SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Bias Detection and Fairness Optimization in Algorithmic Systems

Tresa Maria Josylin[1], Tannmay Gupta[2], Budigi Prabhakar[3], Shubhashish Goswami[4], Sivasankari V[5], Gayathri M[6], Dr. S. T. Santhanalakshmi[7], M. A. Prasanna[8]

[1] Department of Computer Science & Engineering, BMS Institute of Technology & Management, India.
Email: tresamjosylin@bmsit.in ORCID: 0009-0009-6143-7074
[2] Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura 140401, Punjab, India. Email: tannmay.gupta.orp@chitkara.edu.in ORCID: 0009-0001-3147-5848
[3] Department of Physics, Noida International University, Greater Noida, Uttar Pradesh 203201, India. Email: budigi.prabhakar@niu.edu.in
[4] Associate Professor, Department of Computer Science & Engineering, Dev Bhoomi Uttarakhand University, Dehradun, Uttarakhand, India. Email: socse.shubhashish@dbuu.ac.in ORCID: 0000-0002-6129-9822
[5] Assistant Professor, Department of Mathematics, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, India. Email: sivasankariv@maher.ac.in
[6] Assistant Professor, Department of Management Studies, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, India. Email: gayathrimba@maher.ac.in
[7] Assistant Professor Grade I, Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, Tamil Nadu, India. Email: santhanalakshmi.peccse2024@gmail.com
ORCID: 0000-0002-1377-1408
[8] Assistant Professor, Department of Computer Science and Engineering, K. Ramakrishnan College of Technology, Tiruchirappalli, Tamil Nadu, India. Email: mapsan1984@gmail.com

Abstract
Algorithmic decision-making systems are widely deployed in critical domains, where embedded biases can lead to systematic discrimination against sensitive groups and reliability concerns. Existing fairness approaches are often stage-specific and fail to provide a unified mechanism that simultaneously detects bias and optimizes fairness without significantly compromising predictive performance. This research offers an integrated model for bias detection and fairness optimization in algorithmic systems to ensure equitable and robust decision-making. A fairness-aware structured dataset of 5,000 samples includes demographic-sensitive attributes, behavioral features, and decision-related variables. Preprocessing is performed using re-weighting and synthetic data balancing (SMOTE) to address class and group imbalance. Feature extraction employs Principal Component Analysis (PCA) to transform features into an uncorrelated space while preserving variance. The proposed Adaptive Gradient-Constrained Fair Logistic Regression (AGCFLR) is designed to minimize bias by enforcing fairness-aware constraints during model training while maintaining predictive performance. It integrates Adaptive Gradient-Based Constrained Optimization (AGCO) to dynamically adjust model parameters under fairness constraints and Logistic Regression (LR) to provide interpretable and efficient classification for decision-making tasks. The proposed AGCFLR model achieved superior performance with 94.8% accuracy, 93.1% precision, and 92.7% recall using Python 3.10. It effectively reduces bias across sensitive groups while preserving overall classification performance, demonstrating balanced trade-off accuracy. It demonstrates that fairness-aware optimization can effectively mitigate bias while maintaining model reliability.

Keywords: Bias Detection, Fairness Optimization, Statistical Parity, Decision-Making.

1. Introduction

Advanced learning mechanisms have enabled effective data processing, precise forecasting, and flexibility in their various implementations. However, current solutions encounter several drawbacks, associated with unstable bias mitigation techniques, and lower generalization capabilities [1]. Data-driven decision-making models increase efficiency and reliability across different fields of use. The existing problems comprise the limitation of the impact of sensitive attributes, and balanced data representation [2]. Bias-informed models in the generation of predictions make it possible for organizations to achieve fairness, consistency, and justice within their populations. However, some difficulties remain, such as an imbalance in the available

data and ensuring fairness while maintaining predictive accuracy [3]. As algorithmic decision-making has become common practice within crucial domains, the reason behind its prevalence is its ability to increase efficiency, and scalability. Nevertheless, current models exhibit biased data, and differential treatment towards certain sensitive groups [4]. The rise in algorithmic systems is mainly driven by their potential for increasing efficiency and consistency, and for analyzing huge amounts of data. Nonetheless, current decision-making algorithms usually have problems with bias, and difficulties in ensuring equity [5]. Automated decision-making systems in various sectors, help increase efficiency and enable processing of massive amounts of information. Current issues include inequity in dealing with different demographic groups, opacity, and fairness versus efficiency trade-offs [6]. Automated decision systems have gained popularity in healthcare, finance, education, and transportation to improve efficiencies and lower manual effort. However, existing fairness criteria are limited and context dependent [7-8]. To address these challenges, this research designs an Adaptive Gradient-Constrained Fair Logistic Regression (AGCFLR) model, which aims at detecting and optimizing bias in algorithmic decision-making systems.

1.2 Research organization: Introduction to detection of bias, and earlier research-driven fair predictive models are detailed in Sections 1 and 2. Section 3 presents the suggested AGCFLR model and its design process. Section 4 and 5 presents the experimental setup, and performance findings and discussion of this research. The final conclusion with future research directions is provided in Section 7.

2. Related works

Existing fairness-aware approaches improve bias mitigation, fairness metrics, and classification accuracy across domains, but face limitations including scalability, computational complexity, domain dependency, and dataset generalization challenges. Table 1 shows the effectiveness of Fairness-Aware Machine Learning (ML) Approaches.

Table 1: Fairness-Aware Machine Learning Approaches for Bias Detection and Mitigation

Ref	Objective	Method	Results	Limitations
[9]	Improvement of Fairness and generalization of drowsy driving detection algorithms	Generative Adversarial Networks (GAN) for augmenting data through data visualization of population bias.	Improved accuracy for the detection of ethnic minorities	Dependence on the generated image quality and insufficient diversity in evaluation scenarios.
[10]	Guaranteed fairness for multi-class classification by reducing group bias	Used ML model Debiaser for Multiple Variables (DEMV) to reduce unbalanced group bias	DEMV performs better in terms of both multi-class and multi-sensitive	The performance is influenced by generation strategies
[11]	Maximized fairness in Reconfigurable Intelligent Surface (RIS) -assisted Visible Light Communication (VLC) systems	RIS is used to control the refractive index of the module	Enhanced user fairness and light distribution efficiency	Depends on assumptions about accurate channel modeling
[12]	Enhanced Fairness and Sum-Rate and fairness performance	Performance in RIS-Assisted VLC Systems Under Non-Convex Conditions	lower outage rate and improved fairness	High computational complexity and Parameter sensitivity.
[13]	Fairness-aware learning for Artificial Intelligence (AI) bias detection	Random Forest (RF) to detect algorithmic bias	Reduced algorithmic bias in educational datasets	Possible information loss affecting accuracy and interpretability
[14]	Mitigate intersectional bias in educational ML systems	Fairness in Education (FAIREDU) is an ML framework used for the reduction of intersectional bias	The result shows Disparate Impact (DI) by 96.7% and Statistical Parity Difference (SPD) by 88.55%,	Requires validation on heterogeneous datasets and models

[15]	Improvement of fairness in Computer Vision (CV) and Natural Language Processing (NLP) models	AI Fairness 360 (AIF360) based Comparative Fairness Mitigation (CFM) approach	Sequential mitigation reduced bias without major performance loss.	Domain dependency, scalability issues, and high computation
[16]	Improved fairness in cardiovascular risk prediction	Susceptible Carrier Infected Recovered (SCIR) based prediction model using Magnetic Resonance Imaging (MRI) and tabular data	Improved fairness metrics include Equal Opportunity Difference (EOD) and DI.	Limited scalability and high computational requirements

2.1 Research gap

Several major breakthroughs were made regarding fairness-aware machine learning and DL technologies; current models continue to struggle with issues of scalability, computational complexity, and generalizability. In addition, ensuring fairness while retaining accurate and robust predictions is an unsolved problem in algorithmic bias detection systems [11-14]. In this research, AGCFLR addresses all these problems by incorporating adaptive fairness optimization, feature-balanced representation, and consistent bias reduction with reliable classification performance.

3. Proposed Methodology

Develop an adaptive fairness-aware optimization model for detecting algorithmic bias and improving equitable decision-making while preserving predictive accuracy and reliability. Figure 1 show the proposed approach developed in this research. The Fairness and Bias Detection Dataset is preprocessed with techniques that include reweighing and the Synthetic Minority Oversampling Technique (SMOTE), while feature extraction can be done by implementing Principal Component Analysis (PCA). The novel aspect of this research is AGCFLR, which is used for bias-sensitive decision outcomes.

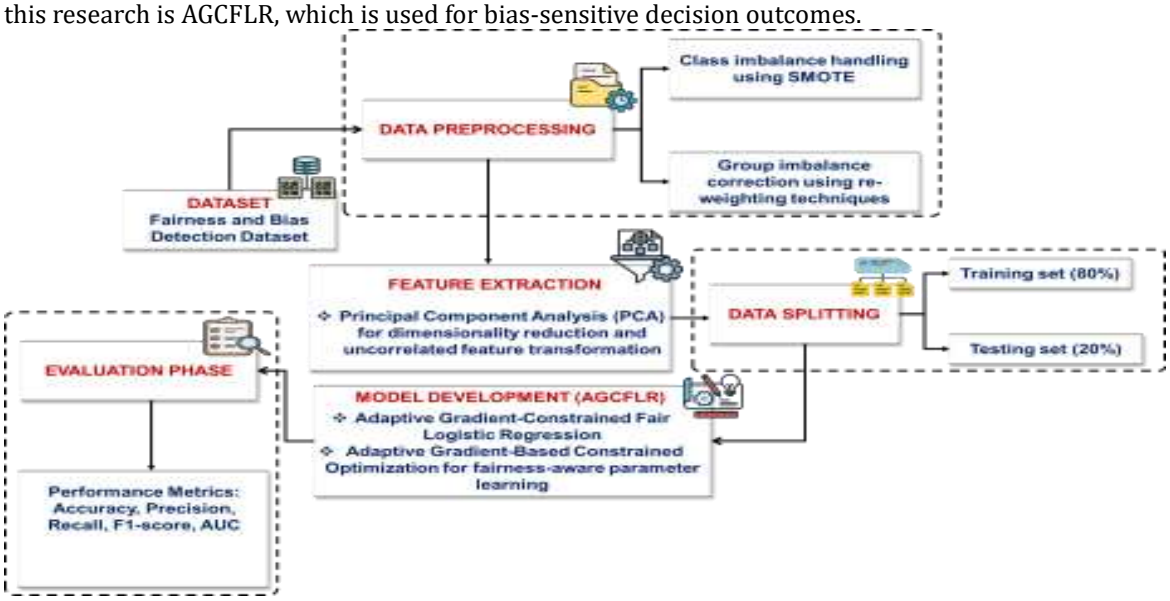

Figure 1: Methodological Workflow for Bias Detection and Fairness Optimization in Algorithmic Decision-Making Systems

3.1 Data collection

The Fairness and Bias Detection Dataset was gathered from the Kaggle platform, Link: https://www.kaggle.com/datasets/programmer3/fairness-and-bias-detection-dataset/data. It contains demographics and decision-making algorithms that help to detect and analyze bias within algorithmic systems. The dataset contains 30000 rows and 27 columns covering structured demographic, employment, and behavioral attributes used to analyze fairness in automated decision processes. The dataset includes sensitive demographic variables such as gender, ethnicity, age, and regional background to evaluate fairness

across multiple protected groups. The dataset was split into 80:20 ratio with the intention of training and testing.

3.2 Data preprocessing

The Fairness and Bias Detection Dataset contain uneven distribution within demographics, unequal representation within groups, and biased distribution of classes. Preprocessing is crucial for achieving fairness through proper data balancing using reweighting method for decreasing demographic bias and SMOTE for synthesizing more minorities.

- **Fairness-Aware Reweighting for Balanced Decision**

Reweighting consists of unequal representation and biased decision-making among the different groups based on demographics prior to the implementation of appropriate fairness-aware weights. In this research, the purpose of the reweighing technique is to reduce demographic bias and discrimination in the form of assigning fairness-aware weights to both over-represented and under-represented groups in order to ensure balanced learning without changing any of the existing class labels.

$$W(Z) = \frac{Q_{exp}(S=Z(S) \wedge Class=Z(Class))}{Q_{obs}(S=Z(S) \wedge Class=Z(Class))} \quad (1)$$

In Equation (1), $W(Z)$represents the reweighing weight assigned to instance Z, Q_{exp}represents the expected probability distribution, Q_{obs}represents the observed probability distribution, S represents the sensitive attribute such as gender, race, or age, $Class$ represents the target class label $Z(S)$represents the sensitive attribute value of instance Z, $\wedge$ represents the logical AND operator, $Z(Class)$represents the class label value of instance Z. The outcome of reweighing achieves balanced and fairness-aware weighted data representation that reduces bias among sensitive groups, improving the efficacy, reliability, and fairness of the proposed AGCFLR model.

- **SMOTE-Based Sensitive Group Balancing for Fairness Enhancement**

An imbalanced structured dataset consisting of samples from the minority and majority classes, including attributes related to demographics, behavior, and decisions for use in fairness-aware classification and detecting biases. The technique of SMOTE is used to handle the issue of class imbalance in the dataset through the creation of synthetic minority class samples.

$$D(a_i, a_j) = \sqrt{\sum_{m=1}^{q}(a_{i,m} - a_{j,m})^2} \quad (2)$$

Equation (2) shows, D shows the Euclidean distance between two minority-class samples, a_i shows the current minority-class sample, a_j shows the neighboring minority-class sample, $a_{i,m}$ shows the value of the m^{th} feature in sample a_i, $a_{j,m}$ shows the value of the m^{th}feature in sample, $a_j q$ shows the total number of features considered in the dataset, m shows the feature index used during distance computation.

$$A_{synthetic} = \delta a_j + (1 - \delta)a_i \quad (3)$$

Where Equation (3), $A_{synthetic}$ shows the newly generated synthetic minority-class sample, a_i shows the selected minority-class data instance from the original dataset, a_j shows the nearest neighboring minority-class sample, δ shows the random interpolation, and $(1-\delta)$ shows the complementary interpolation weight assigned to the original sample. However, the SMOTE optimizes balanced classes with better distribution of minority classes, and better fairness and generalization of downstream models.

3.3 PCA for Redundant Feature Reduction

After the preprocessing phase is completed, the improved dataset is then passed on to the next step, which involves PCA to reduce redundant features. PCA consists of balanced and normalized feature vectors with sensitive demography attributes, behavior features, and decision-related features produced after preprocessing, reweighting, and SMOTE minority class balancing. PCA is used for reducing dimensions and removing correlation, where the features that were correlated initially are transformed into an orthogonal space using PCA, maximizing variance at the same time.

$$W_l = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_m \end{bmatrix} \quad (4)$$

Equation (4) denotes W_l, the input feature matrix used for PCA transformation, W_1, W_2, … , W_m represent individual feature vectors, while m refers to the number of features used in the reduction process.

$$D_{a,a} = \begin{bmatrix} \sigma_{1,1} & \cdots & \sigma_{1,m} \\ \vdots & \ddots & \vdots \\ \sigma_{m,1} & \cdots & \sigma_{m,m} \end{bmatrix} \quad (5)$$

In Equation (5), $D_{a,a}$ Covariance matrix representing relationships among features, $\sigma_{1,1}, \sigma_{m,m}$, variance of individual features indicating feature-wise dispersion contributing to fairness-aware variability analysis.

$$\sigma_{j,i} = \frac{1}{M-1}\sum_{o=1}^{M}(DN_{o,j} - \mu_j)\,(DN_{o,i} - -\mu_i) \quad (6)$$

In Equation (6), $\sigma_{j,i}$ Covariance between the j^{th} and i^{th} features, M Total number of samples in the dataset, o Index representing each sample instance, DN Numerical feature value representing the data intensity, $DN_{(o,j)}$ Value of the j^{th} feature for the o^{th} sample, $DN_{(o,i)}$ Value of the i^{th} feature for the o^{th} sample, μ_j Mean value of the j^{th} feature, μ_i Mean value of the i^{th} feature.

$$det(D - \lambda J) = 0 \quad (7)$$

Equation (7), det Determinant operation, D Covariance matrix, λ Eigenvalue representing variance, captured by each component, J Identity matrix.

$$JZ_l = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix} = \begin{bmatrix} \omega_{1,1} & \cdots & \omega_{1,m} \\ \vdots & \ddots & \vdots \\ \omega_{m,1} & \cdots & \omega_{m,m} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix} \quad (8)$$

In Equation (8), J Index represents a specific principal component, Z_l Principal component output matrix after transformation, $z_1, z_2, \dots, z_m$: Individual principal components, $\omega_{1,1}, \omega_{m,m}$ Eigenvector coefficients forming the transformation matrix, $\omega_{j,i}$: Weight associated with the i^{th} feature in the j^{th}principal component, $w_1, w_2, \dots, w_m$ Original input feature vectors used in PCA transformation. However, the outcome is that of principal components, which capture most discriminatory and structural information, without much noise and multicollinearity.

3.4 Bias Detection and Fairness Analysis

The process of detecting bias examines fairness before and after the classification process. It is used to detect the levels of discrimination such as, SPD, DI, and EOD. Detecting bias serves the purpose of identifying bias, optimizing AGCFLR, and ensuring fairness**.**

3.4.1 Fairness Evaluation using SPD

SPD measures fairness based on the assessment of the positive results from the prediction that are distributed equally among privileged and unprivileged demographic populations within the algorithmic decision-making system.

$$DP = P(\hat{C} = 1 \mid A = 0) - P(\hat{C} = 1 \mid A = 1) \quad (9)$$

Equation (9), shows DP represents the Demographic Parity difference, $\hat{C}$ represents the predicted class label, A represents the sensitive attribute or protected group, $A = 0$ shows the unprivileged group, $A = 1$ shows the privileged group, $P(\hat{C} = 1 \mid A = 0)$represents the probability of predicting a positive outcome for the unprivileged group $P\left(\hat{C} = 1 \mid A = 1\right)$ shows the probability of predicting a positive outcome for the privileged group.

3.4.2 DI-Based Demographic Fairness Assessment

DI measures fairness based on comparing the rate of predictive outcomes for privileged vs unprivileged demographic categories in order to detect possible discriminatory bias in decision-making algorithms.

$$DI = \frac{P(\hat{Z}=1|G=1)}{P(\hat{Z}=1|G=0)} \quad (10)$$

In equation (10), DI represents the Disparate Impact ratio, $\hat{Z}$ denotes the predicted outcome indicating a favorable decision such as approval selection, G represents the sensitive attribute corresponding to group membership, $G = 1$denotes the disadvantaged group, $G = 0$ denotes the privileged group, $P(\hat{Z} = 1 \mid G = 1)$ represents the probability of the protected group, $and\ P(\hat{Z} = 1 \mid G = 0)$represents the probability of the privileged group.

3.4.3 EOD for Bias Evaluation in Predictive Models

The EOD measures discrimination based on the differences in the True Positive Rate for privileged versus unprivileged groups. The model uses EOD to ensure that equal numbers of positives are identified among the various sensitive attributes without underestimating.

$$\text{EOD} = P(X = 1 \mid B = b, X = 1) - P(X = 1 \mid B = b', X = 1) \quad (11)$$

Equation (11), EOD is defined as the difference in true positive rates, P denotes predicted outcome label of the model, where $X = 1$ represents a positive prediction, B denotes the protected attribute, such as gender, caste, or age group, B=b denotes the privileged group within the protected attribute, and $B = b'$denotes the unprivileged group within the protected attribute.

3.5 AGCFLR: Bias Reduction and Fairness Optimization for Reliable Decision-Making

The proposed AGCFLR model minimizes algorithmic bias through adaptive fairness constraints while maintaining accurate, interpretable, and reliable decision-making performance. Specifically, it integrates Adaptive Gradient-Based Constrained Optimization (AGCO) dynamically updates model parameters under fairness constraints to minimize bias while preserving classification performance, and Fair Logistic Regression (FLR) performs interpretable binary classification by estimating decision probabilities and supporting efficient fairness-aware predictive decision-making processes.

3.5.1 FLR: Interpretable Classification for Fair and Accurate Decision Prediction

The FLR is used to perform efficient and interpretable binary classification by estimating decision probabilities from extracted features. In this research, it supports fairness-aware prediction, enables transparent analysis of decision outcomes, and maintains stable classification performance while integrating adaptive optimization constraints to reduce algorithmic bias.

$$P(D = 1 \mid F_1, F_2, \dots, F_n) = \frac{exp(\alpha_0+\alpha_1F_1+\alpha_2F_2+\cdots+\alpha_nF_n)}{1+exp(\alpha_0+\alpha_1F_1+\alpha_2F_2+\cdots+\alpha_nF_n)} \quad (12)$$

In Equation (12), P denotes the binary target variable, where $D = 1$ indicates the positive class, $F_1, F_2, \dots, F_n$ represent the predictor features. α_0 denotes the intercept term of the FLR model. $\alpha_1, \alpha_2, \dots, \alpha_n$ denote the regression coefficients associated with each feature. exprepresents the exponential function. $(\alpha_0 + \alpha_1F_1 + \alpha_2F_2 + \cdots + \alpha_nF_n)$ represents the instance to positive class. Although this model provided probabilistic outputs and class labels, where the cut-off was 0.5, and fairness-aware optimization to minimize any bias between sensitive groups, it also provided classification accuracy.

3.5.2 AGCO: Dynamic Fairness Constraint Optimization for Bias Mitigation and Performance Preservation

The gradient-based constrained optimization (GCO) approach is used for minimizing the classification loss subject to fairness constraints. Hence, there are many limitations associated with the use of standard gradient-based optimization approaches, dependence on the choice of learning rate, and poor performance in handling high-dimensional data. AGCO was introduced to overcome these limitations and to optimize parameter updates during training, converge faster, decrease oscillation issues, and perform fair learning optimization while maintaining accurate classification among different demographics.

$$minimize\ loss\ (\theta\ \epsilon\ \mathbb{R}^c)\ =\ loss\ (Q_c\ , W_c\ , H_c,\ A_c\ , Q_{dc}\ ,\ W_{dc}) \quad (13)$$

Equation (13), shows θ Parameter vector optimized during model training, ϵ symbol representing membership, $\mathbb{R}^c$ dimensional real-valued parameter space, c Total number of trainable model parameters, Q_c Classification-related model parameter component, W_c Weight parameter associated with fairness-aware learning, H_c Hidden optimization or feature interaction parameter, A_c Adaptive fairness constraint parameter, Q_{dc} Decision correction parameter used for bias reduction, W_{dc} Weighted fairness adjustment parameter during constrained optimization.

$$n_s\ =\ \emptyset(h_1, h_2, \dots, h_s) \quad (14)$$

In Equation (14), n_s First-order moment estimates at iteration s, $\emptyset$ Function used to compute the first-order gradient moment, $h_1, h_2, \dots, h_s$ Gradient values obtained from previous optimization iterations.

$$u_s\ =\ \psi(h_1, h_2, \dots, h_s) \quad (15)$$

In Equation (15), u_s is the second-order moment estimate representing adaptive variance information at iteration s, ψ is the function used to compute the second-order gradient moment, $h_1, h_2, \dots, h_s$ hs are the gradient values obtained from previous optimization iterations.

$$\theta_{s+1} = \theta_s - \frac{1}{\sqrt{u_s+\varepsilon}}\ n_s \quad (16)$$

In Equation (16), θ_s Parameter vector at the current iteration s, $\theta_{(s+1)}$ updated parameter vector after the current optimization step, ϵ Small smoothing constant preventing division instability, $\sqrt{u_s + \epsilon}$ Adaptive normalization factor controlling learning stability, u_s Second-order moment estimate representing adaptive variance information at iteration s, n_s Gradient-based momentum term used for parameter updating. Therefore, the proposed AGCFLR model achieves reliable and unbiased decision classification as well as minimizes statistical bias, enhances fairness and consistency, and preserves classification accuracy across sensitive demographic groups.

4. Experimental Setup and Evaluation Metrics

The development of the AGCFLR model is conducted in the Python 3.10 with NumPy, Scikit-learn, TensorFlow libraries. The following devices are used to conduct the experiments: 16 GB RAM, 512 GB SSD hard drive space, Intel Core i7 processor and Windows 11 operating system. Table 2 shows the evaluation metrics.

Table 2: Performance Evaluation Metrics for Bias Detection and Fairness

Metrics	Explanation
Accuracy (%)	Measures the total percentage of accurate classifications.
Precision (%)	Measures the percentage of true positive values.
Recall (%)	Measures the accuracy of true positive classification within all the positive instances.
F1-Score (%)	Measures the ability of the model.
Receiver Operating Characteristic – Area Under the Curve. (ROC-AUC) (%)	Measures the positive and negative classes.

SPD ↓	Measures disparity in terms of positivity for privileged versus underprivileged groups.
DI ↑	Measures bias of comparative evaluation.
EOD ↓	Measures the difference in target population groups.
Bias Reduction (%)	Represents the percentage of bias among sensitive groups.
Fairness Consistency (%)	Measures fairness prediction in demographic groups.
Error Rate (%)	It represents the proportion of misclassified cases.
Fairness Score (%)	Measures the overall fairness performance.
Computational Complexity	Represents the processing time for model training.

5. Results

The integrated model for bias detection and fairness optimization in algorithmic systems was effectively offered to ensure equitable and robust decision-making. This section offers the findings of AGCFLR performance, fairness improvement, bias reduction, and comparative analysis against existing models.

Analysis of demographic and transactional data distributions, as shown in Figure 2, is essential for bias-aware decision modeling. It analyses age distribution and transactional distribution to detect balanced population distribution and behavioral diversity for bias-aware classification.

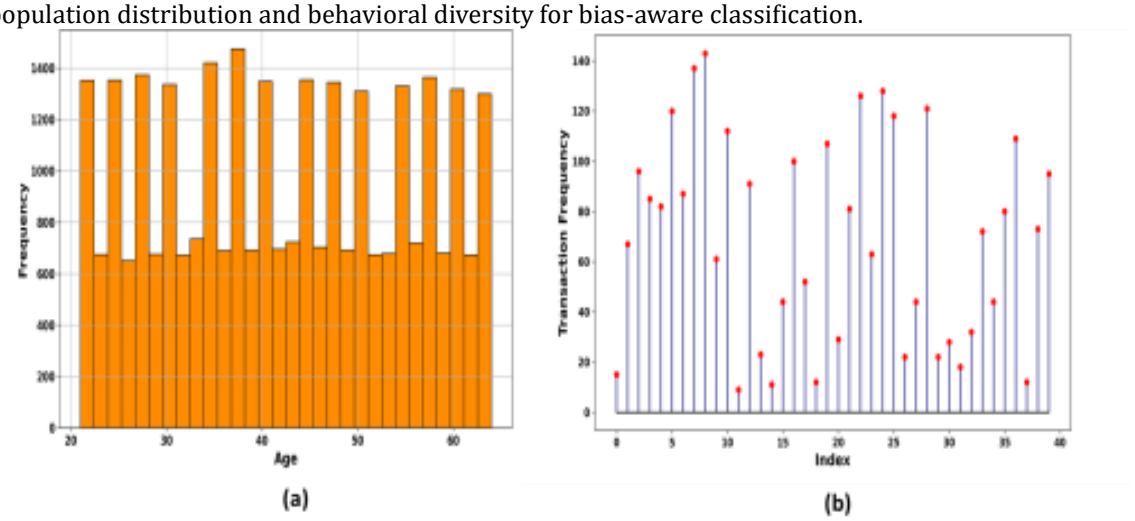

Figure 2: Graphical representation of (a) Age Distribution Analysis and (b) Transaction Frequency Distribution

Figure 2(a) shows the age distribution between 21 and 65 years, with frequency counts ranging between 1300 and 1480 per group, and Figure 2(b) shows the frequency distribution of transactions varies between 10 and 145 in sample indices.

5.1 Comparative Analysis

The comparative analysis of the AGCFLR model is conducted by comparing its performance with the existing Fairness-Constrained Neural Network (FCNN) model [16] in fairness-aware classification techniques. For this comparison, both models were trained and evaluated using proposed Fairness and Bias Detection Dataset with same splits. Table 3 shows the fairness and the bias comparison, and the performance. Figure 3 shows the comparison of classification performances between the FCNN model and the proposed AGCFLR model. Better performance was obtained by the proposed model were 94.8% accuracy, 93.1% precision, 92.7% recall, 92.9% F1-score, and 94.1% ROC-AUC.

Table 3: Comparative Analysis of Performance Metrics and Fairness Metric Comparison and Bias Reduction Performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)	SPD↓	DI ↑	EOD ↓
FCNN	89.4	88.8	88.3	88.5	90.2	0.084	0.084	0.081
AGCFLR [Proposed]	94.8	93.1	92.7	92.9	94.1	0.051	0.094	0.046

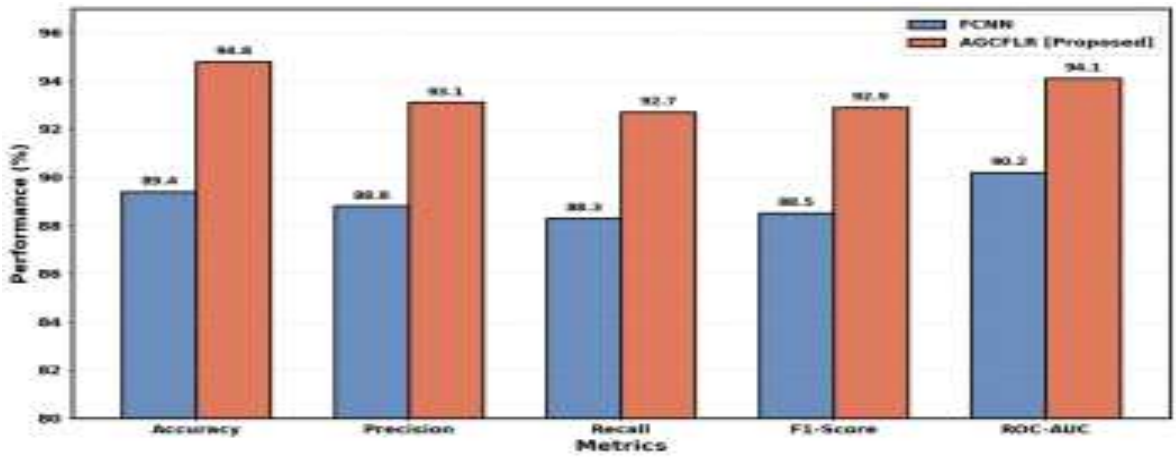

Figure 3: Graphical Comparison of existing and proposed model using performance metrics

Table 4 shows Fairness–Accuracy Trade-Off Comparison. The comparative trade-off analysis shows that the proposed AGCFLR model achieves higher performance than FCNN, with accuracy improved to 94.8% and fairness score increased to 95.1%. For bias reduction, there was an increased to 91.7%. Similarly, for fairness consistency, it increased to 94.2%. The error rate decreased to 6.2.

Table 4: Fairness–Accuracy Trade-Off Comparison

Model	Bias Reduction (%)	Fairness Consistency (%)	Error Rate (%)	Accuracy (%)	Fairness Score (%)	Computational Complexity
FCNN	72.4	86.5	10.6	89.4	88.3	High
AGCFLR [Proposed]	91.7	94.2	6.2	94.8	95.1	Medium

5.2 Discussion

The integrated model for bias detection and fairness optimization in algorithmic systems was effectively offered to ensure equitable and robust decision-making. Existing approaches that include the use of FCNN [16] model were generally adopted for the detection of biases; nevertheless, their application was hindered by several shortcomings. Moreover, previous models had poor fairness results, biases, and a lack of reliability in decision-making processes. Conversely, this proposed model was shown to effectively enhance fairness scores while ensuring the accuracy of predictions. Therefore, it concluded that the proposed AGCFLR model is a reliable means for resolving biases.

6. Conclusion

The bias detection and fairness optimization in algorithmic systems was effectively developed to ensure equitable and robust decision-making using AGCFLR model. It utilizes Fairness and Bias Detection dataset and pre-processing done by reweighting and SMOTE with PCA-based feature extraction. The proposed model AGCFLR shows enhanced results in accuracy (94.8%), F1-score (92.9%), and ROC-AUC score (94.1%), while showing minimal bias metrics (SPD: 0.051, DI: 0.94, EOD: 0.046). The model maybe sensitive to parameter tuning and dataset variability, while future work will concentrate on scalable and real-time fairness optimization.

References

[1] Patil, P. and Purcell, K., 2022. Decorrelation-based deep learning for bias mitigation. Future Internet, 14(4), p.110. https://doi.org/10.3390/fi14040110

[2] Li, S., Yu, J., Du, X., Lu, Y., and Qiu, R., 2022. Fair outlier detection based on adversarial representation learning. Symmetry, 14(2), p.347. https://doi.org/10.3390/sym14020347

[3] Yang, J., Soltan, A.A., Eyre, D.W., Yang, Y. and Clifton, D.A., 2023. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. NPJ digital medicine, 6(1), p.55. https://doi.org/10.1038/s41746-023-00805-y

[4] Popoola, G. and Sheppard, J., 2024. Investigating and Mitigating the Performance–Fairness Tradeoff via Protected-Category Sampling. Electronics, 13(15), p.3024. https://doi.org/10.3390/electronics13153024

[5] Nagpal, R., Khan, A., Borkar, M., and Gupta, A., 2024. A multi-objective framework for balancing fairness and accuracy in debiasing machine learning models. Machine Learning and Knowledge Extraction, 6(3), pp.2130-2148. https://doi.org/10.3390/make6030105
[6] Wang, X., Chang, C.H. and Yang, C.C., 2024. Achieving equity via transfer learning with fairness optimization. IEEE Access, 12, pp.195229-195241. https://doi.org/10.1109/ACCESS.2024.3519465
[7] Perera, A., Aleti, A., Tantithamthavorn, C., Jiarpakdee, J., Turhan, B., Kuhn, L., and Walker, K., 2022. Search-based fairness testing for regression-based machine learning systems. Empirical Software Engineering, 27(3), p.79.https://doi.org/10.1007/s10664-022-10116-7
[8] Ngxande, M., Tapamo, J.R., and Burke, M., 2020. Bias remediation in driver drowsiness detection systems using generative adversarial networks. IEEE Access, 8, pp.55592-55601. https://doi.org/10.1109/ACCESS.2020.2981912
[9] d'Aloisio, G., D'Angelo, A., Di Marco, A. and Stilo, G., 2023. Debiaser for Multiple Variables to enhance fairness in classification tasks. Information Processing & Management, 60(2), p.103226. https://doi.org/10.1016/j.ipm.2022.103226
[10] Zhang, X.Y., Zhang, J., and Wu, Q., 2024. Fairness optimization for VLC systems with liquid crystal-based RIS-enabled transmitters. IEEE Photonics Journal, 16(4), pp.1-8. https://doi.org/10.1109/JPHOT.2024.3429187
[11] Abdeljabar, S., Eltokhey, M.W., and Alouini, M.S., 2024. Sum rate and fairness optimization in RIS-assisted VLC systems. IEEE Open Journal of the Communications Society, 5, pp.2555-2566. https://doi.org/10.1109/OJCOMS.2024.3389093
[12] Moon, D. and Ahn, S., 2025. Metrics and algorithms for identifying and mitigating bias in AI design: a counterfactual fairness approach. IEEE Access, vol. 13, pp. 59118-59129, 2025 https://doi.org/10.1109/ACCESS.2025.3556082
[13] Pham, N., Do, M.K., Dai, T.V., Hung, P.N. and Nguyen-Duc, A., 2025. FAIREDU: A multiple regression-based method for enhancing fairness in machine learning models for educational applications. Expert Systems with Applications, 269, p.126219. https://doi.org/10.1016/j.eswa.2024.126219
[14] Rashed, A., Kallich, A., and Eltayeb, M., 2025. Analyzing Fairness of Computer Vision and Natural Language Processing Models. Information, 16(3), p.182. https://doi.org/10.3390/info16030182
[15] Sufian, M.A., Alsadder, L., Hamzi, Wss., Zaman, S., Sagar, A.S., and Hamzi, B., 2024. Mitigating algorithmic bias in AI-driven cardiovascular imaging for fairer diagnostics. Diagnostics, 14(23), p.2675. https://doi.org/10.3390/diagnostics14232675
[16] Nagaraj, S.K.S., 2025. An Analytical Framework for Bias Mitigation in Credit Scoring Systems through Fairness-Constrained Neural Optimization. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 6(1), pp.186-195. https://doi.org/10.63282/3050-9262.IJAIDSML-V6I1P120