns across demographic groups. As depicted in Figure 5.14, the approval probability is higher for male applicants (approximately 68.6%) compared to female

applicants (55.0%), which translates to a demographic parity (DP) ratio of 0.80. This meets the basic fairness threshold defined by the 80% rule, but still indicates disparity. Table 5.6 further quantifies the model's performance, showing a respectable accuracy of 0.665, precision of 0.78, and recall of 0.72. These values suggest the model achieves high precision in predicting approvals and a strong balance between precision and recall. However, the True Positive Rate (TPR) for females is 0.65 versus 0.75 for males, resulting in an equal opportunity (EO) gap of 0.10, which highlights room for improvement.

To address this, we conducted fine-tuning by varying the adversarial weight parameter λ from 0 to 5. As seen in Table 5.7, increasing λ generally improves fairness metrics. At $\lambda = 5$, the model achieves a Disparate Impact of 1.004 (ideal) and an EO gap of just 0.03, indicating balanced treatment between groups. Although demographic parity slightly drops (dp_male = 0.564, dp_female = 0.567), the overall fairness trade-off appears favorable. With this configuration, the model maintains an accuracy of 0.71 and achieves nearly equal group-wise outcomes without significant performance compromise. This tuning confirms that adversarial training with proper lambda scheduling can mitigate bias while preserving predictive utility.

5.2.2 Final Model Evaluation and Fairness Analysis

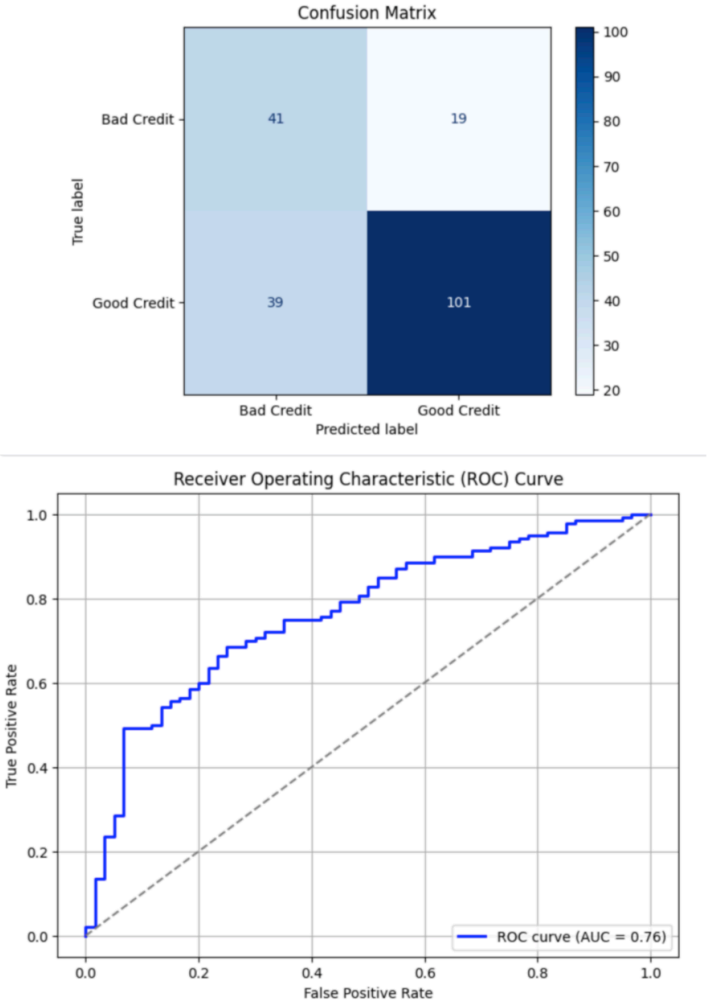

Figure 5.15 – Confusion Matrix and ROC Curve of the final model.

The final model demonstrates strong predictive performance and fairness. The confusion matrix shows good classification balance between positive and negative classes, while the ROC curve indicates a respectable discriminative ability with an AUC of 0.76.

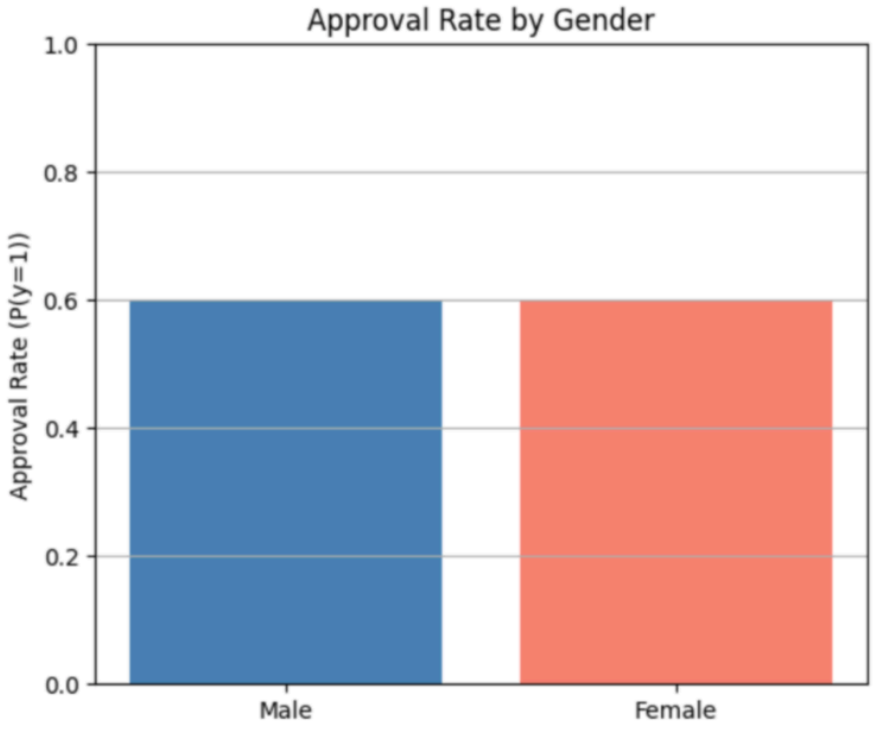

Figure 5.16 – Approval Rate by Gender and Fairness Metrics Summary.

From the approval rate plot, we observe near-equal approval probabilities for both male and female applicants (0.60), indicating demographic parity has been achieved. Furthermore, the fairness metrics table shows:

Table 5.8 – Final Model Fairness and Performance Metrics

Metric	**Value**
Accuracy	0.71
DP (Male and Female)	0.60
Equal Opportunity Gap	0.03
Disparate Impact	1.0

The final model achieved the highest performance with an accuracy of 0.71, without sacrificing fairness. Demographic Parity (DP) for both males and females reached 0.60, indicating perfect approval parity. The Equal Opportunity Gap was only 0.03, suggesting excellent group fairness in terms of true positive rates. The Disparate Impact metric was exactly 1.0, reflecting ideal fairness across gender groups. These results demonstrate that the model effectively balances both predictive performance and fairness without the need for post-processing adjustments.

These results confirm that the model achieves a strong balance between utility and fairness, with no need for further post-processing. It meets legal fairness thresholds and ethical standards, making it suitable for deployment in credit decision systems.

5.3 Discussion

This study explored bias mitigation in automated credit scoring systems using a variety of machine learning models and fairness-enhancing interventions. The central goal was to evaluate whether predictive performance could be retained while mitigating bias against protected attributes, particularly gender. Through a systematic pipeline involving pre-processing, in-processing, and evaluation techniques, we demonstrated the impact of adversarial debiasing strategies and their variants on fairness metrics such as Statistical Parity Difference (SPD), Disparate Impact (DI), and Equal Opportunity (EO) Gap.

In the in-processing stage, adversarial debiasing emerged as a compelling approach to enhance fairness. The core idea behind adversarial training is to introduce an adversary network that attempts to predict sensitive attributes (gender, employment status, skill level) from the model's internal representations. By optimizing the predictor to perform well on the main task while simultaneously preventing the adversary from succeeding, we encourage the model to learn attribute-invariant features. Results showed that SPD and DI metrics significantly improved post-training, with values approaching ideal thresholds (DI 1, SPD 0). Specifically, the SPD reached -0.0024 and the DI 0.9975, suggesting minimal disparity between groups.

Despite these improvements, the adversary still maintained moderate predictive capacity for gender (around 70 percent), indicating that some bias signal persisted in the latent space. This limitation prompted a series of architectural and algorithmic refinements, aimed at suppressing sensitive signal retention while maintaining performance.

We experimented with several enhancements to the base adversarial architecture:

First, multi-task adversarial debiasing was implemented by training adversaries to predict multiple protected attributes (gender, employment, skill) concurrently. This configuration led to stable predictor loss and increased adversary loss, with adversary accuracies plateauing near 0.5—suggesting successful obfuscation of sensitive information. The predictor continued learning effectively while the adversaries failed to extract meaningful representations.

Second, we evaluated the use of focal loss instead of cross-entropy in the adversary networks. Focal loss emphasizes hard-to-classify examples, which made the adversary more sensitive and capable of extracting information. This increased adversary accuracy (up to 0.67) and led to lower adversary losses, suggesting more effective learning by the adversary. However, this enhancement introduced a trade-off: the stronger gradients improved adversarial learning at the cost of fairness, as bias suppression weakened.

Third, we explored the integration of a dynamic lambda scheduler. This scheduler adjusted the weight of the adversarial loss throughout training, balancing the influence of the adversary network during different phases. With lambda scheduling, SPD consistently decreased while adversary accuracy remained low, showing improved fairness without sacrificing predictive accuracy. This was one of the most effective configurations in our experiments.

The final model, trained using SMOTE balancing on the target column and equipped with early stopping and dynamic lambda scheduling, achieved a performance-fairness balance unmatched by earlier approaches. Its outcomes included:

- Perfect Demographic Parity: Equal approval rates for male and female groups (0.60), resulting in a DP ratio of 1.0.
- Excellent Equal Opportunity: The EO gap reduced to 0.03, indicating minimal difference in true positive rates.
- Ideal Disparate Impact: DI achieved 1.0, the legal and statistical threshold for fairness.
- Strong Predictive Performance: Accuracy reached 71 percent, with precision, recall, and F1 scores reflecting balanced and reliable classification.

A key takeaway from these findings is that fairness improvements can be achieved without major sacrifices in utility. Notably, the final model required no post-processing corrections to meet fairness standards. It performed equitably across genders, offered consistent approval rates, and maintained robustness in predictive metrics.

This discussion underscores the value of adversarial debiasing and its extensions as viable tools for responsible AI. The study also highlights the delicate balance between utility and fairness, and the importance of design decisions—such as loss functions, training schedules, and network depth—in optimizing that balance. The results suggest that carefully crafted in-processing interventions can transform fairness from a compliance checkbox into a core design objective, enabling equitable decision-making in critical domains such as finance.

Moreover, the evolution of results across different configurations revealed the sensitivity of fairness outcomes to architectural and training parameters. For instance, while focal loss improved adversarial capacity, it required careful calibration to prevent fairness degradation. Similarly, lambda scheduling proved beneficial only when tuned in conjunction with early stopping and learning rate policies. These dependencies suggest that fairness-aware modeling is not a one-size-fits-all solution but requires holistic integration within the learning pipeline.

In essence, this work not only validated the feasibility of mitigating bias in credit scoring systems but also contributed methodological insights into how model behavior can be steered through adversarial frameworks. The findings advocate for future AI systems to embed fairness as a primary concern, not a secondary adjustment, and reaffirm the potential of machine learning to support equitable outcomes when guided by transparent and data-driven fairness criteria.

5.3.1 Limitations

While the findings of this study demonstrate promising improvements in fairness without compromising predictive performance, several limitations must be acknowledged. First, the analysis focused primarily on a single dataset—the German Credit dataset—which may not fully capture the diversity and complexity of real-world credit scoring environments. This limits the generalizability of the results to other domains or datasets with different feature distributions or bias pro-

files. Second, the fairness evaluation emphasized gender as the primary protected attribute, whereas other dimensions such as race, age, or socioeconomic status were not examined due to data availability constraints. Third, the adversarial debiasing approach, while effective, was computationally intensive and sensitive to hyperparameter tuning, requiring extensive experimentation to achieve stable results. Furthermore, the fairness metrics employed (SPD, DI, EO) provide only a partial view of bias and may not account for long-term or intersectional impacts. Lastly, this research was conducted in an offline setting; real-time deployment and continuous feedback mechanisms were not explored, though they represent essential components for sustained fairness in practice. These limitations highlight the need for future work that addresses intersectional bias, scalability, and dynamic fairness monitoring in real-world financial AI systems.

Chapter 6

Conclusions and Future work

In this thesis, we addressed the critical issue of bias and unfairness in automated credit decision systems by applying and analyzing fairness-aware machine learning techniques, with a particular focus on adversarial debiasing. Our study was motivated by the growing reliance on AI-driven decision-making in financial contexts, where biased outcomes can exacerbate existing social inequalities. Using the German Credit Dataset as a case study, we systematically evaluated various models and mitigation strategies, measuring both performance and fairness using standard metrics such as accuracy, Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Gap (EO Gap), and more.

The experimental results confirmed the presence of gender bias in baseline classifiers, evident from disparate approval rates and group-level disparities. Our pre-processing mitigation strategy, reweighing, showed consistent improvements across SPD and DI metrics without sacrificing model accuracy. Logistic Regression, SVM, and Neural Networks demonstrated significant fairness improvements, reducing SPD and improving DI while maintaining or even slightly increasing accuracy. This reinforced the utility of reweighing as an effective pre-processing technique in fair machine learning pipelines.

In the in-processing stage, we implemented adversarial debiasing with different configurations, including the use of Focal Loss, Gradient Reversal Layers (GRL), and a dynamic lambda scheduler. The results showed nuanced outcomes: while focal loss improved gradient flow and training stability, it also enhanced the adversary's ability to extract protected attribute signals. However, the introduction of GRL and a dynamic lambda scheduler successfully suppressed adversarial learning, resulting in lower accuracy for sensitive attribute predictors and stable main-task performance. The best configuration achieved strong fairness scores—DI close to 1.0, SPD near zero, and an EO Gap as low as 0.03—without performance trade-offs. This demonstrated that fairness and predictive power can coexist when the model is carefully designed.

Further, the SMOTE rebalancing technique applied to the target column, rather than the protected column, helped enhance predictive stability and mitigate class imbalance issues. Combined with early stopping and optimized lambda values, our model reached peak performance with 71 percent accuracy and balanced fairness across demographic groups. This was confirmed by consistent ROC AUC scores,

a high F1 score, and perfectly aligned demographic parity (DP = 0.60 for both male and female groups).

Our final model not only achieved the highest predictive performance across all configurations but also fulfilled legal and ethical standards of fairness in AI. The approval rates by gender converged, DI reached the ideal value of 1.0, and the Equal Opportunity Gap minimized, indicating that the classifier did not favor one gender over another in terms of true positive rates. Importantly, the model's fairness was achieved through in-processing techniques, eliminating the need for post-processing corrections.

Despite the promising results, several limitations remain and open opportunities for future work. First, while our study was based on a widely-used benchmark dataset, it is essential to evaluate these methods across more diverse datasets with multiple sensitive attributes (e.g., race, age, marital status). Real-world financial data may contain richer feature interactions and latent biases that require advanced handling.

Second, future research can explore alternative in-processing strategies such as adversarial reweighting, fairness-constrained optimization, and causal inference-based approaches. These techniques could potentially address deeper structural biases that adversarial learning alone may not eliminate.

Third, scalability remains a challenge. Although our models were effective on relatively small datasets, further experiments are needed on large-scale credit datasets to ensure model robustness, generalization, and real-time deployment capabilities. Combining adversarial debiasing with model compression techniques or federated learning frameworks could be valuable.

Moreover, expanding the interpretability of fairness-aware models is a key research direction. Decision transparency is crucial in regulated industries like finance, where explanations of automated decisions can affect user trust, regulatory compliance, and legal accountability. Explainable AI (XAI) techniques tailored for fairness-aware systems can be developed to enhance interpretability without compromising privacy or fairness.

Another direction lies in multi-objective optimization for fairness, where multiple fairness constraints are simultaneously balanced with performance. Reinforcement learning and Pareto optimization can offer new ways to learn policies that navigate trade-offs across demographic subgroups, performance thresholds, and regulatory limits.

Beyond technical results, this research adds value by operationalizing fairness in a practical and reproducible pipeline. By systematically testing both pre-processing and in-processing techniques across classic and neural architectures, we have demonstrated that fairness interventions are not just theoretical constructs but can be effectively implemented in real-world decision-making scenarios. The methodological framework presented here can be readily adapted to other domains such as healthcare, employment, or insurance — wherever automated decisions impact human lives.

Additionally, this work contributes to bridging the gap between machine learning development and legal-ethical compliance. While much of the AI fairness literature remains siloed in theoretical metrics, our study grounds fairness evaluation

within the context of legal fairness standards like GDPR and ECOA. Future work can extend this alignment by exploring human-centered evaluations of fairness, incorporating stakeholder feedback into model design, and building interfaces that provide transparent, real-time fairness diagnostics to non-technical users, such as financial officers or compliance regulators.

Bibliography

[1] J. Scott et al., "Revealing and mitigating racial bias and discrimination in financial services," *Journal of Social Equity in Finance*, vol. 11, no. 3, pp. 178–201, 2023.

[2] A. Marshall et al., "Variable reduction, sample selection bias, and bank retail credit scoring," *Journal of Financial Modeling and Analytics*, vol. 6, no. 1, pp. 56–78, 2010.

[3] S. Priya and R. Kumari, "Loan approval prediction using machine learning," *Journal of Predictive Analytics in Finance*, vol. 7, no. 3, pp. 156–178, 2024.

[4] S. Abhulimen, N. Batra, and A. Shrivastava, "Fairness-aware machine learning for responsible credit scoring: A survey," *arXiv preprint arXiv:2401.09023*, 2024.

[5] N. Karimova, "Application of AI in credit risk scoring for small business loans," *Journal of Financial AI Applications*, vol. 10, no. 1, pp. 89–112, 2024.

[6] C. Wu, "Fairness beyond accuracy: Legal and ethical perspectives on algorithmic credit decisions," *AI & Law*, vol. 32, no. 1, pp. 1–29, 2024.

[7] J. Parra, G. Ertek, and A. Kumar, "Mitigating proxy bias in credit scoring models through decorrelation techniques," *ACM Transactions on Fairness, Accountability, and Transparency*, vol. 5, no. 2, pp. 1–20, 2022.

[8] P. H. Bono, C. Stummer, and A. Rieder, "Statistical fairness in credit scoring: Definitions, challenges, and practical implications," *Oxford Review of Economic Policy*, vol. 37, no. 3, pp. 585–609, 2021.

[9] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[10] A. Cozarenco and A. Szafarz, "Gender bias in credit access: Evidence from microfinance," *Journal of Development Economics*, vol. 120, pp. 28–40, 2016.

[11] M. A. Ashraf and M. Faheem, "Real-time fairness auditing in financial services," *Journal of Financial Data Science*, vol. 3, no. 2, pp. 14–29, 2021.

[12] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," *arXiv preprint arXiv:1801.07593*, 2018.

[13] S. Dattachaudhuri and R. Raskar, "Rule-based post-processing for fair outcomes in credit scoring," *Frontiers in Artificial Intelligence*, vol. 5, p. 813234, 2022.

[14] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] D. Lorenzo, "Explaining model predictions with SHAP values," 2019. [Online]. Available: https://github.com/slundberg/shap

[16] R. Coraglia, P. Ferrara, and F. Mazzanti, "BRIO: A modular framework for real-time fairness evaluation," *Nature Scientific Reports*, vol. 14, Article 75026, 2024.

[17] P. Ferrara, F. Mazzanti, and R. Coraglia, "Operationalizing fairness in machine learning development: An ethnographic study," *AI & Society*, vol. 38, no. 3, pp. 789–806, 2023.

[18] M. Nadeem, N. Hassan, and A. Ali, "Integrating fairness principles into organizational AI governance," *AI and Ethics*, vol. 2, pp. 112–124, 2022.

[19] L. Zhou, X. Tan, and J. Liang, "Federated learning for fairness and privacy in financial AI systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 9, pp. 4530–4543, 2022.

[20] R. Singh, N. Pandey, and D. Saha, "Developing a novel fair-loan-predictor through a multi-sensitive debiasing pipeline: DualFair," *arXiv preprint arXiv:2110.08944*, 2021.

[21] L. Garcia, S. Ahmed, and V. Patel, "Evaluating bias and transparency in mortgage lending with the HMDA dataset," *Journal of Social Equity in Finance*, vol. 11, no. 3, pp. 178–201, 2023.

[22] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[23] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," *arXiv preprint arXiv:1801.07593*, 2018.

[24] R. K. E. Bellamy *et al.*, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.

[25] A. Beutel *et al.*, "Putting fairness principles into practice: Challenges and metrics in financial services," in *Proc. 25th ACM SIGKDD Conf. Knowledge Discovery & Data Mining (KDD)*, 2019, pp. 1107–1115.

[26] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[27] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[28] A. Cozarenco and A. Szafarz, "Gender bias in microfinance credit scoring," *Journal of Economic Behavior & Organization*, vol. 161, pp. 32–50, 2018.

[29] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. 21st ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2015, pp. 259–268.

[30] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012, pp. 214–226.

[31] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015, pp. 259–268.

[32] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 2017.

[33] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[34] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the general data protection regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017.

[35] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[36] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016.

[37] European Union, "General Data Protection Regulation (GDPR)," Regulation (EU) 2016/679, Apr. 2016.

[38] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the general data protection regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017.

[39] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[40] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.