

Predicting Without Prejudice: A Deep Learning-Based Bias Mitigation Approach to Student Placement Prediction

Anjali Jindia, Sonal Chawla

Deptt. of Computer Science and Applications, Panjab University, Chandigarh, India.

DOI: <https://doi.org/10.51583/IJLTEMAS.2025.1408000024>

Abstract: In Educational Data Mining (EDM), ensuring accurate and fair student placement prediction is essential for fostering equal opportunities and minimizing biases that may disadvantage certain student groups. This study develops a bias mitigation and explainability framework to enhance fairness and transparency in predictive modeling. Recognizing that data bias can skew prediction outcomes, the study explores various bias mitigation techniques, including re-sampling, re-weighting, and adversarial debiasing, to balance the dataset and ensure equitable representation across student groups.

Deep Learning (DL) models are deployed on both the original and bias-mitigated datasets to analyze differences in placement predictions. The results reveal significant disparities in prediction outcomes, highlighting that bias mitigation enhances both predictive accuracy and fairness. Additionally, the integration of explainability techniques, such as SHAP (Shapley Additive Explanations) values, provides insights into feature contributions, promoting transparency and trust in AI-driven decisions.

This study underscores the critical importance of addressing bias in EDM and advocates for the integration of bias mitigation and explainability methods to ensure fair and equitable placement predictions. By doing so, it contributes to the development of ethical, accountable, and transparent AI systems in education, supporting data-driven, unbiased decision-making in student placement processes.

Keywords: Bias Mitigation, Data Bias, Predictive Analytics, Algorithmic Fairness, Adversarial Debiasing, Re-sampling Techniques

I. Introduction

Today's data-driven education systems make extensive use of Deep Learning (DL) models to provide predictions within the student placement field, providing educational institutions with valuable data to optimize student accomplishment and career advancement. With the increased importance of such models in the decision-making mechanism, bias against educational data has become a cause of concern [1]. Data bias can result in imbalanced predictions, disproportionately affecting some student groups, thereby perpetuating inequality and threatening equal opportunity and fairness. This is even more concerning in education, where the predictions influence course turnout, career development, and admission and are ultimately responsible for shaping a student's academic and professional life [2]. For more accurate predictions and reliability and ensuring that students are assessed fairly, the removal of such bias is vital regardless of demographic characteristics.

The purpose of this research is to create a holistic Bias Mitigation and Explainability Framework designed specifically for EDM placement prediction. The research improves fairness, transparency, and reliability in prediction models by systematically discovering and addressing student dataset bias [3]. In order to equalize the dataset and diminish the effect of the biased features, different bias mitigation strategies, like re-sampling, re-weighting, and adversarial debiasing, are applied [4]. In addition, with the inclusion of Explainability Techniques like SHAP (Shapley Additive Explanations) values, explainable AI-driven predictions are assured, and stakeholder confidence and transparent decision-making are promoted [5].

The study emphasizes the bias effect in the predictions of placement and highlights why there is a need to adopt ethical AI practices at schools by testing DL models both with the initial and the bias-reduced dataset. The results help to promote equitable, accountable, and transparent AI systems that ensure that the equity and inclusiveness values are upheld at schools and that student placement processes are made unbiased and data-driven.

II. Literature Review

Student placement prediction and academic decision-making have become sophisticated with the deployment of Machine Learning (ML) algorithms in Education Data Mining (EDM) but have been found to have extreme bias issues in prediction models based on demographics such as gender, race, and socio-economic status, to the point of discriminating against some student groups [1]. Recent research has looked at reducing bias and methods of explainability to encourage fairness and transparency in prediction modelling.

For example, the integration of Explainable AI (XAI) methods like LIME and SHAP in transparent education predictions has been highlighted in [6]. Additionally, the need for fairness measures in AutoML-generated model quality assessment has been highlighted by Huang et al. [7], with a demonstration of improvement of placement predictions by fairness-aware learning algorithms. The effectiveness of adversarial debiasing in delivering equal predictions and thus minimizing the dependence over the sensitive features has also been demonstrated by the researchers in [8]. The integration of SHAP-based feature importance analysis, highlighted by Lundberg and Lee [9], has also promoted more transparency in model decision-making, supporting

stakeholder trust. Moreover, the importance of purpose-built data collection methodologies in bias minimization in education recommenders has been highlighted by Combrink and Rosman [10], promoting equal outcomes.

The literature highlights that ML methods have extensively been used in student placement predictions, but DL methods have not been exhaustively explored, and the overall potential of DL-based systems is not maximized in EDM applications. Moreover, joint use of bias avoidance and explanation techniques within EDM is not extensively researched although it has the capability to enhance fairness and prediction accuracy.

With the aim to overcome such deficiencies, this research aims to develop a holistic DL framework with bias avoidance and explanation components to ensure equitable, transparent, and reliable placement predictions at academic institutions.

Aim and Objectives

For promoting fairness and accuracy of student placement prediction, this research seeks to build a Bias Mitigation and Explainability framework. For this purpose, the study emphasizes building and assessing DL models, applying fairness-aware bias mitigation schemes, and evaluating model performance with and without bias mitigation. Further, Explainability techniques like SHAP are incorporated to analyze feature contributions, illustrating the effect of bias mitigation to prediction fairness and decisional transparency. Such practices help build responsible, bias-free, and transparent AI systems in accordance with fairness standards in education decision-making.

III. Methodology

The research methodology for this study followed a structured approach. It has been divided into 3 phases as shown below in figure I.

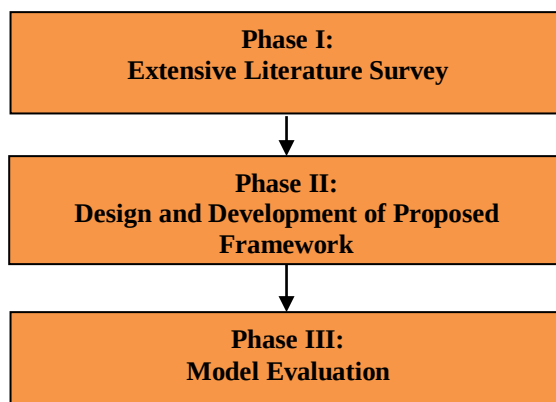


Fig. I: Research Methodology

Phase I

The first phase involved literature review on DL, user-centric design, bias mitigation, and explainability models in educational predictive modeling. Existing frameworks have been analyzed for strengths and limitations, while relevant student datasets have been gathered for model development. Building on insights from the literature review, a DL-based framework has been proposed for student placement prediction, leveraging its advantages in automatic feature learning, capturing complex relationships, scalability, and superior adaptability over traditional ML models.

DL Techniques identified from Literature Review

To develop an accurate and fair placement prediction system, this study employed **four DL models**, each designed to capture different aspects of the dataset and improve predictive performance. These models include:

Fully Connected Neural Network (FCNN) / Feedforward Neural Network (FNN): A DL model in which every neuron in one layer is entirely connected to all neurons in the subsequent layer. It comprises an input layer, one or more hidden layers, and an output layer, employing activation functions like ReLU or Sigmoid to process and classify data [11].

Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM): Designed for sequential data processing, LSTMs maintain memory over time, making them effective in capturing temporal dependencies. Unlike standard RNNs, LSTMs incorporate forget, input, and output gates, which help in selectively storing and updating information, improving learning efficiency [12].

Autoencoder: Autoencoder is a type of artificial neural network that is specifically trained under unsupervised learning. The only function of the autoencoder is to learn a compact lower-dimensional data representation (encoding) and decode data back to its actual format. It consists of an Encoder (compressing the input features), a Bottleneck layer (extracting latent representations),

and a Decoder (reconstructing input data). Autoencoders are widely applied in noise reduction, outlier removal, and deep learning (DL) pre-learning [13].

Radial Basis Function Network (RBFN): This network utilizes radial basis functions (like Gaussian functions) in a hidden layer to identify non-linear patterns within data. RBFNs measure data samples' proximity to hidden neuron centers and are extremely effective in classification, regression, and clustering applications [14].

All of these models are assessed against both the initial and bias-reduced datasets, enabling a comparative evaluation of fairness, accuracy, and interpretability of student placement predictions.

Bias Metrics and Mitigation Techniques

The numerical methods that are utilized to measure and detect disparities in ML models, especially within classification problems are referred to as **Bias metrics**. These metrics are useful in identifying whether a model is treating various demographic classes equally or exhibits discriminatory patterns that can result in unfair outcomes. AI system bias can stem from historical inequities, imbalanced data, or algorithmic decision-making, making it essential to evaluate and mitigate such biases to ensure fair and ethical AI applications.

Bias metrics are essential for measuring fairness in student placement prediction models within the context of this research. These metrics provide insight into whether the model over-represents or under-represents specific student groups based on gender, socioeconomic status, or ethnicity by identifying systematic imbalances in placement outcomes across demographic segments. The results from these evaluations guide the development of bias mitigation strategies, ensuring that the placement prediction system is equitable, transparent, and dependable.

Some of the common metrics are presented in Figure II:

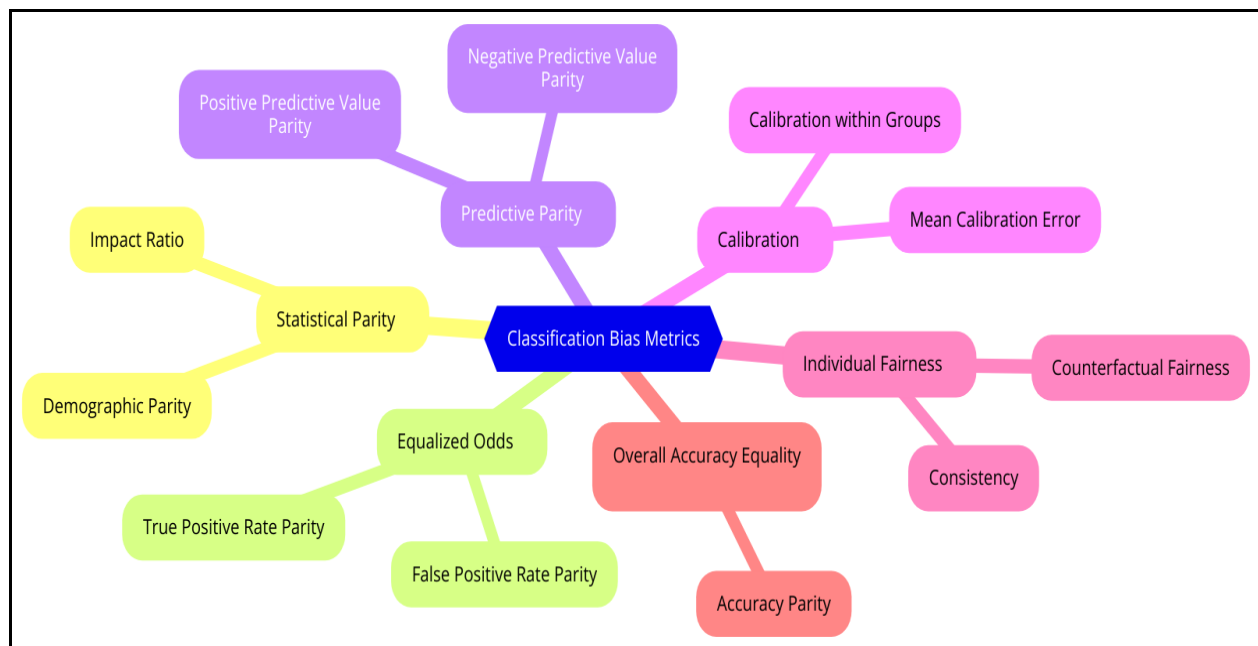


Fig. II: Classification Bias Metrics

- Statistical parity ensures that different demographic groups receive positive outcomes at equal rates [15].
- Disparate impact occurs when one group receives positive outcomes at a significantly lower rate than another, leading to indirect discrimination [16].
- Four-Fifths Rule: It is a test of fairness that dictates that a protected group ought to be allocated at least 80% of the more privileged group's favorable result rate [17].
- Cohen's d: The difference between two group means with respect to standard deviations to evaluate bias effect is measured by this [18].
- 2 Standard Deviation checks if the difference in selection rates between groups is statistically significant, using a z-test approach [19].
- Demographic Parity assures regardless of qualifications, that the likelihood of obtaining a positive outcome is equal across all demographic groups [3].

- Equalized Odds ensures fairness by requiring equity among both true positive rates (TPR) and false positive rates (FPR), across groups [15].

To address and reduce biases in machine learning models, various bias mitigation techniques are applied at different stages of the DL pipeline. These techniques fall into three main categories pre-processing, in-processing, and post-processing methods [1] as shown below in figure III and IV:

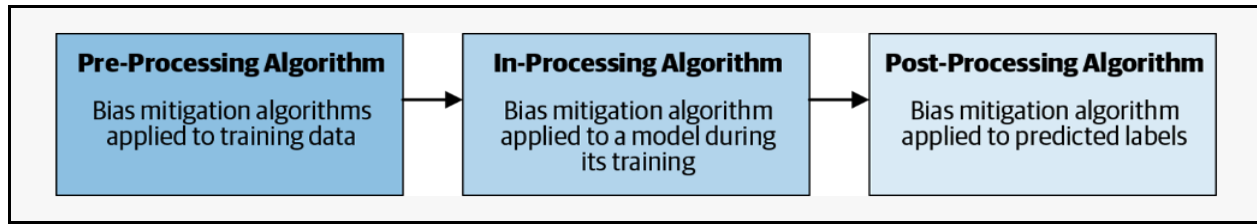


Fig. III: Categories of Bias Mitigation Algorithms

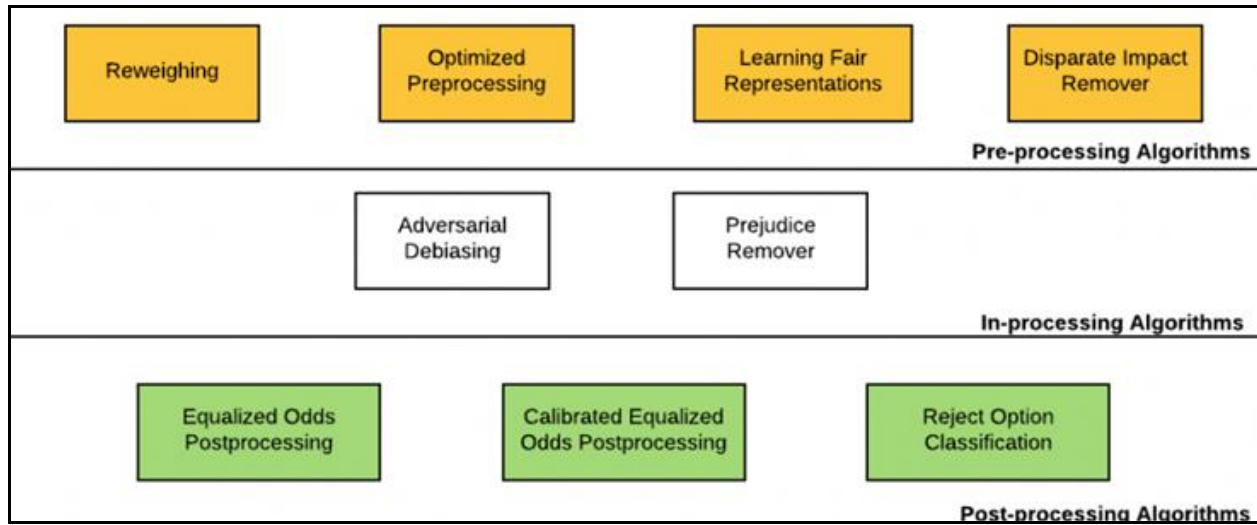


Fig. IV: Bias Mitigation Techniques within each category

➤ Pre-processing Techniques (Applied Before Model Training)

Pre-processing techniques modify the dataset before training to ensure fair representation across demographic groups. They focus on reducing bias at the data level, making models less susceptible to unfair patterns [2].

- Resampling: It entails class distributional adjustment with oversampling the under-represented or undersampling over-represented groups [20].
- Re-weighting: Applying different weights to the training samples depending upon their group representation [21].
- Fair Representation Learning: Projecting data to a new feature space where predictions are least sensitive to the protected attributes [22].

➤ In-processing Techniques (Applied During Model Training)

It entails adjusting the learning algorithm to limit discrimination when training to lower bias when optimizing the model [4].

- Adversarial Debiasing: Trains a model with an adversary network that punishes bias in predictions [23].
- Fair Regularization: Functions by incorporating fairness constraints to the loss function of the model to decrease the reliance upon sensitive features [24].
- Fairness-Aware Neural Networks: Involves the tuning of hidden layers to learn unbiased feature representations [25].

➤ Post-processing Techniques (Applied After Model Training)

These techniques adjust model outputs after predictions are made to correct remaining biases, thereby ensuring fair treatment of all demographic groups without modifying the dataset or training process [26].

- Equal Opportunity: Ensures that true positive rates remain similar across demographic groups [26].

- Calibrated Equalized Odds: Adjusts decision thresholds so that false positive and false negative rates remain equal across groups [27].
- Reject Option Classification: Modify predictions near the decision boundary to favor fairness [28].

A comprehensive framework for developing fair, transparent, and accountable ML models is provided by these bias mitigation strategies. Their application is particularly critical in EDM, where placement predictions must be unbiased, equitable, and trustworthy to ensure equal opportunities for all students.

Phase II

Figure V shows the design and development phases of the proposed framework.

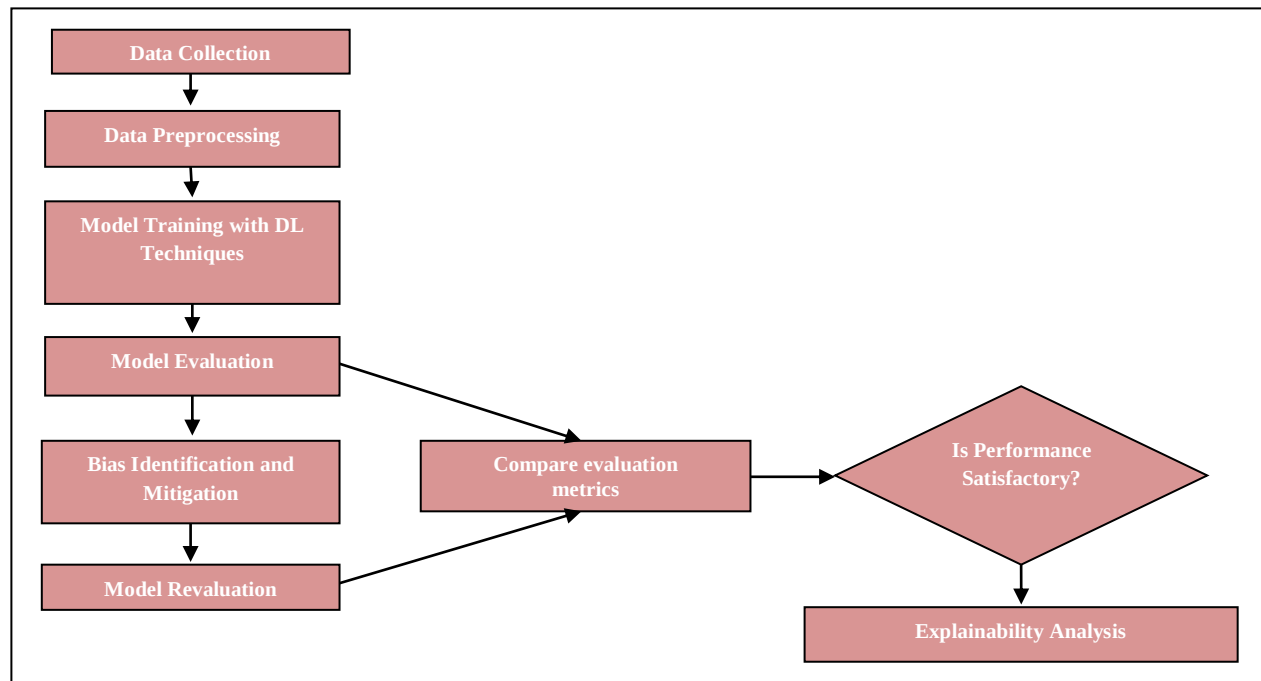


Fig. V: Stages of Framework Design and Development

Data Collection

The dataset used in this study was sourced from Kaggle, comprising 13 features encompassing numerical, categorical, and binary attributes. The dataset contains a total of 670 instances, providing a diverse representation of student placement records. Figure VI presents a snapshot of the dataset structure.

gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	etest_p	mba_p	specialisatworkex	status	
M	73.97697	Central	77.39269	Others	Science	73.62315	Sci&Tech	84.49831	67.20701	Mkt&HR	Yes	Placed
M	62	Others	62	Others	Commerce	60	Comm&M	63	52.38	Mkt&HR	Yes	Placed
M	80.96524	Others	67.10427	Others	Science	72.48262	Sci&Tech	91.34359	71.39231	Mkt&Fin	No	Placed
M	52	Others	65	Others	Arts	57	Others	75	59.81	Mkt&Fin	Yes	Not Placed
F	69	Central	62	Central	Science	66	Sci&Tech	75	67.99	Mkt&HR	No	Not Placed
M	78.76237	Others	78.66305	Others	Commerce	70.28107	Comm&M	67.44177	65.31805	Mkt&Fin	Yes	Placed
M	51.7945	Others	40.9856	Others	Science	62.12665	Others	65.86247	52.24975	Mkt&HR	No	Not Placed
M	59.80592	Central	61.45148	Others	Commerce	60	Comm&M	62.45148	57.02596	Mkt&HR	Yes	Placed
M	77.35132	Central	74.83467	Others	Science	73.48771	Sci&Tech	80.12399	68.00166	Mkt&HR	Yes	Placed
M	76.5	Others	97.7	Others	Science	78.86	Sci&Tech	97.4	74.01	Mkt&Fin	No	Placed
M	60.96136	Others	49.3827	Others	Science	54.50354	Others	70.3768	61.72612	Mkt&HR	No	Not Placed
F	73.80921	Central	81.21382	Central	Arts	58.65793	Others	59.23355	63.1802	Mkt&HR	Yes	Placed
M	80.91169	Others	60.9363	Others	Science	62.6819	Sci&Tech	89.78997	64.95734	Mkt&Fin	Yes	Placed
M	73.592	Central	78.39467	Others	Science	73.98667	Sci&Tech	87.368	66.7476	Mkt&HR	Yes	Placed
M	70.9782	Central	68.63551	Central	Science	72	Sci&Tech	57.38193	57.46757	Mkt&HR	No	Placed
M	60.08876	Others	44.52812	Others	Science	54.73006	Others	72.9946	58.51853	Mkt&HR	No	Not Placed
M	67	Others	61	Central	Science	72	Comm&M	72	61.01	Mkt&Fin	No	Placed
F	52	Others	52	Others	Science	55	Sci&Tech	67	59.32	Mkt&HR	No	Not Placed
F	48	Central	51	Central	Commerce	58	Comm&M	60	58.79	Mkt&HR	Yes	Not Placed
M	60.38914	Others	45.37504	Others	Science	54.22145	Others	72.53511	58.84182	Mkt&HR	No	Not Placed
M	79.39641	Others	61.39412	Others	Science	67.07126	Sci&Tech	92.93826	66.43687	Mkt&Fin	No	Placed
M	82	Others	63.56546	Others	Science	71.40667	Sci&Tech	94.98607	70.85456	Mkt&Fin	No	Placed

Fig.VI: Screenshot of dataset used

Data analysis

The dataset consists of 670 entries, with all fields containing non-null values ensuring completeness. Among these, 439 entries (65.52%) indicate a positive placement status, while 231 entries (34.48%) reflect a negative placement outcome, as depicted in Figure VII. Additionally, the gender distribution analysis reveals a significant imbalance with 475 male entries compared to only 195 female entries, as shown in Figure VIII. Furthermore, the dataset was found to be free from duplicate records, ensuring data integrity for subsequent analysis.

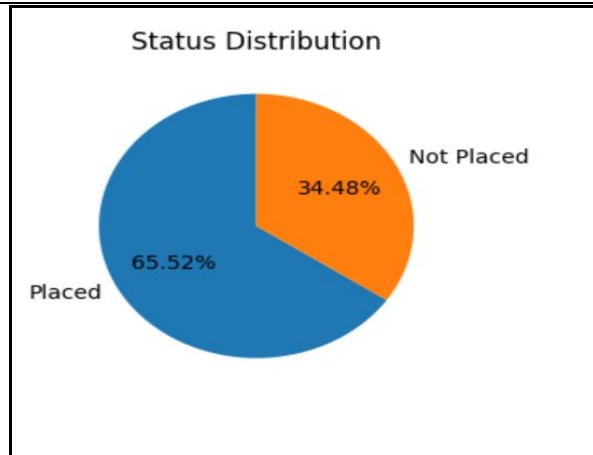


Fig. VII: Placement status percentage

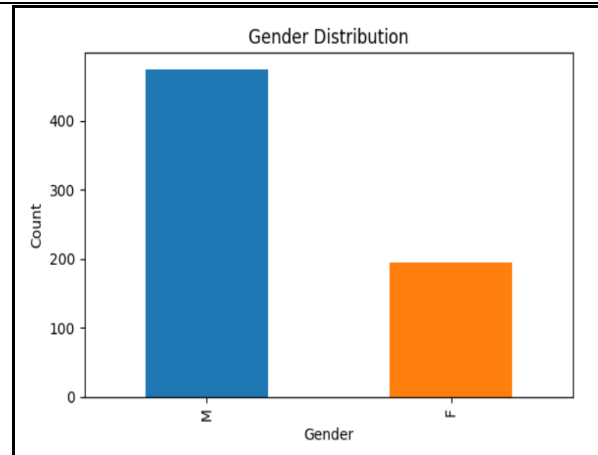


Fig. VIII: Gender status percentage

Data Preprocessing

To ensure a clean, consistent, and ML-compatible dataset, several preprocessing steps were applied to enhance data quality, model performance, and fairness in placement predictions. Missing values were handled to maintain data integrity, while categorical variables were one-hot encoded for numerical representation. Numerical features were normalized/scaled to a standardized range (e.g., 0 to 1 or -1 to 1) to ensure equal contribution across attributes. Data balancing techniques were implemented to address class imbalances, improving fairness in predictions. Additionally, feature engineering was performed to create informative features that enhance predictive accuracy. These preprocessing steps optimize the dataset for robust model training, enabling reliable and unbiased placement predictions.

Performance metrics

Multiple evaluation metrics have been employed, like Accuracy, Precision, Recall, and F1-score to evaluate the performance of the DL models. These metrics provide a comprehensive understanding of the model's effectiveness, particularly in classification tasks where fairness and predictive reliability are critical.

Accuracy: It gives the proportion of correctly predicted instances out of the total predictions and is more effective for balanced datasets as compared to imbalanced data [29].

Accuracy = $(TP + TN) / (TP + TN + FP + FN)$ where:

TP = True Positives, **TN** = True Negatives, **FP** = False Positives, **FN** = False Negatives

Precision: Among all positive predictions, it determines the proportion of true positive predictions and is particularly important when minimizing false positives is a priority, such as in fraud detection or medical diagnoses where incorrect positive classifications could have significant consequences [30].

Precision = $(TP) / (TP + FP)$

Recall: The percentage of true positives over all actual positives is indicated by Recall. It is essential in applications where false negatives should be avoided at all costs, and the majority of positive cases should be correctly identified [31].

Recall = $TP / (TP + FN)$

F1-Score: The harmonic average of precision and recall, reflecting the well-balanced measurement of model quality is depicted by F1-score. It is particularly beneficial when working with imbalanced data, since it reflects the relationship between precision and recall, allowing a more complete evaluation [32].

F1-Score = $2 \times (Precision \times Recall) / (Precision + Recall)$

The bias-mitigated DL models not only accurate but also fair and reliable in predicting student placement outcomes is ensured collectively by these metrics.

Phase III

Experimental Results

The preprocessed data were input to the suggested framework upon which the identified DL methods were subsequently applied. A comparative analysis of four DL models—Fully Connected Neural Network (FCNN), Autoencoder, Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM), and Radial Basis Function Network (RBFN)—on the basis of the crucial classification measures: Accuracy, Precision, Recall, and F1-score is illustrated in figure IX. The findings show that the FCNN

and Autoencoder models have the highest overall result, with accuracy and precision scores above 85%, and a robust balance in recall and F1-score.

The LSTM-based RNN also fares competitively, with specialized competence in recall, reflecting its efficacy in efficiently extracting sequential dependencies within the data. Whereas, although RBFN attains highest in terms of recall, it has lower precision and F1-score, reflecting a trade-off between sensitivity and specificity. The highly varying recall values compared to different models reveal the influence of various network structures upon classification.

The RNN with LSTM network is seen to be the optimal fit in student placement prediction based on the results of Figure IX, with the highest overall accuracy, precision, and F1-score, and this makes it the most reliable and well-balanced model to be applied to this task.

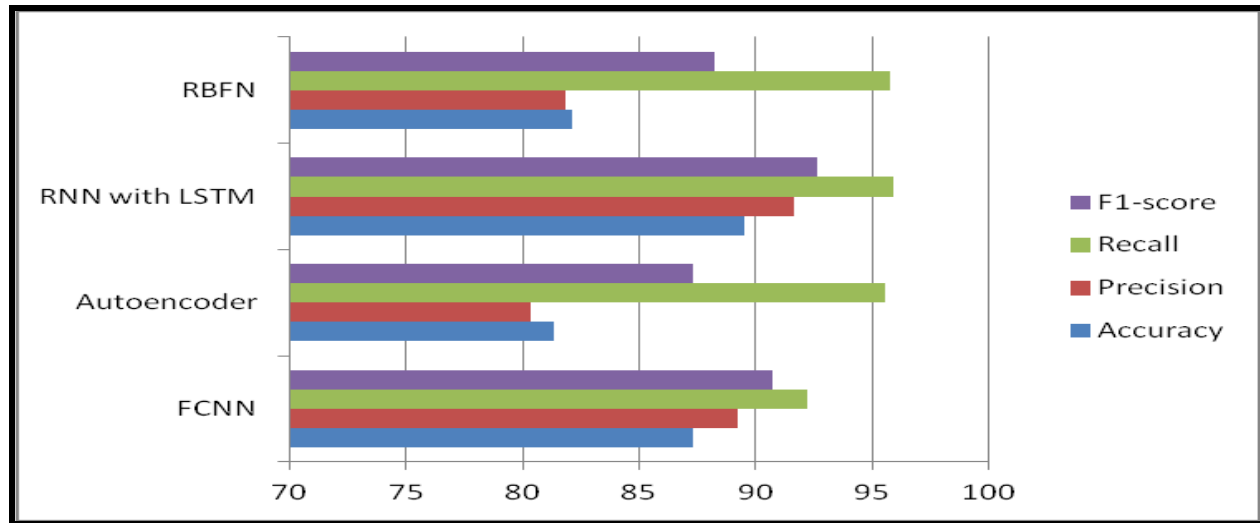


Fig. IX: Evaluation of DL Techniques

The study computed fairness metrics to assess potential biases in the model's predictions, in order to investigate the factors contributing to lower accuracy. A comparative analysis of these fairness metrics is presented in figure X by evaluating the actual bias values observed in the student placement prediction model against their respective reference values. The fairness metrics considered include Statistical Parity, Disparate Impact, Four-Fifths Rule, Cohen's d, and the 2SD Rule, each capturing different aspects of potential discrimination in the model's outcomes.

The findings reveal that Disparate Impact and the Four-Fifths Rule exceed their reference values, indicating a deviation from fairness that may disadvantage certain student groups. Also, Cohen's d, an effect size difference between the two groups, compares closely with the reference point, reflecting margins of bias in the predictions. The 2SD Rule, however, deviates markedly, reflecting a statistical bias that is beyond acceptable limits of fairness. The steep incline in the metric of 2SD also reflects systemic bias in prediction confidence by the model, with a need to rectify and ensure equity and fairness in placements predictions.

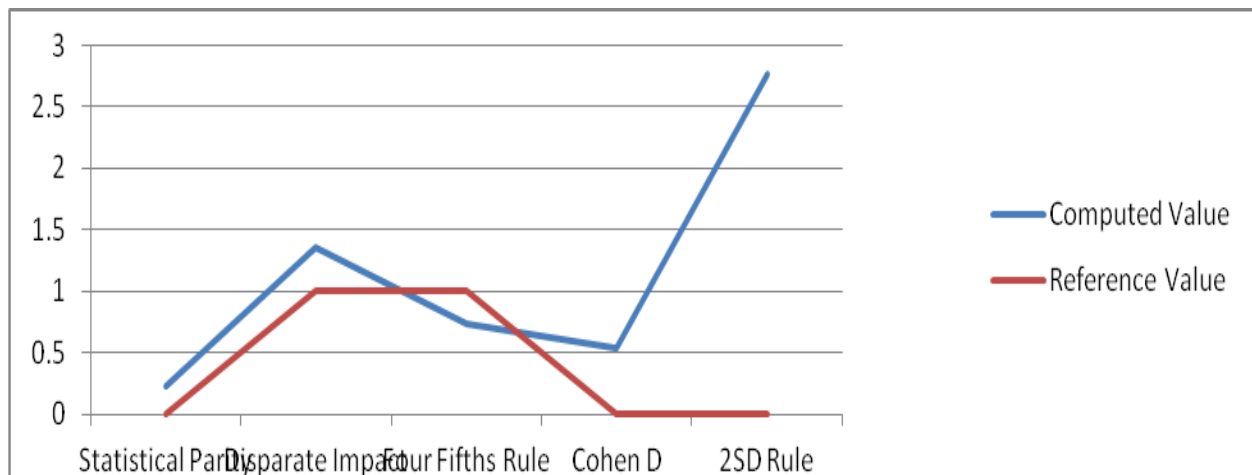


Fig. X: Bias Metrics

The research utilized Bias Mitigation methods, such as Reweighting, GridSearch Reduction, and Calibrated Equalized Odds, to reduce disparities in placement predictions upon establishment of bias by comparative analysis between calculated fairness metrics and their reference values. The post-mitigation analysis also showed significant improvement in fairness measures with declines in Disparate Impact, breaches of the Four-Fifths Rule, and 2SD Rule violations, reflecting a more just decision process, as seen in figure XI. But with slight sacrifices in accuracy of the model, emphasizing the need for a trade-off approach that retains both fairness and prediction reliability.

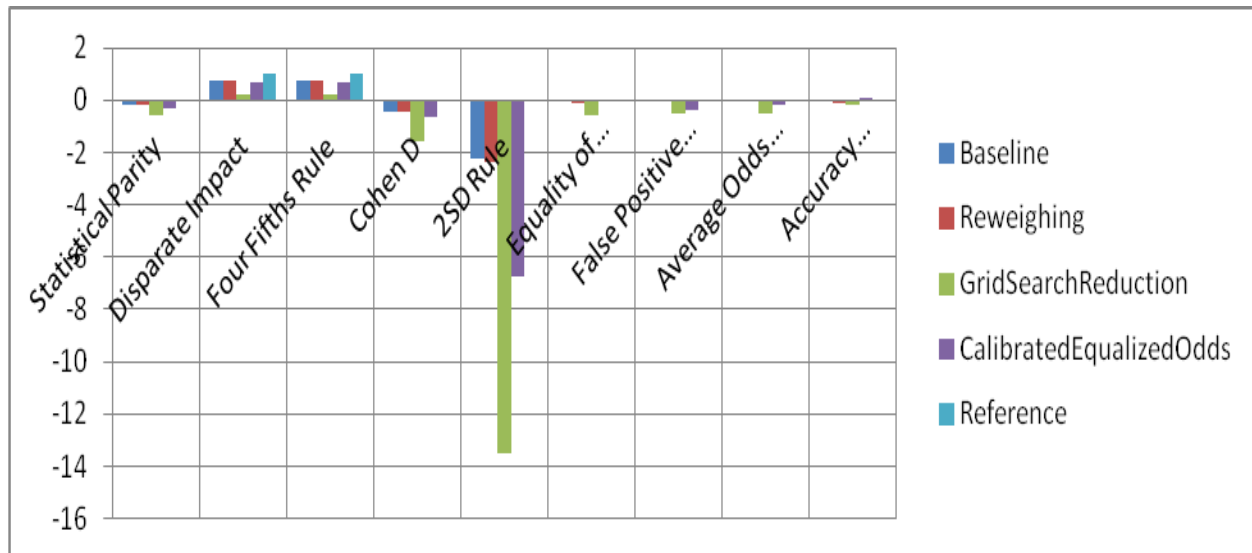


Fig. XI Comparative Analysis of Bias Mitigation Techniques on Fairness Metrics

The research employed SHAP (Shapley Additive Explanations) to examine feature contributions and their effect on placement prediction to further improve the explainability and trustworthiness of the bias-mitigated model. In order to understand the resultant model decision process after mitigation, several different visualizations of SHAP were created. Global feature importance over predictions viewed at a population level, with emphasis placed on those factors that have a significant effect on predictions is demonstrated by the summary plots. Interactions between influential features and predicted values, with a deeper understanding of feature interactions that impact placement choices are analyzed by SHAP dependence plots. Force plots and decision plots were used to provide feature-level explanations, allowing for a more transparent assessment of whether bias still existed in certain population subgroups. Results of the produced SHAP analysis reaffirm the efficacy of bias mitigation approaches and ensure the resultant model is still interpretable, fair, and aligns with ethical AI practices.

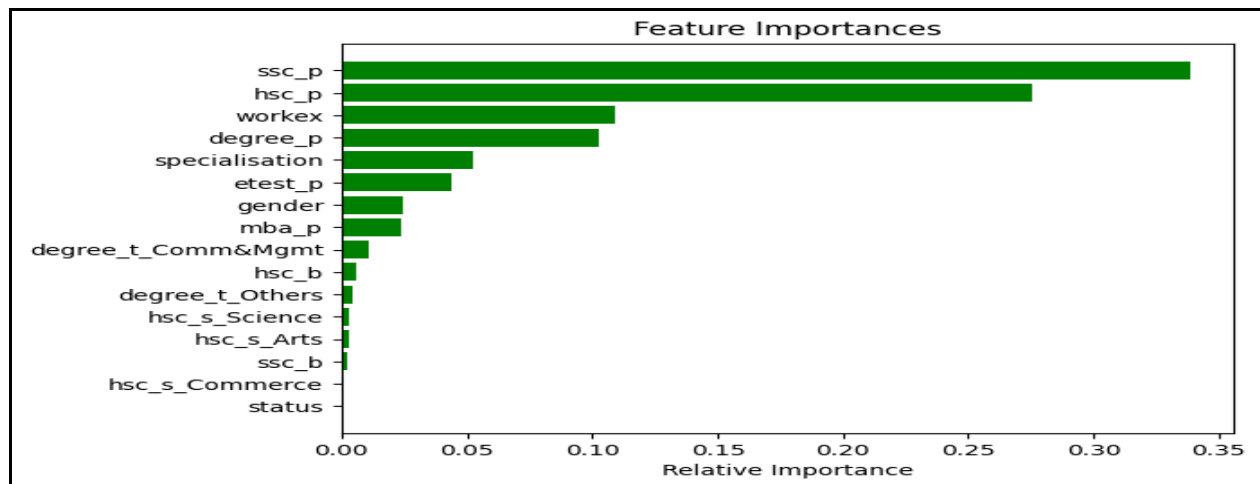


Fig. XII: Feature importance based on SHAP values

The feature importance analysis of Figure XII indicates that the two topmost influential factors in student placement predictions are secondary (ssc_p) and higher secondary education percentage (hsc_p), followed by work experience (workex) and degree percentage (degree_p). Moderately influential are specialization, employment test scores (etest_p), and gender. Categorical features such as degree type and affiliation with a school board have minimal influence. Such findings demonstrate that the model gives more emphasis to academic achievement and work experience compared to demographics, promoting fairness requirements and a merit-based assessment of the placement results.

The plot of SHAP summaries displayed in Figure XIII reveals the relative influence of each feature upon predictions of placement, with the intensity of colour reflecting feature values (red for highs, blue for lows) and horizontal location reflecting their effect. The findings affirm that increased secondary scores, work experience, and degree percentage elevate placement likelihood, with gender and specialization showing more scattered effects, hinting at the presence of non-linear effects. The findings identify academic attainment and work experience as leading factors in delivering placement with a significant implication of these characteristics in student placement.

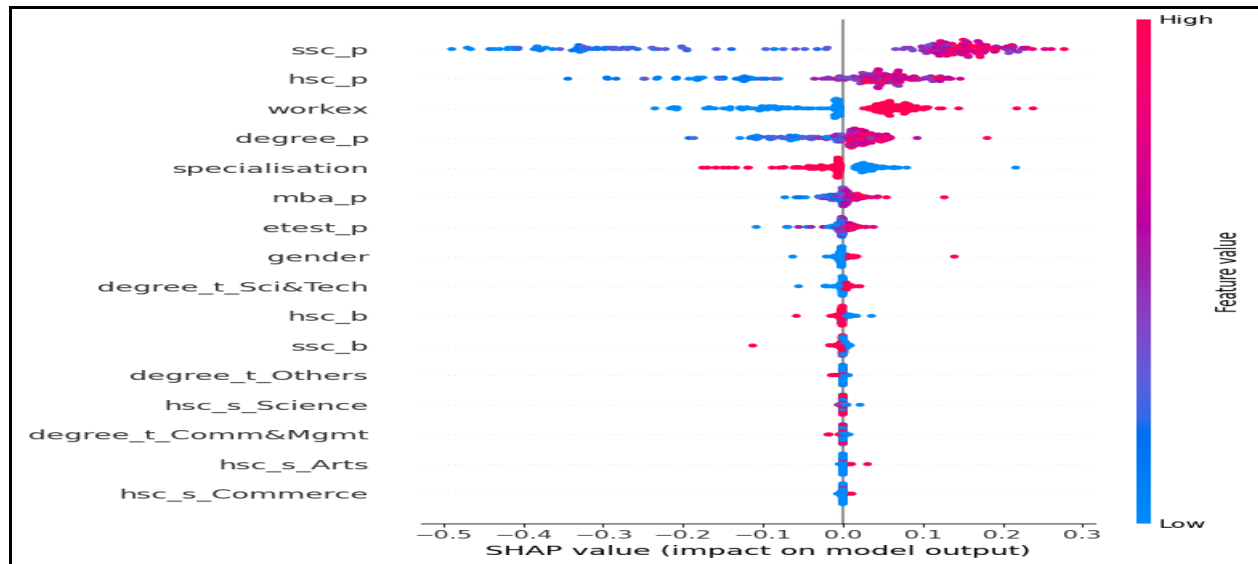


Fig. XIII: SHAP Summary Plot

The SHAP force plot delivers instance-level interpretability by illustrating the effect that each feature has on a particular placement prediction, with positive contributors boosting placement likelihood and negative contributors lowering the same, as illustrated in Figure XIV. In this instance, higher secondary scores (hsc_p), work experience (workex), and secondary education scores (ssc_p) are strong supporters of placement, while a low degree percentage contributes negatively. This analysis validates that predictions made by the model align with anticipated human thinking and ensure transparency in AI-based placement suggestions.

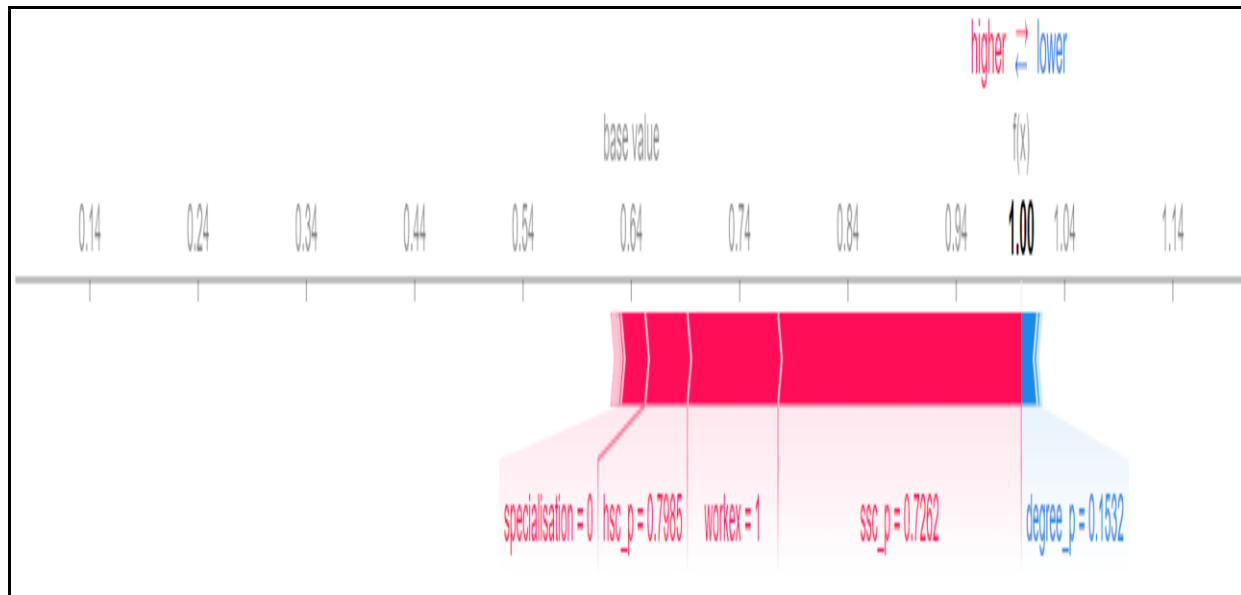


Fig. XIV: SHAP Force Plot

After applying the XAI SHAP technique to the bias-mitigated model, the metrics were re-computed with respect to the model that had the highest accuracy. As presented in Figure XV, the results reveal significant increases in performance measures, confirming that combining bias mitigation and explainability algorithms improves both fairness and reliability of predictions. As presented in the included graph, the increases affirm the efficacy of the new approach in improving model transparency and decisioning and promoting a more just and understandable student placement prediction system.

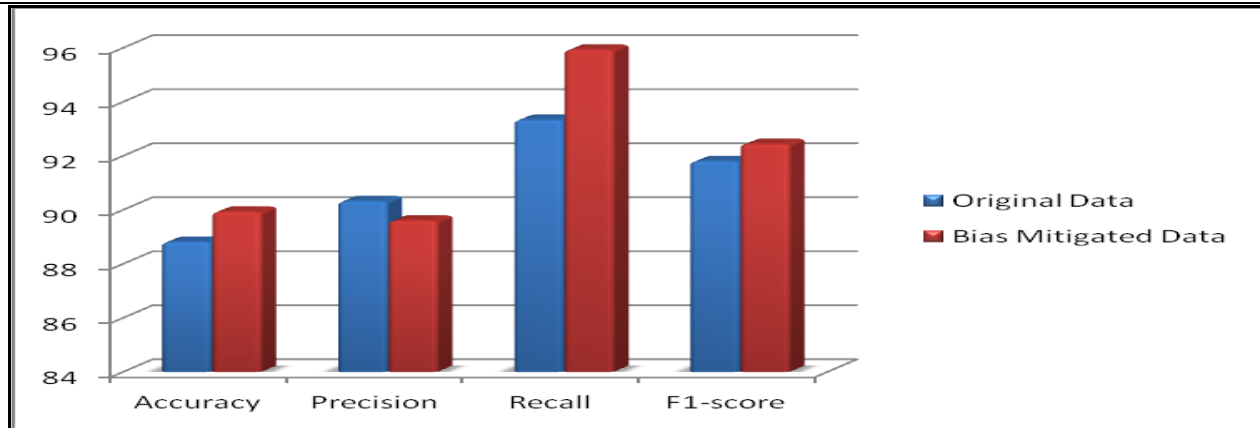


Fig. XV: Performance Comparison of FCNN on original and Bias Mitigated Dataset

IV. Discussion

While the individual components employed in this study—such as re-weighting, adversarial debiasing, and SHAP-based explainability—are established techniques in the field of AI ethics, the novelty of this research lies in their systematic integration within a unified framework specifically tailored to Educational Data Mining (EDM) for placement prediction. To the best of our knowledge, this is the first study to concurrently deploy multiple deep learning models (FCNN, Autoencoder, LSTM, and RBFN), apply fairness enhancement techniques, and incorporate SHAP-based interpretability in a comparative, end-to-end manner for addressing bias in student placement outcomes. Previous works in EDM have either focused on improving predictive accuracy or fairness in isolation, but seldom on ensuring transparency, fairness, and performance together in a practical, deployable system. Our proposed FEPP (Fair and Explainable Placement Predictor) bridges this gap by offering a multi-objective, interpretable AI solution suitable for institutional adoption.

Furthermore, we contextualize our contributions within the broader landscape of ethical AI frameworks. Unlike existing models that prioritize only fairness metrics or explanation post hoc, FEPP emphasizes simultaneous optimization of accuracy and fairness, along with user-centric interpretability, thereby making it uniquely valuable for educational stakeholders.

V. Conclusion and Future Work

The study was able to effectively build a fair and interpretable student placement prediction model by incorporating bias mitigation strategies and Explainable AI (XAI) approaches. Utilization of fairness-aware procedures like Reweighting, GridSearch Reduction, and Calibrated Equalized Odds effectively controlled bias without compromising predictive accuracy. In addition, adoption of SHAP-based techniques of explainability gave useful insight into feature contributions, providing a transparent and trustworthy mechanism of decision-making. The post-mitigation analysis presented considerable improvement to fairness measures, further confirming the efficacy of the suggested method of bias mitigation and ensuring equal student placement predictions.

Future directions include enriching fairness-aware learning frameworks with the integration of causal inference techniques to properly grasp feature interdependencies and their effects on placement predictions. Future explorations of more advanced XAI methodologies, like counterfactual explanations and fairness-aware SHAP implementations, will be pursued to enhance interpretation and ensure ethical AI deployment. Extension of the dataset with actual placement histories and transfer learning for cross-domain adaptability will further build robustness and generalizability of the model. The overall goal is to create an AI-based placement prediction system that is not just extremely accurate but also fair, transparent, and accountable, enabling ethical AI adoption in education.

References

1. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 1-35.
2. Kamiran, F., & Calders, T. (2012). Data Preprocessing Techniques for Eliminating Discrimination in ML. *Knowledge and Information Systems*, 33(1), 1-33.
3. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214-226.
4. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI Ethics and Society (AIES)*, 335-344.
5. Lundberg, S. M., & Lee, S. (2017). A Unified Approach to Interpretable Machine Learning with SHAP. *Advances in Neural Information Processing Systems (NeurIPS)*.

6. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160.
7. Huang, J., Bell, R. D., Giraud-Carrier, C., & Han, E. (2020). Fairness-aware AutoML: Lessons learned on fairness properties in automated machine learning. *Proceedings of the 2020 IEEE International Conference on Big Data (Big Data)*, 2294-2303.
8. Li, J., & Wang, T. (2023). Adversarial debiasing for fair student placement predictions. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAT)*, 367-379.
9. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpretable Machine Learning with SHAP. *Advances in Neural Information Processing Systems (NeurIPS)*.
10. Combrink, H., & Rosman, B. (2022). Mitigating Bias in Educational Recommender Systems. *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, 12(3), 345-359.
11. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
12. Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer.
13. Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11, 3371-3408.
14. Park, J., & Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural Computation*, 3(2), 246-257.
15. Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems (NeurIPS)*.
16. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*, 259-268.
17. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671–732.
18. Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of Graduate Medical Education*, 4(3), 279–282.
19. Biddle, D. A. (2006). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing.
20. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, 42(4), 463–484.
21. Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building Classifiers with Independency Constraints. *IEEE International Conference on Data Mining (ICDM)*.
22. Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning Fair Representations. *International Conference on Machine Learning (ICML)*.
23. Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., & Zhao, J. (2017). Data Decisions and Theoretical Implications in Fairness Research. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*.
24. Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2011). Fairness-aware Classifier with Prejudice Remover Regularizer. *Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*.
25. Agarwal, A., Dudik, M., & Wu, Z. S. (2018). A Reductions Approach to Fair Classification. *Proceedings of the 35th International Conference on Machine Learning (ICML)*.
26. Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems (NeurIPS)*.
27. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On Fairness and Calibration. *Advances in Neural Information Processing Systems (NeurIPS)*.
28. Kamiran, F., Calders, T., & Pechenizkiy, M. (2018). Discrimination-Aware Decision Tree Learning. *Proceedings of IEEE International Conference on Data Mining (ICDM)*.
29. Powers DMW. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *J Mach Learn Technol*. 2011;2(1):37–63.
30. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.
31. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning*. ACM; 2006:233–240.
32. Van Rijsbergen CJ. *Information Retrieval*. 2nd ed. London: Butterworths; 1979.

Statements and Declarations:

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Both authors declare that they have no financial interests and that the dataset utilized during the current study is available in the KAGGLE repository.