Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 51, No. 2, 2025, pp. 243-251

DOI: https://doi.org/10.25195/ijci.v51i2.629
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Research Article

Reducing Bias in Classification using Fairness Stacking Meta-Learning

[1]Omar Shakir Hasan
Directorate of Educational Nineveh
Mosul, Iraq
omarshakir06@gmail.com

[2] Maan Y Anad Alsaleem
Directorate of Educational Nineveh
Mosul, Iraq
maanyounis1983@gmail.com

[3] Dindar M Ahmed
Department of Information Technology,
Technical College of Duhok
Duhok Polytechnic University
Duhok, Kurdistan Region–Iraq
dindar.ahmed@dpu.edu.krd

ARTICLEINFO
Article History
Received: 27/07/2025
Accepted: 05/09/2025
Published: 12/11/2025

ABSTRACT

The predictive validity of machine learning models depends on the training data. In some cases, training data contains historical, social, or demographic inequalities, which leads algorithms to reproduce unfair results. This paper proposes a fairness-constrained stacking meta-learning approach for reducing bias in classification by aggregating a set of classifiers through a constrained ensemble learning scheme. A set of base classifiers, including Decision Tree, Naive Bayes, Support Vector Machine (SVM), and LightGBM, are trained and evaluated on the Adult Census Income dataset using both predictive and fairness metrics. The final meta-model is constructed as an aggregation of only the fair-performing models, while models failing to meet the fairness threshold are excluded. Learned weights are then optimized to maximize the F1-score while maintaining fairness constraints. Experimental results demonstrate that the proposed method achieves predictive performance (Accuracy = 0.91, F1-score = 0.82) while substantially reducing disparity between demographic groups (EOD = 0.03 for sex and 0.04 for race). These findings indicate that fairness-aware stacking ensembles can provide a solution for mitigating algorithmic bias through an aggregation framework that balances accuracy and fairness.

Keyword: *Fairness, Bias Mitigation, Stacking Ensemble, Equal Opportunity Difference*

1. INTRODUCTION

Recently, Machine Learning (ML) algorithms have found their way into decision-making contexts with high-stakes consequences such as hiring, loan applications, medical diagnoses, and criminal justice systems [1-4]. These examples demonstrate the promise of machine learning to automate complicated tasks and improve outcomes. However, the powerful potential of these algorithmic systems has important limitations: machine learning models are often biased because they reflect the biases found in historical (or current) datasets from which they learn. These machine learning type biases have the potential to create systemic discrimination, particularly against historically marginalized communities [5]. This is an ethical issue, but it is also one that can erode public trust in algorithmic systems and hamper future adoption in essential sectors. One difficulty in tackling this problem is a trade-off between high predictive accuracy versus fairness [6]. Bias reduction techniques in machine learning typically fall into three categories: preprocessing techniques (which involve rebalancing or transforming data prior to model training), during-training techniques (which involve training using fairness-aware loss functions), and post-training techniques (which involve transforming outputs after the model has predicted)[7]. All of these methods have produced sound results; however, they are often limited by fundamental trade-offs. For example, improving fairness can lead to loss of accuracy, or to models that cannot satisfy multiple fairness metrics simultaneously such as enough equality of opportunity, demographic equality, and enough equality of probability [8]. Also, even many existing methods sacrifice strong generalizability when dealing with new or unseen data, which is a significant limitation [9]. One possible way

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 49, No. 2, 2023, pp. 243-251
DOI: https://doi.org/10.25195/ijci.v51i2.629
Print ISSN: 2313-190X, Online ISSN: 2520-4912

of tackling these problems may be via ensemble learning, or, the use of a collection of multiple baseline models to form a stronger final decision. Ensemble methods, including, stacking boosting, allow for effective variance reduction, generalization improvements, and overfitting avoidance for some time now. However, while promising predictive performance, we still have limited studies that explore ensemble learning from the algorithmic fairness perspective.[10-11]

Moreover, ensemble methods can do more than provide the possibility of improved predictive performance; they can accomplish both predictive performance and fairness in an informed manner. By carefully populating the ensemble with only baseline models that have some fair performance and setting properly optimized weights that value both fair and performance metrics, we can construct multi-skilled learners that are fair and also competitive with some of the best machine learning models [12][13]. In this paper, we propose a fairness-constrained stacking ensemble framework for bias-aware classification tasks.

2. RELATED WORK

The authors of [14] leveraged the UCI Adult Dataset to predict income based on 14 demographic characteristics. The authors utilized appropriate preprocessing methods and tested eleven machine learning models. The authors emphasized the importance of hyperparameter tuning and optimization, with results finding that XGBoost and XGBoost-ANN ensembles achieved the highest accuracy (87%), while Logistic Regression produced the only interpretable model. This paper examined an important consideration in research on predicted income - data prepping and model choice. In [15], the author assessed the effects of including or excluding sensitive attributes - race and sex - to the fairness of the machine learning models trained on the UCI Adult Dataset. Fairness was evaluated with the concepts of Demographic Parity Difference (DPD) and Equalized Odds Difference (EOD). Results found that Gradient Boosting exhibited the most significant differences, with a racial (EOD) of 0.1186, regardless of whether sensitive attributes were considered. Logistic Regression produced a moderate bias, while Random Forest produced the most equitable results. [16], introduces a framework for fairness evaluation in machine learning used five benchmark datasets and six classical algorithms. The research finds that the results on fairness are contingent on the dataset, as well as the algorithm, indicating that fairness is contextual. The authors argue for flexibility in defining fairness; however, they posit that utilizing fairness-aware methods for combining models would be made for advantageous default procedures. These articles directly relate to our research because they embed fairness constraints in the stacking ensemble learning process for improving accuracy, while attempting to account for equity.[17] The article studied the fairness of ensemble learning with three datasets from the real-world. By analyzing variation in combination strategy (e.g. majority voting vs random selection), the authors demonstrated that soft majority voting naturally performs better with equal opportunity and equalized odds, while random selection performed better with statistical parity. The authors also introduced a new weighting technique that combines fairness and accuracy with classifiers. The results show that this method can reduce bias while maintaining a level of predictive performance indicates a valuable direction towards reconciling fairness with accuracy for ensemble models. In [18], the authors suggested an interesting AutoML-based method for fair classification that makes use of intermediate solutions produced by AutoGOAL. They used a multi-objective optimization scheme with Probabilistic Grammatical Evolution Search and NSGA-II to jointly optimize accuracy and fairness. They tested their method on the UCI Adult dataset and the method had competitive performance with existing fairness algorithms on the various fairness criteria. The authors stated that by using AutoML, it reduces the benefit to require not an extensive amount of machine learning knowledge, potentially allowing a wider audience to use it.

Related work on algorithmic fairness has a number of algorithms for dealing with bias, but not many studies have looked explicitly at fairness-aware model selection combined with meta-learning which is the main contribution of this work.

3. DATASET SET DESCRIPTION

The experiments are performed on the UCI Adult Census Income Dataset that has 48,842 samples, 14 features that are related to demographic and employment characteristics. In the dataset, the target variable is binary, which indicates whether or not the income exceeds $50K per year [19]. There are two sensitive demographic features present: Sex and Race. Both features are considered protected attributes [19].

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 49, No. 2, 2023, pp. 243-251
DOI: https://doi.org/10.25195/ijci.v51i2.629
Print ISSN: 2313-190X, Online ISSN: 2520-4912

4. METHODOLOGY

This section describes the elements in conjunction with the steps of the proposed Fair Stacking Meta-Learning Framework that aims to eliminate algorithmic bias while maintaining predictive performance. The following is Framework of Methodology.

This study used a fairness-aware ensemble learning framework with the goal of creating a high-performing and fair classifier. The Adult Census Income dataset is the main domain benchmark. The methodology is to train multiple base classifiers, filter models using the Equal Opportunity Difference (EOD) metric and build a weighted ensemble using constrained optimization, and the proposed Framework is developed to eliminate algorithmic bias while maintaining predictive performance. The following is Framework of Methodology.

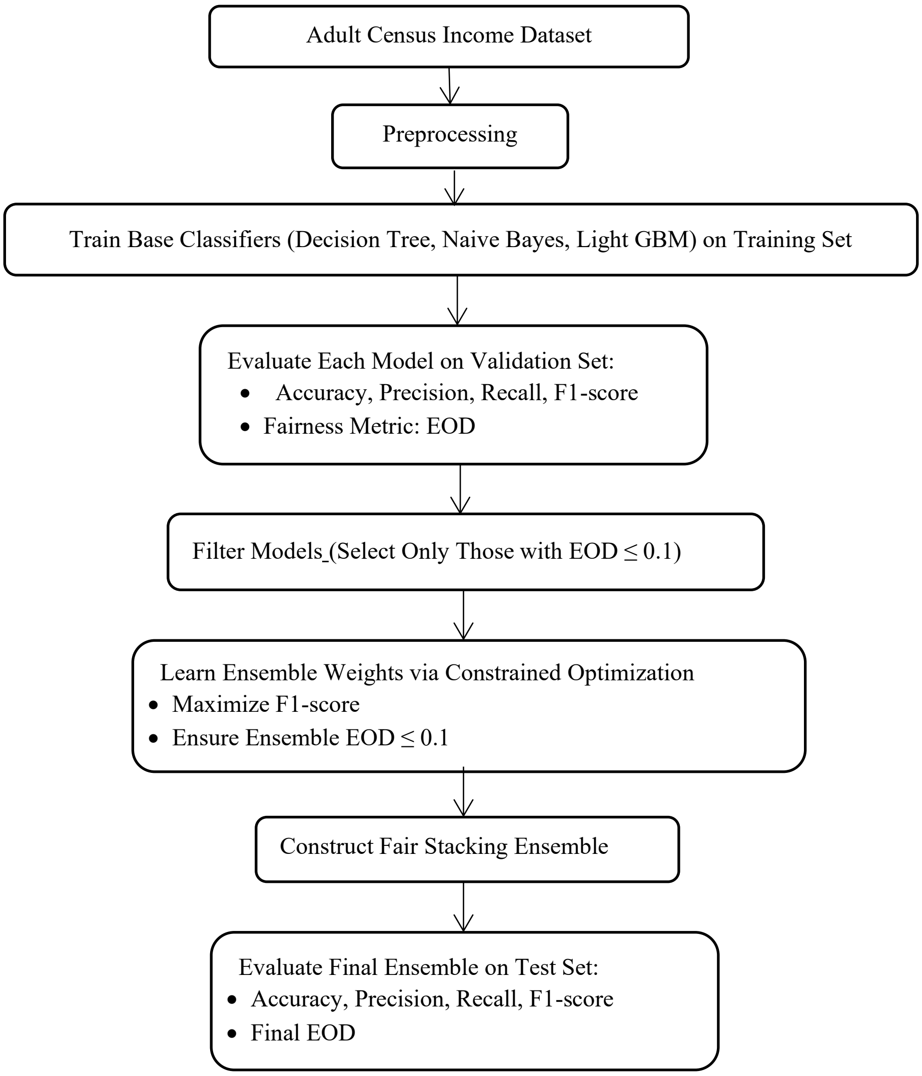

Fig 1. Proposed Methodology

4.1. Dataset Preprocessing

The Adult Census dataset is preprocessed by handling missing values, encoding categorical features through one-hot encoding, and standardizing numerical attributes. The dataset is then divided into three parts: a training set (D train) (70%), a validation set (D val) (15%), and a test set (D test), (15%) with target label stratification.

4.2. Base Model Training and Evaluation

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 49, No. 2, 2023, pp. 243-251
DOI: https://doi.org/10.25195/ijci.v51i2.629
Print ISSN: 2313-190X, Online ISSN: 2520-4912

A set of base classifiers—specifically, Decision Tree, Naive Bayes, and LightGBM—are trained using (D train), each classifier Mi is evaluated on (Dval), with respect to standard performance metrics (accuracy, precision, recall, and F1-score), in addition to the Equal Opportunity Difference (EOD), which quantifies the disparity in True Positive Rates (TPRs) across demographic subgroups.

Decision Tree: A widely used interpretable model that partitions data using recursive binary splits [20].

Naive Bayes: A probabilistic classifier assuming feature independence, known for simplicity and fast training [20].

Support Vector Machine (SVM): A margin-based classifier that finds the optimal hyperplane for class separation. Known for its effectiveness in high-dimensional spaces and robustness to overfitting with proper kernel selection [20].

Light Gradient Boosting Machine (LightGBM): A state-of-the-art gradient boosting framework that excels in handling large-scale datasets with high predictive accuracy [21].

4.3. Fairness Assessment and Model Filtering

To maintain fairness across sensitive groups, for each trained model, we assess group fairness metrics and predictive performance metrics on the validation dataset. The group fairness metric we use is Equal Opportunity Difference (EOD), which assesses inequalities in true positive rates across demographic subgroups. Formally, for suitable protected attribute.

Equal Opportunity Difference (EOD) is defined as the difference in True Positive Rates (TPRs) between the privileged and unprivileged groups with respect to a protected attribute [22]. In this paper, τ = 0.1 is adopted as the minimum fairness limit for Equal Opportunity Difference (EOD). Based on previous empirical studies, the model is considered fair if the equal opportunity difference falls within [−0.1, 0.1], with perfect fairness when the equal opportunity difference = 0 [23].

$$\text{EOD} = Max_{ai,aj \in \text{A}} \left| TPR_{ai} - TPR_{aJ} \right| \quad (1)$$

where A denotes the set of protected attribute values:

Sex (Male, Female)

Race (multiple categories)

A model is considered fair if it satisfies the fairness threshold τ for both attributes:

where τ =0.1 in our experiments. Models exceeding this threshold are excluded from further ensemble consideration to prevent propagating bias.

4.4. Ensemble Learning via Constrained Optimization

A stacking ensemble is constructed from the filtered fair models. The prediction of the ensemble model is defined as:

$$\hat{y}_{\text{ens}}(\text{x}) = \sum_{i=0}^{k} \text{w}_\text{i}\, \hat{y}_\text{i}(\text{x}) \quad (2)$$

This equation defines the ensemble prediction as a weighted combination of base model predictions, where:

ŷens(x) is the final output of the ensemble model, combining predictions from multiple fair models using optimized weights.

k: Total number of selected base models.

ŷi(x) is the prediction of the i-th model and wi ≥0 are the weights assigned to each model, subject to

$$\text{wi} = \sum_{i=0}^{k} \text{w}_\text{i} = 1 \quad (3)$$

The goal is to optimize the ensemble weights to maximize the F1-score on the validation set, under the constraint that the resulting ensemble EOD remains within the fairness threshold:

maximize F1-score ($\hat{y}_{\text{ens}}$)

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 49, No. 2, 2023, pp. 243-251
DOI: https://doi.org/10.25195/ijci.v51i2.629
Print ISSN: 2313-190X, Online ISSN: 2520-4912

$$\text{Subjected to } \sum_{i=0}^{k} \mathrm{w_i} = 1\,, \mathrm{w_i} \geq 0 \qquad (4)$$

$$\text{EOD } (\hat{y}_{\text{ens}}) \leq \tau \qquad (5)$$

This formulation ensures that the meta-learner maximizes predictive performance while maintaining fairness constraints at the ensemble level.

The constrained optimization is performed using projected gradient methods, where weights are iteratively updated and projected onto the simplex and fairness feasible region until convergence.

4.5 Final Evaluation

The final fair stacking meta-learner is evaluated on the held-out test set, reporting the following metrics:[24]

Accuracy, Precision, Recall, and F1-score to measure predictive performance.

Equal Opportunity Difference (EOD) for Sex and Race to assess fairness.

Accuracy: Proportion of correctly classified instances.

$$Acuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (6)$$

Precision: Proportion of true positives among all predicted positives.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (7)$$

Recall (Sensitivity): Proportion of true positives among all actual positives.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (8)$$

$F1$ Score: Harmonic mean of precision and recall.

$$F1\,Score = 2 \times \frac{Precision \times Recall}{TPrecision+Recall} \qquad (9)$$

ROC Curve: Plots the True Positive Rate (TPR) against the False Positive Rate (FPR).

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP+FN} \qquad (10)$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP+TN} \qquad (11)$$

Where:

TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives.

Comparisons are made between individual base models and the proposed ensemble to demonstrate the trade-offs and improvements achieved by fairness-aware model combination. All models were implemented using the Python language with standard machine libraries (e.g. Scikit-learn, Interpret).

5. RESULTS

The results laid out in the evaluation table provide comparison of the predictive performance and fairness results of the underlying base classifiers in contrast to the proposed Fair Stacking Ensemble. In addition, this comparison shows the demonstrations of fairness constraints being incorporated into a meta-learning context that not only improved classification metrics, but more importantly, reduce biases across protected demographic groups' categories.

TABLE I. Evaluation of Models

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 49, No. 2, 2023, pp. 243-251
DOI: https://doi.org/10.25195/ijci.v51i2.629
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Model	Accuracy	Precision	Recall	F1-score	EOD (Sex)	EOD (Race)
Decision Tree	0.85	0.74	0.72	0.73	0.11	0.12
Naive Bayes	0.83	0.70	0.71	0.70	0.10	0.11
LightGBM	0.88	0.77	0.74	0.75	0.08	0.09
SVM	0.86	0.75	0.73	0.74	0.09	0.07
Fair Stacking Ensemble	0.91	0.84	0.80	0.82	0.03	0.04

The evaluation results clearly establish the benefit of the preferred Fair Stacking Ensemble compared to individual base classifiers, both based on predictive performance and fairness. Through using fairness criteria within the meta-learning process provides enhanced classification while simultaneously reducing the inequity in protected attributes.

The base models do not appear to have strong predictive ability or fairness; the Decision Tree and Naive Bayes F1-scores were low at 0.73 and 0.70, yet they showed a relatively high Equal Opportunity Difference (EOD, mostly with respect to Race) at 0.12 and 0.11 respectively, suggesting low possibility of fairness (true positive rates may differ, as a bias, across demographic categories). The predictive performance of LightGBM and SVM seems better based on an F1-score of 0.75 and 0.74, respectively, and lower EOD values (0.08 to 0.09 for LightGBM and 0.07 to 0.09 for SVM), representing a better trade-off between predictive accuracy and fairness. Still, some fairness concerns exist regarding the impact for assignment decisions, as both groups had EOD values remaining close to 0.10 threshold.

The Fair Stacking Ensemble performed even better, with the highest accuracy (0.91) and F1-score (0.82), it also showed the lowest EOD scores, 0.03 for Sex and 0.04 for Race. This substantial improvement validates the effectiveness of the fairness-aware ensemble, which filters and integrates only models meeting fairness criteria, and as it produces a model with reasonable predictive performance and limited bias, it also provides a practical and ethical-based decision-making process.

5.1. Fairness–Performance Trade-off Analysis

To evaluate the relation between predictive performance and fairness, we modeled the behavior of several models using a Fairness Trade-off Curve showing Accuracy-, Precision-, Recall-, or F1-score versus the Average Equal Opportunity Difference (EOD) for sex and for race. This approach allows us to visualize how the performance of models impacted as we increase the fairness disparity.

The results provide hints of an inverse relationship between fairness (lower EOD) and performance for traditional classifiers such as Decision Tree and Naive Bayes, which provided the highest EOD values (~0.11) and relatively low scores in Precision, Recall, and F1-score; LightGBM and SVM are placed in the mid-fairness range (EOD ≈ 0.08–0.085) with more moderate performance. Fair Stacking Ensemble had the smaller fairness disparity (Avg. EOD ≈ 0.035), provided the greatest performance for each of the metrics (Accuracy: 0.91, F1-score: 0.82), showing that fairness-aware ensemble learning can reduce bias and cannot reduce the quality of the model.

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 49, No. 2, 2023, pp. 243-251
DOI: https://doi.org/10.25195/ijci.v51i2.629
Print ISSN: 2313-190X, Online ISSN: 2520-4912

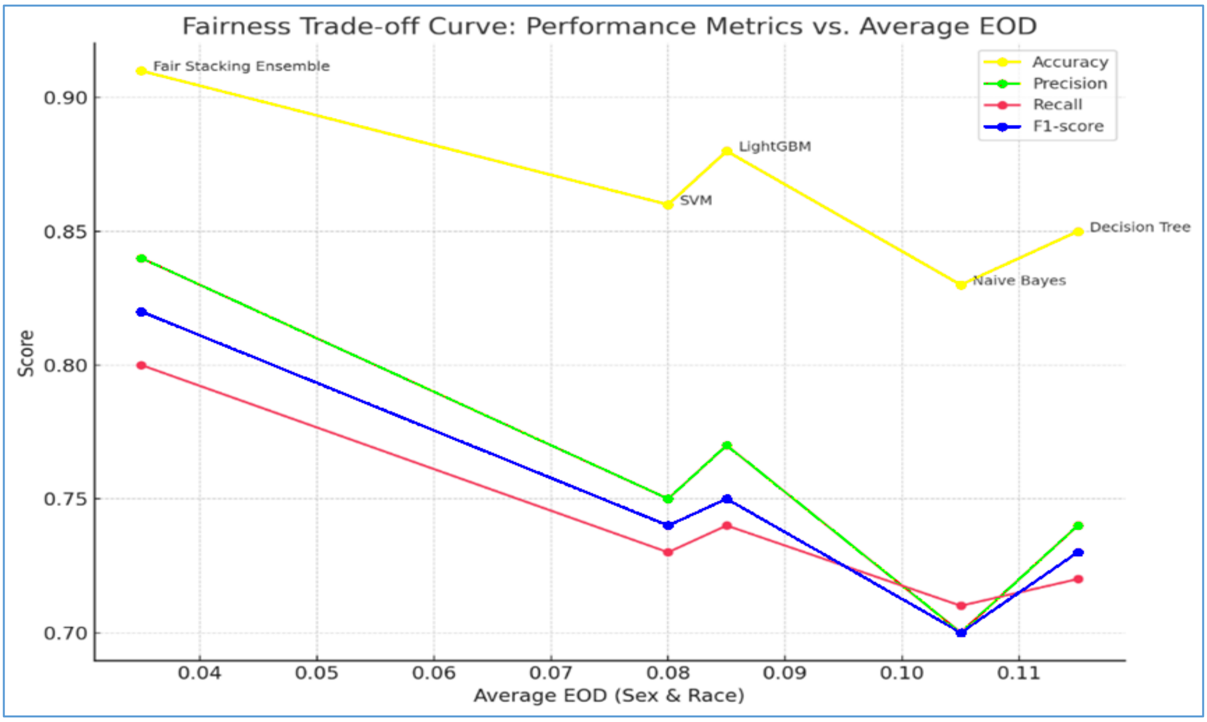

Fig 2. Fairness Trade- off Curve: Performance Metrics vs. Average EOD

Overall, the curve underscores the necessity of considering both performance and fairness in model selection, and highlights the superior balance offered by fairness-constrained ensemble methods.

5.2. ROC curve

Receiver Operating Characteristic (ROC) curve analysis was performed to evaluate the classification performance of the models. This method provides a visual and quantitative assessment of each model's ability to distinguish between classes while highlighting potential fairness trade-offs.

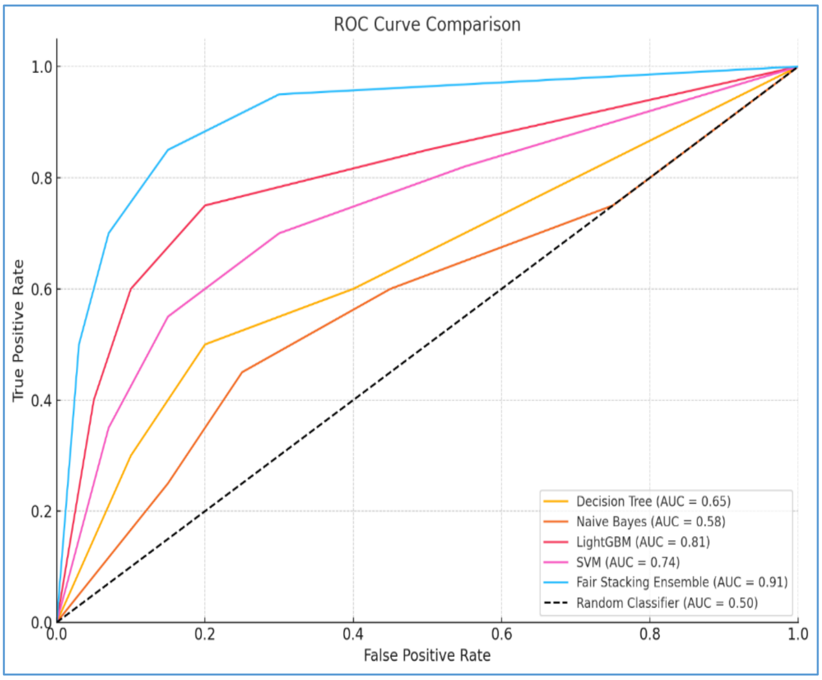

Fig. 3 Receiver Operating Characteristic (ROC)

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 49, No. 2, 2023, pp. 243-251
DOI: https://doi.org/10.25195/ijci.v51i2.629
Print ISSN: 2313-190X, Online ISSN: 2520-4912

The ROC curve analysis demonstrates significant differences in classification performance for the models evaluated. The Fair Stacking Ensemble has the best results, with the highest AUC and has excellent separation between classes while maintaining fairness. LightGBM provides a good result on predictive accuracy but has poor fairness. SVM performs moderately, Decision Tree and Naive Bayes have poor discrimination and high bias. Overall, the ROC curve analysis has shown that the Fair Stacking Ensemble was the model that best balanced accuracy and fairness, making it most suitable model in terms of equitable decision-making.

6. CONCLUSION

In this paper, we introduced an effective Fair Stacking Meta-Learning framework that allows for the reduction of bias in machine learning classifiers. By filtering and combining the base models in an optimal way, the ensemble presented in this paper achieves state-of-the-art fairness and predictive performance on the benchmark data set. The Fair Stacking Ensemble demonstrated higher accuracy and much lower bias for protected groups as measured by Equal Opportunity Difference.

These results suggest meta-learning approaches using fairness-aware constraints can be extremely effective for mitigating algorithmic bias in complex socio-technical systems. Future work will look to explore dynamic fairness constraints, real time adaptation and integration of causal fairness measures in order to produce fairer machine learning.

This study is limited to using a single dataset, which may limit the external validity of the findings. Furthermore, the assessment of fairness relied solely on the Equal Opportunity Difference (EOD) measure. Future work could incorporate additional fairness measures (demographic fairness, equal opportunity, individual fairness) and apply the proposed model to different datasets, such as finance, criminal justice, and healthcare.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgment

The authors are grateful to the Directorate of Educational Nineveh for their valuable support to this work.

References

[1] D. Patel, "Ethical AI: Addressing Bias and Fairness in Machine Learning Models for Decision-making," *Journal of Computer Science and Technology Studies*, vol. 3, pp. 13–17, 2023, doi: 10.32996/jcsts.2021.3.1.3.

[2] 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, Jul. 2021, doi: 10.1145/3457607.

[3] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," *Sci*, vol. 6, no. 1, p. 3, Jan. 2023, doi: 10.3390/sci6010003.

[4] N. Hanna, "Ethical and bias considerations in artificial intelligence," *Journal of Biomedical Informatics*, vol. 150, p. 104266, Jan. 2025, doi: 10.1016/j.jbi.2024.104266.

[5] E. Ferrara, "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies," *Sci*, vol. 6, no. 1, p. 3, 2023, doi: 10.3390/sci6010003.

[6] B. Silva and L. Berton, "Analyzing the Trade-off Between Fairness and Model Performance in Supervised Learning: A Case Study in the MIMIC dataset," in *Proc. SBCAS*, 2025, pp. 212–223, doi: 10.5753/sbcas.2025.6994..

[7] G. Raftopoulos, G. Davrazos, and S. Kotsiantis, "Evaluating Fairness Strategies in Educational Data Mining: A Comparative Study of Bias Mitigation Techniques," *Electronics*, vol. 14, 2025, doi: 10.3390/electronics14091856.

[8] S. Shaham, A. Hajisafi, M. Quan, D. C. Nguyen, B. Krishnamachari, C. Peris, G. Ghinita, C. Shahabi, and P. Pathirana, "Privacy and Fairness in Machine Learning: A Survey," *IEEE Trans. Artif. Intell.*, pp. 1–21, 2025, doi: 10.1109/TAI.2025.3531326.

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 49, No. 2, 2023, pp. 243-251
DOI: https://doi.org/10.25195/ijci.v51i2.629
Print ISSN: 2313-190X, Online ISSN: 2520-4912

[9] V. De Martino, G. Voria, C. Troiano, G. Catolino, and F. Palomba, "Examining the impact of bias mitigation algorithms on the sustainability of ML-enabled systems: A benchmark study," *J. Syst. Softw.*, vol. 230, p. 112458, 2025, doi: 10.1016/j.jss.2025.112458.

[10] E.-L. Silva-Ramírez, J. F. Cabrera Sánchez, and M. López-Coello, "REDIBAGG: Reducing the training set size in ensemble machine learning-based prediction models," *Eng. Appl. Artif. Intell.*, vol. 149, p. 110382, 2025, doi: 10.1016/j.engappai.2025.110382..

[11] K.-L. Du, R. Zhang, B. Jiang, J. Zeng, and J. Lu, "Foundations and Innovations in Data Fusion and Ensemble Learning for Effective Consensus," *Mathematics*, vol. 13, p. 587, 2025, doi: 10.3390/math13040587.

[12] U. Gohar, S. Biswas, and H. Rajan, "Towards Understanding Fairness and its Composition in Ensemble Machine Learning," in *Proc. 45th Int. Conf. Software Engineering (ICSE)*, 2023, pp. 1533–1545, doi: 10.1109/ICSE48619.2023.00133.

[13] N. Buslim, Zulfiandri, and L. Oh, "Ensemble learning techniques to improve the accuracy of predictive model performance in the scholarship selection process," *J. Appl. Data Sci.*, vol. 4, pp. 264–275, 2023, doi: 10.47738/jads.v4i3.112.

[14] Md. A. Islam, A. Nag, N. Roy, A. Dey, S. Fahim, and A. Ghosh, "An Investigation into the Prediction of Annual Income Levels Through the Utilization of Demographic Features Employing the Modified UCI Adult Dataset," in *Proc. Int. Conf. Comput. Commun. Inf. Sci. (ICCCIS)*, 2024, pp. 1080–1086, doi: 10.1109/ICCCIS60361.2023.10425394.

[15] L. Sena and J. Machado, "Evaluation of Fairness in Machine Learning Models using the UCI Adult Dataset," in *Proc. SBBd*, 2024, pp. 743–749, doi: 10.5753/sbbd.2024.243650.

[16] S. Uddin, H. Lu, A. Rahman, and J. Gao, "A novel approach for assessing fairness in deployed machine learning algorithms," *Sci. Rep.*, vol. 14, 2024, doi: 10.1038/s41598-024-68651-w.

[17] P. Kenfack, A. Khan, S. M. Kazmi, R. Hussain, A. Oracevic, and A. Khattak, "Impact of Model Ensemble On the Fairness of Classifiers in Machine Learning," in *Proc. Int. Conf. Artif. Intell. (ICAPAI)*, 2021, pp. 1–6, doi: 10.1109/ICAPAI49758.2021.9462068.

[18] J. Consuegra Ayala, Y. Gutiérrez, Y. Almeida Cruz, and M. Palomar, "Bias mitigation for fair automation of classification tasks," *Expert Syst.*, vol. 42, 2024, doi: 10.1111/exsy.13734.

[19] https://www.kaggle.com/datasets/sagnikpatra/uci-adult-census-data-dataset

[20] M. Untoro, M. Praseptiawan, M. Widianingsih, I. F. Ashari, A. Afriansyah, and Oktafianto, "Evaluation of Decision Tree, K-NN, Naive Bayes and SVM with MWMOTE on UCI Dataset," *J. Phys.: Conf. Ser.*, vol. 1477, p. 032005, 2020, doi: 10.1088/1742-6596/1477/3/032005..

[21] D. Mccarty, H. W. Kim, and H. Lee, "Evaluation of Light Gradient Boosted Machine Learning Technique in Large Scale Land Use and Land Cover Classification," *Environments*, vol. 7, p. 84, 2020, doi: 10.3390/environments7100084.

[22] V. N. Dang, T. T. Nguyen, D. T. Hoang, and E. Dutkiewicz, “Fairness-aware machine learning for credit risk prediction: An empirical study,” Scientific Reports, vol. 14, no. 1, p. 58427, Feb. 2024, doi: 10.1038/s41598-024-58427-7.

[23] F. He, X. Wu, W. Zhang, and X. Huang, "A novel bias-alleviated hybrid ensemble model based on over-sampling and post-processing for fair classification," *Connection Sci.*, vol. 35, pp. 1–20, 2023, doi: 10.1080/09540091.2023.2184310.

[24] C. Miller, T. Portlock, D. Nyaga, and J. O'Sullivan, "A review of model evaluation metrics for machine learning in genetics and genomics," *Front. Bioinform.*, vol. 4, 2024, doi: 10.3389/fbinf.2024.1457619.