

# Fairness and Bias Analysis in Loan Approval Systems Using Machine Learning Models

J Andrea<sup>1</sup>, Shaila Mary J<sup>2</sup>

<sup>1</sup>Student, Department of Computer Science, Mount Carmel College, Autonomous, Bengaluru, India

<sup>2</sup>Assistant Professor, Department of Computer Science, Mount Carmel College, Autonomous, Bengaluru, India

\*\*\*

**Abstract** – Automated loan approval systems using machine learning are widely used to assess credit risk and support lending decisions. While these models achieve high predictive performance, they may also inherit historical biases, leading to unfair outcomes across demographic groups. This paper presents a fairness-aware framework that jointly evaluates accuracy and algorithmic fairness. A dataset of 20,000 loan applications is used to train Logistic Regression, Decision Tree, Random Forest, and hybrid ensemble models. Fairness is measured using Demographic Parity Difference, Disparate Impact Ratio, and Equal Opportunity Difference. Results show that Random Forest achieves the highest accuracy but exhibits measurable bias. A reweighting-based mitigation approach is applied to develop a Fair Random Forest model, achieving 0.8965 accuracy with reduced disparities in approval rates and true positive rates. Hybrid ensembles provide competitive performance but do not outperform the fairness-aware model. The results demonstrate that fairness can be improved with minimal impact on accuracy, supporting ethically aligned loan approval systems.

**Key Words:** Fairness in machine learning models, algorithmic bias, loan approval, credit scoring, Random Forest, reweighting, disparate impact, ethical AI.

## 1. INTRODUCTION

### 1.1 Background

Machine learning (ML) has become central to credit risk assessment and loan approval, where predictive models estimate the likelihood of loan repayment based on historical application data [1, 2]. Such systems offer scalability, consistency, and improved predictive power compared to manual or rule-based credit scoring. However, ML models trained on historical data can reflect and reinforce existing social and institutional biases, particularly along sensitive attributes such as gender, race, or age [3–5]. In high-stakes domains such as finance, this raises serious ethical and regulatory concerns, including potential violations of anti-discrimination laws and fairness regulations such as the Equal Credit Opportunity Act.

### 1.2 Problem Statement

Conventional loan approval models primarily optimize predictive accuracy or related performance measures, often ignoring fairness considerations. As a result, they may assign systematically different approval probabilities to groups that differ only in protected attributes, leading to disparate treatment or disparate impact. The core problem addressed in this work is:

*How can a loan approval prediction model be designed to maintain high predictive performance while reducing unfair bias across protected demographic groups, with a focus on gender?*

### 1.3 Objectives

This research pursues the following objectives:

- Develop machine learning model for binary loan approval prediction.
- Evaluate models using standard classification metrics (accuracy, precision, recall, F1-score).
- Quantify algorithmic bias using group fairness metrics.
- Apply fairness mitigation via a reweighting technique at training time.
- Compare baseline, hybrid ensemble, and fairness-aware models.
- Identify a model that provides a favorable trade-off between accuracy and fairness.

### 1.4 Scope

- Binary classification of loan approval (approved vs. not approved).
- Fairness analysis primarily with respect to gender as a protected attribute.
- Classical ML models (Logistic Regression, Decision Tree, Random Forest, and ensembles) rather than deep learning.

## 1.5 Paper Organization

Section 2 reviews literature on credit scoring and fairness-aware ML. Section 3 describes the dataset. Section 4 presents preprocessing steps. Section 5 introduces the proposed methodology. Section 6 outlines system implementation. Section 7 details the experimental setup. Section 8 reports results and analysis. Sections 9–12 discuss applications, limitations, conclusions, and future work.

## 2. LITERATURE REVIEW

### 2.1 Machine Learning for Credit Risk and Loan Approval

Traditional credit scoring relies on statistical models such as Logistic Regression due to their interpretability and reliability [1, 2]. More advanced approaches, including Decision Trees, Random Forests, and boosting methods, often achieve better predictive performance on structured financial data [6, 7].

### 2.2 Algorithmic Fairness and Bias

Machine learning models can inherit and amplify biases present in historical data, leading to unfair outcomes across demographic groups [3, 5, 8]. Several fairness definitions exist, including demographic parity and equal opportunity [4, 8]. Tools such as IBM's AI Fairness 360 (AIF360) support fairness evaluation and mitigation [10].

### 2.3 Fairness Metrics in Classification

Fairness metrics measure disparities across groups. Demographic Parity ensures equal approval rates [4], while Disparate Impact evaluates their ratio using the "four-fifths rule" [9]. Equal Opportunity focuses on equal true positive rates across groups [4, 5].

### 2.4 Fairness Mitigation Approaches

Fairness techniques include pre-processing, in-processing, and post-processing methods [3, 8, 11]. Reweighting, a pre-processing approach, assigns sample weights to reduce bias by balancing protected groups and outcomes [10, 11]. This study adopts reweighting due to its simplicity and model-agnostic nature.

### 2.5 Research Gap

Many prior works in credit risk modeling emphasize accuracy and calibration while neglecting fairness metrics and mitigation strategies. Even when fairness is

considered, comprehensive comparisons between baseline, ensemble, and fairness-aware models in a single loan approval pipeline are limited. This work addresses this gap by systematically:

- Evaluating multiple models on both performance and fair-ness.
- Applying reweighting-based mitigation for a widely used ensemble model (Random Forest).
- Comparing fairness-aware and hybrid ensembles in a unified experimental framework.

## 3. DATASET DESCRIPTION

### 3.1 Data Source and Structure

The dataset consists of 20,000 historical loan application records from a publicly available credit dataset (e.g., Kaggle/UCI-style loan prediction data) [12]. Each record includes demographic, employment, and financial attributes, along with a binary target indicating whether the loan was repaid (loan paid back: 1) or not (0).

### 3.2 Features and Target Variable

Key features include:

- **Demographic:** gender (protected attribute), age.
- **Financial:** income, loan amount, credit-related variables.
- **Employment:** employment status and related indicators.

The target variable is a binary label: loan paid back (1) vs. not paid back (0), used to approximate loan approval behavior.

### 3.3 Protected Attribute

Gender is treated as the primary protected attribute due to its historical relevance in discrimination within lending systems [3, 5]. Attributes such as income and credit score, although potentially correlated with socio-economic status, are considered legitimate risk factors directly related to loan repayment ability and are therefore not treated as protected attributes.

In addition to gender, other demographic attributes are also analyzed to examine potential bias in model predictions. The protected attributes analyzed include:

- Gender
- Age
- Marital Status
- Education Level

Bias detection and fairness evaluation metrics are computed separately for each of these attributes to assess disparities in model outcomes across different demographic groups.

Furthermore, an intersectional analysis is performed by combining multiple protected attributes into a single composite feature. This enables the evaluation of how overlapping demographic characteristics impact fairness and model decisions.

While multiple attributes are analyzed for bias detection, fairness mitigation using reweighting is applied primarily with respect to gender. This ensures methodological clarity while still providing a comprehensive fairness assessment across multiple dimensions.

## 4. DATA PREPROCESSING

### 4.1 Data Cleaning

Data cleaning steps include:

- Handling missing and null values through imputation or row removal where necessary.
- Resolving inconsistent or out-of-range values.
- Ensuring that data types and formats are consistent across features.

### 4.2 Data Transformation

Numerical features are standardized (e.g., using z-score scaling) to support algorithms sensitive to feature magnitude (e.g., Logistic Regression). Categorical variables (excluding the protected attribute for fairness analysis) are encoded using appropriate schemes such as one-hot encoding or ordinal encoding.

### 4.3 Feature Engineering and Selection

Features relevant to loan repayment and not introducing target leakage are retained. Protected attributes (e.g., gender) are excluded from model inputs when appropriate but retained for fairness evaluation. However, indirect proxy variables may still encode similar information. *Addressing such latent correlations requires advanced methods such as causal analysis or adversarial debiasing, which are beyond the scope of this study.*

### 4.4 Train-Test Split

The dataset is split into 80% training and 20% testing data using stratified sampling to preserve the class distribution. The training set is used for model fitting, hyperparameter

tuning, and fairness-aware training; the test set is used solely for final evaluation.

## 5. METHODOLOGY

### 5.1 Overall Pipeline

The proposed pipeline consists of the following stages:

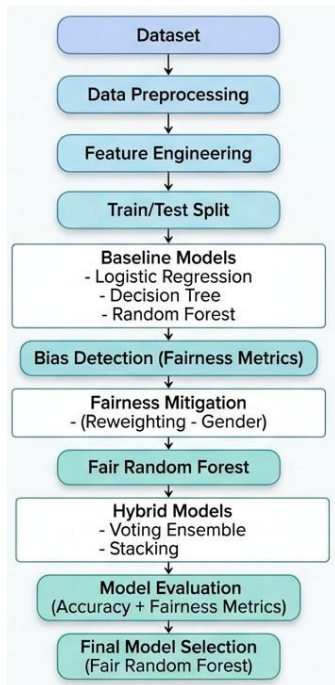
1. Data preprocessing and feature engineering.
2. Training baseline ML models.
3. Quantifying bias using fairness metrics on baseline models.
4. Applying reweighting-based fairness mitigation.
5. Training fairness-aware models (Fair Random Forest).
6. Training hybrid ensemble models.
7. Evaluating all models on performance and fairness.
8. Selecting the model that best balances accuracy and fairness.

### 5.2 System Architecture

The proposed system architecture for loan approval prediction ensures both high predictive performance and fairness through an integrated pipeline. It includes data preprocessing, bias detection, fairness mitigation, and model evaluation. The system operates in the following stages:

- **Input Layer:** Raw loan application data is collected and fed into the system.
- **Preprocessing Layer:** Data cleaning, encoding of categorical variables, and feature scaling are performed to prepare the dataset for modeling.
- **Protected Attribute Identification:** Sensitive attributes such as gender are identified for fairness analysis.
- **Model Training Layer:** Baseline models and fairness-aware models are trained using processed data.
- **Bias Detection Layer:** Fairness metrics are computed to identify disparities in predictions across different demo-graphic groups.
- **Fairness Mitigation Layer:** A reweighting-based technique is applied during training to assign weights to samples to reduce bias.
- **Evaluation Layer:** Models are evaluated based on both performance and fairness metrics.

- **Output Layer:** The final model (Fair Random Forest) is selected for prediction.



**Fig - 1:** Input-Process-Output (IPO) Based Fairness-Aware Loan Prediction System Architecture

This architecture embeds fairness throughout the pipeline rather than treating it as a post-processing step. Bias is detected using fairness metrics and reduced through reweighting, while hybrid models are explored for performance. The final model is selected based on a balance between accuracy and fairness.

### 5.3 Baseline Models

Three primary baseline classifiers are used:

**Logistic Regression:** a linear, interpretable model often used as a baseline in credit scoring [1, 2].

**Decision Tree:** a tree-based model that produces human-readable decision rules.

**Random Forest:** an ensemble of decision trees constructed via bootstrap sampling and feature subsampling, typically offering strong performance on tabular data [6, 7].

### 5.4 Hybrid Ensemble Models

To explore performance improvements, ensemble methods are considered:

- Voting (Logistic Regression, Decision Tree, Random Forest)
- Voting (Logistic Regression + Random Forest)
- Stacking ensemble

These models combine multiple learners to enhance predictive performance.

### 5.5 Fairness Metrics

The following group fairness metrics, computed for gender groups, are used:

**Demographic Parity Difference:** difference in the rates of favorable outcomes (approvals) across groups.

**Disparate Impact Ratio:** ratio of favorable outcome rates, often compared to the 0.8 threshold [9].

**Equal Opportunity Difference:** difference in true positive rates (TPR) across groups [4].

Smaller absolute values of differences and ratios closer to 1 indicate improved fairness. These fairness metrics help evaluate whether the model treats different demographic groups equitably. While a model may achieve high accuracy, it can still produce biased outcomes if predictions are not evenly distributed across protected groups. Therefore, these metrics are essential to assess the trade-off between model performance and fairness, ensuring that the system does not disadvantage any particular group.

### 5.6 Reweighting-Based Fairness Mitigation

A pre-processing reweighting technique inspired by AIF360 [10, 11] is employed:

1. The joint distribution of protected attribute (gender) and outcome (loan repaid) is estimated.
2. Each sample is assigned a weight inversely proportional to the empirical probability of its (gender, outcome) combination.
3. The classifier (Random Forest) is trained using these sample weights, reducing correlation between gender and pre-dictions.

This produces a Fair Random Forest model that explicitly accounts for group fairness during training.

### 5.7 Evaluation Strategy

All models are evaluated on the held-out test set. Performance is assessed using accuracy, precision, recall, and F1-score. Fairness is measured via the metrics above. Comparison focuses on:

- Baseline Random Forest vs. Fair Random Forest.
- Baseline Random Forest vs. hybrid ensembles.

- Accuracy–fairness trade-offs across all models.

## 6. SYSTEM IMPLEMENTATION

### 6.1 Development Environment

The system is implemented in Python using:

- Scikit-learn for ML models and metrics [6].
- Pandas and NumPy for data handling.
- Matplotlib and Seaborn for visualization.

Optionally, AIF360 or similar libraries for fairness metrics and reweighting [10].

## 7. EXPERIMENTAL SETUP

### 7.1 Hardware and Software

Experiments are run on a standard workstation with:

- CPU: Intel Core i3.
- RAM: 8–16 GB.
- Storage: SSD-based system.

Software environment:

- Programming Language: Python
- Environment: Jupyter Notebook and Anaconda
- Libraries:
  - Scikit-learn (model training and evaluation)
  - Pandas and NumPy (data processing)
  - Matplotlib and Seaborn (data visualization)
  - Jupyter (model persistence)

### 7.2 Performance Metrics

The following standard classification metrics are computed to evaluate model performance:

**Accuracy:** Measures the overall proportion of correctly classified instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

**Precision:** Measures the proportion of predicted positive in-stances that are actually positive.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

**Recall:** Measures the proportion of actual positive instances that are correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

**F1-score:** Harmonic mean of precision and recall, providing a balance between the two.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

### 7.3 Fairness Metrics

For each model, fairness metrics across gender groups are computed as described in Section 5. Approval rates and confusion matrices are also analyzed per group to provide interpretability.

## 8. RESULTS AND ANALYSIS

### 8.1 Baseline Performance

Among baseline models, Random Forest achieves the highest accuracy (around 0.89+), outperforming Logistic Regression and Decision Tree, consistent with prior work on credit risk prediction [6, 7]. The baseline Random Forest thus serves as the primary reference for subsequent comparisons.

### 8.2 Fairness Evaluation Before Mitigation

When evaluated on the test set, the baseline Random Forest model exhibits differences in approval rates and true positive rates across gender groups. The representative approval rates are:

Male: 0.8937

Female: 0.8768

Other: 0.8857

These values correspond to encoded categories (Male = 0, Female = 1, Other = 2). The fairness metrics – Demographic Parity Difference ( $\approx 0.0169$ ), Disparate Impact Ratio ( $\approx 0.98$ ), and Equal Opportunity Difference ( $\approx 0.0107$ ) – indicate the presence of mild but non-negligible bias. Specifically, one group receives slightly lower approval rates and reduced true positive rates compared to others.

### 8.3 Fairness-Aware Random Forest

After applying the reweighting-based mitigation technique during training, the Fair Random Forest model achieves an accuracy of 0.8965, which is comparable to the baseline Random Forest. The approval rates across gender groups become more aligned:

Male: 0.8922

Female: 0.8805

Other: 0.8857

As before, these correspond to encoded categories (Male = 0, Female = 1, Other = 2). The reduction in variation across groups reflects improved fairness. The fairness metrics move closer to their ideal values, with lower Demographic Parity and Equal Opportunity differences and a Disparate Impact Ratio closer to 1. This shows the reweighting approach effectively reduces bias while maintaining comparable predictive performance.

### 8.4 Hybrid Ensemble Models

Hybrid models achieve strong performance but do not surpass the fairness-aware Random Forest:

- Voting (Logistic + Random Forest): accuracy  $\approx$  0.89475
- Stacking: accuracy  $\approx$  0.89575

While these models perform well, they exhibit slightly higher bias. This indicates that increased model complexity does not necessarily improve fairness.

**Table - 1: Comparison of All Models**

| Model                    | Accuracy | Observation                |
|--------------------------|----------|----------------------------|
| Random Forest            | 0.8975   | Best baseline performance  |
| Hybrid 1 (Voting - All)  | 0.8930   | No significant improvement |
| Hybrid 2 (Logistic + RF) | 0.89475  | High recall, not superior  |
| Hybrid 3 (Stacking)      | 0.89575  | Slight bias observed       |
| Fair Random Forest       | 0.8965   | Best fairness performance  |

The results indicate that while hybrid models perform competitively, they do not outperform the original Random Forest model in terms of overall accuracy. Although some hybrid approaches improve recall and maintain strong performance, they do not provide a significant advantage over the baseline. This suggests that model complexity does not necessarily guarantee better results, especially when the base model is already highly optimized.

### 8.5 Confusion Matrix and Error Analysis

The Fair Random Forest’s confusion matrix indicates:

Low false negatives and high true positives, important for minimizing denial of credit to truly creditworthy applicants.

Balanced trade-off between false positives and false negatives across groups, reducing disparate misclassification burdens.

### 8.6 Comparative Summary

The Random Forest models provide the best overall performance. The Fair Random Forest:

- Maintains high accuracy (0.8965)
- Reduces bias across demographic groups
- Offers the best balance between accuracy and fairness. This makes it suitable for fairness-sensitive application.

### 9. APPLICATIONS

The proposed framework can be applied in:

- Automated loan approval systems.
- Credit risk assessment platforms.
- Financial systems requiring fairness compliance.

### 10. LIMITATIONS

- Fairness evaluated mainly with respect to gender; intersectional fairness is not addressed.
- Use of a single dataset may limit generalization.
- Only reweighting is explored; advanced methods may further improve fairness.
- Proxy variables for gender may still exist

### 11. DISCUSSION

**Table - 2: Final Model Comparison: Accuracy vs. Fairness**

| Model              | Accuracy       | Fairness           |
|--------------------|----------------|--------------------|
| Random Forest      | Highest        | Moderate           |
| Fair Random Forest | Slightly lower | Best               |
| Stacking Hybrid    | High           | Slightly more bias |

The Fair Random Forest model provides the best balance between accuracy and fairness. While the baseline Random Forest achieves slightly higher accuracy, it exhibits more bias. The fairness-aware model reduces disparities with minimal impact on performance, making it more suitable for ethical decision-making.

#### Reasons for Selection:

- Accuracy remains very high (approximately 0.8965).
- Lowest demographic bias among all models.
- Incorporates fairness mitigation using reweighting.
- Strong justification for ethical AI deployment.

## 12. CONCLUSIONS

This paper presented a systematic fairness and bias analysis of machine learning models for loan approval decision systems. Baseline models, particularly Random Forest, achieved strong predictive performance but exhibited measurable disparities across gender groups. By incorporating reweighting-based fairness mitigation into the training of a Random Forest classifier, a Fair Random Forest model was obtained that significantly reduced these disparities while maintaining an accuracy of 0.8965. Hybrid ensemble models offered competitive performance but did not provide a better accuracy-fairness balance. The results demonstrate that fairness-aware machine learning can meaningfully mitigate bias in credit decision systems without substantial sacrifices in accuracy, supporting the development of more trustworthy and compliant financial AI systems.

## 13. FUTURE WORK

Future research can focus on:

- Extending fairness to multiple protected attributes.
- Exploring advanced fairness mitigation techniques.
- Deploying the model in real-world systems.

## REFERENCES

- [1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.
- [2] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring," *J. Royal Stat. Soc.: Series A*, vol. 160, no. 3, pp. 523–541, 1997.
- [3] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016.
- [4] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3315–3323.
- [5] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness," *arXiv preprint arXiv:1808.00023*, 2018.
- [6] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [7] Y. Xia *et al.*, "A novel tree-based dynamic heterogeneous ensemble method for credit scoring," *Expert Systems with Applications*, vol. 159, p. 113615, 2020.
- [8] S. Verma and J. Rubin, "Fairness definitions explained," in *IEEE/ACM Intl. Workshop on Software Fairness*, 2018, pp. 1–7.
- [9] R. B. Schwab, "The four-fifths rule and legal standards of disparate impact," *Labor Law Journal*, vol. 42, no. 8, pp. 495–502, 1991.
- [10] R. K. N. Bellamy *et al.*, "AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM J. Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.
- [11] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.
- [12] Kaggle, "Loan prediction dataset," [Online]. Available: <https://www.kaggle.com/>. Accessed: Mar. 24, 2026.
- [13] J. N. Crook, D. B. Edelman, and L. C. Thomas, "Recent developments in consumer credit risk assessment," *European J. Operational Research*, vol. 183, no. 3, pp. 1447–1465, 2007.
- [14] C. Dwork *et al.*, "Fairness through awareness," in *Proc. 3rd Innovations in Theoretical Computer Science Conf.*, 2012, pp. 214–226.
- [15] R. Zemel *et al.*, "Learning fair representations," in *Intl. Conf. on Machine Learning*, 2013, pp. 325–333.
- [16] M. Feldman *et al.*, "Certifying and removing disparate impact," in *ACM SIGKDD Conf.*, 2015, pp. 259–268.
- [17] A. Chouldechova, "Fair prediction with disparate impact," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.
- [18] N. Mehrabi *et al.*, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.
- [19] A. Agarwal *et al.*, "A reductions approach to fair classification," in *Intl. Conf. on Machine Learning*, 2018, pp. 60–69.
- [20] T. Calders and S. Verwer, "Three naive Bayes methods for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.
- [21] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *FAT\* Conf.*, 2018, pp. 149–159.
- [22] S. Lessmann *et al.*, "Benchmarking state-of-the-art classification algorithms for credit scoring," *European J. Operational Research*, vol. 247, no. 1, pp. 124–136, 2015.
- [23] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems*,

2017, pp. 5680–5689.

- [24] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, “Counter-factual fairness,” in *Advances in Neural Information Processing Systems*, 2017, pp. 4066–4076.
- [25] C. Corbett-Davies *et al.*, “Algorithmic decision making and the cost of fairness,” in *ACM SIGKDD Conference*, 2017, pp. 797–806.
- [26] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, “The (im)possibility of fairness,” *arXiv preprint arXiv:1609.07236*, 2016.