

Original Article

# An Integrated Framework for Fair and Interpretable Predictive Modeling in Educational Analytics

Meenakshi Devi<sup>1</sup>, Rakesh Kumar<sup>2</sup>

<sup>1,2</sup>Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, Haryana, India.

<sup>1</sup>Corresponding Author : [meenakshi.chauhan.2001@kuk.ac.in](mailto:meenakshi.chauhan.2001@kuk.ac.in)

Received: 28 February 2026

Revised: 12 September 2026

Accepted: 18 September 2026

Published: 27 October 2026

**Abstract** - As learning analytics becomes more widely used in education, there is a need to develop predictive models that account for a range of real-world factors. However, existing predictive systems emphasize model performance more than the design of transparent, trustworthy models. This study combines model performance, fairness, and interpretability in a single pipeline. The integrated predictive framework evaluates five classification models under stratified cross-validation on the Open University Learning Analytics Dataset (OULAD). Two data files are selected from the dataset for model development. The processed data comprises 32,593 merged records with demographic and behavioral features of the learners. SVM-RBF and gradient boosting classifiers show similar, high predictive performance in mean Accuracy, AUC, and F1-score. Fairness is assessed using gender as the sensitive attribute. Random forest shows the lowest group disparities, with a mean Demographic Parity Difference (DPD) of 0.1147 and an Equalized Odds Difference (EOD) of 0.1610. Statistical analysis further identifies significant classifier differences in all metrics except DPD; however, pairwise tests show no significant difference for any metric. SHAP provides feature importance plots for model interpretation based on the sensitive attribute. They show that the behavioral features contribute more to predictions than demographics and primarily drive model decisions. Finally, the study offers practical guidance for developing transparent and fair educational AI systems.

**Keywords** - Educational analytics, Predictive modeling, Fairness in machine learning, Model interpretability, SHAP, Cross-validation.

## 1. Introduction

Learning Analytics (LA) monitors student progress, identifies learning difficulties, and supports at-risk students. It also supports personalized learning and helps teachers improve learning activities through diagnostic and predictive tools. Institutions invest in LA resources to develop technologies according to their educational requirements. Learning Analytics has shown potential to improve student performance, success, and retention [1]. The increasing use of predictive models in education also raises fairness concerns. A study highlighted the existing biases at multiple stages from data measurement to model learning [2].

Another concern is the limited transparency of complex models, which makes their decisions difficult to understand. Explainable AI (XAI) has addressed these challenges and supports human-AI interaction. Different research communities view the goals and applications of XAI differently, but agree on its need [3]. In education, explainability also depends on the needs of different stakeholders, including students, teachers, administrators, and developers. It requires both general AI capabilities and context-specific information. Thus, this presents both

opportunities and challenges in addressing stakeholders' diverse needs. Common XAI approaches include global and local explanations and feature importance [4]. Therefore, fairness and interpretability are important considerations when predictive models support educational decisions.

Fair classroom assessment has become an increasingly important concern for educators. The systematic review identified seven key themes of fairness and integrated them into a coherent conceptual model. It viewed fairness as a continuous process across the different stages of data processing. The study reported that propagating side effects from one stage to the next violates fairness. It supports designing and evaluating assessment practices that promote fairness [5]. XAI further strengthens the focus on fairness, accountability, transparency, and ethics. Existing descriptive and quantitative studies show a strong inclination toward using XAI. At the same time, the focus needs to shift to ensure accountability and transparency in educational AI systems. XAI has been widely applied in different experimental settings to detect and reduce bias. Reinforcement plays a role in promoting responsible and ethical use of AI in education [6]. Beyond fairness and transparency, AI prediction models



support collaborative learning environments. The prediction models are combined with learning analytics to form an integrated approach. Its broader impact includes improved learner engagement and satisfaction [7].

A recent systematic review reports machine learning as a core component in Learning Analytics. Machine Learning and Deep Learning models are among the most used for predictions. The confusion matrix is the commonly used model performance evaluation metric [8]. As prediction models in Learning Analytics grow, recent research shows a shift toward incorporating XAI. Explainability approaches were examined in LMS and EDM contexts to support both global and local explanations. Thus, different levels of explanation improve model understanding by addressing separate stakeholder needs [9]. The significance of prediction and explanation was reported with valuable insights for both teachers and students. However, most systems prioritized prediction accuracy while giving limited attention to explainability [10].

Adoption of advanced AI in Learning Analytics remains slow. Many educational systems still focus on improving model performance rather than its transparency. Moreover, studies rarely examined bias related to sensitive attributes and addressed performance and fairness separately. Neither was integrated into an analytics system with model explainability. The reviewed studies show limited depth in analyzing the influence of sensitive attributes. The explanations generated by the developed models were not always understandable. However, current educational applications require models to be accurate, fair, and transparent together. Educators and decision-makers need clear and trustworthy evidence. Thus, an integrated framework is developed to combine prediction, fairness, and explainability to support reliable, transparent decision-making. The main contributions of this study are as follows:

- It develops a predictive framework that evaluates model performance, fairness, and interpretability simultaneously.
- SHAP examines how behavioral and demographic features influence predictions across sensitive groups.
- It analyzes fairness-performance trade-offs and provides statistical evidence across multiple learning models.
- It assesses potential group-based biases with explicit evaluation with respect to a sensitive attribute.
- It establishes a reliable, reproducible evaluation framework using five-fold cross-validation.

## 2. Related Work

The literature shows a growing shift from AI to XAI in educational predictive analytics. Badal and Sungkur [11] predicted students' final grades and engagement levels using profile data and interactions in the online learning platform.

However, the approach was limited in explainability and also did not explicitly address potential bias in the predictive models. Guleria and Sood [12] modeled a career counseling framework including XAI. It analyzed academic and employability factors and supported decision-making for both white-box and black-box models. However, the limited dataset may affect the model's generalizability. Nnadi et al. [13] used XAI approaches for multi-facet interpretability to predict students' adaptivity levels. Several factors influencing the adaptivity levels were identified, related to dynamic socio-economic and institutional environments.

Model interpretability is not the only concern; algorithmic bias is also an important issue in educational AI systems. Recent studies have further emphasized transparency in AI-supported education. Chai et al. [14] examined the risk of unfairness in an AI-based evaluation system. They compared students' and teachers' perceptions of AI algorithms as evaluators. The findings showed that transparency and explanations made the system appear fairer. However, the study still had limited attention to ethical and privacy issues. Further, Sailer et al. [15] integrated multiple aspects of learning analytics within a learner-centric environment. They combined multimodal data acquisition with theoretical perspectives to gain multidisciplinary insights. Thus, they highlighted a balance between performance and interpretability. Liu et al. [16] found that including XAI increased trust in AI-based educational decision-making. Each educator's unique role was included while observing the trade-off between accuracy and interpretability. Cukurova [17] presented a framework for AIED concepts by conceptualizing AI in education. The hybrid intelligence system was designed to promote a fair, accountable, and transparent predictive model. It was supported by human and artificial intelligence to uphold diversity and human ethical values.

To integrate XAI, SHAP methods were introduced in educational analytics using OULAD. Oyedotun et al. [18] analyzed feature relevance, which provided transparency and fairness. The results showed improved predictive accuracy compared with earlier studies. Further, Gunasekara and Saarela [19] highlighted the importance of quantitatively assessing XAI methods. A limited set of machine learning models was evaluated using a limited set of metrics. The naturally interpretable models struggled to capture complex relationships. In contrast, high-performing models required interpretability. XAI methods were presented as a suitable solution in these cases.

Recent research has also increasingly integrated fairness into predictive educational systems. Raftopoulos et al. [20] compared fairness-supporting algorithms at different stages of model development. However, only the trade-off between performance and fairness was reported without addressing transparency. Kesgin et al. [21] used fairness-aware modeling focused on sensitive attributes. The results revealed disparities

and a contrast between performance and fairness. The study stressed the importance of bias mitigation in fairness-aware machine learning. Gunasekara and Saarela [22] also combined fairness and performance measurement in the machine learning pipeline. Several performance and fairness metrics were used along with explainability. However, the focus was only on students' demographic features, and performance and fairness were evaluated separately without considering their interaction.

The literature shows that learning analytics has limited integration with advanced AI technologies. The educational domain suffers from a lack of transparency and trust. The predictive modeling largely emphasizes performance over other aspects. However, XAI research can help fill this gap by providing explanations and visualizations to support decision-making. As a result, developing an integrated framework that includes all important aspects has greater potential and scope.

### 3. Materials and Methods

This section describes the dataset and the phases of the proposed integrated framework. It comprises data acquisition, preprocessing, model training and evaluation, statistical significance testing, and interpretability analysis. Figure 1 shows specific tasks for each phase.

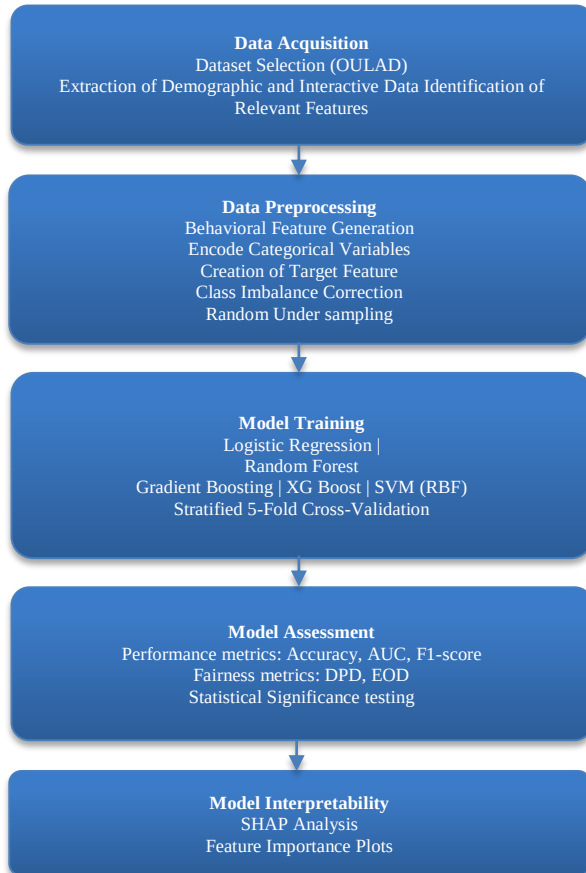


Fig. 1 The main phases of the proposed framework

#### 3.1. Data Acquisition

The study uses the Open University Learning Analytics Dataset (OULAD) [23], from which two files-studentInfo and studentVle- are selected for analysis. The studentInfo file provides demographic and academic information about students. The studentVle file records students' interactions with online learning materials through the sum\_click attribute. The final dataset contains 32,593 instances after merging the two files using common identifiers. Table 1 shows the six selected features used in modeling, three from each data file. The last three features, including total\_clicks, avg\_clicks, and interaction\_count, are derived from the sum\_click attribute. These behavioral features represent student engagement. The combined dataset was then used for preprocessing and model training.

Table 1. Data features used for predictive modeling

| Source File | Feature Name      | Description                               |
|-------------|-------------------|-------------------------------------------|
| studentInfo | Region            | Student's geographical region             |
|             | highest_education | Highest qualification level               |
|             | age_band          | Age category of the student               |
| studentVle  | total_clicks      | Total number of clicks (sum of sum_click) |
|             | avg_clicks        | Average clicks per interaction            |
|             | interaction_count | Number of interaction records             |

#### 3.2. Data Pre-Processing

Preprocessing activities prepare the dataset for the next stage of model development and later performance assessment. These include encoding categorical features, engineering behavioral features, running consistency checks, and target encoding. Categorical features, including region, highest\_education, age\_band, and gender, are numerically encoded using label encoding. Although gender is not part of the predictive features, it is retained separately as the sensitive attribute for group-based fairness and interpretability analyses. The derived behavioral features are aggregated and verified. The final merged dataset is checked for missing and inconsistent values. The target is encoded from the attribute named final\_result. For target encoding, a binary label was assigned as follows: 1 if the result is "Pass" or "Distinction", otherwise 0.

#### 3.3. Class Balancing

The dataset has a class imbalance in the target variable. This imbalance leads to model bias towards the majority class. It is handled before model training with a random undersampling technique. The number of instances for each class of the target variable is computed to identify the smallest class. Now, an equal number of samples is randomly selected

from each target class based on this minimum value. After undersampling, the dataset contains 29,228 instances. The final dataset is balanced and has uniform class distribution. Also, a fixed random state was used to ensure reproducibility. Undersampling is preferred to oversampling because it avoids the insertion of artificially generated samples. This balanced dataset is subsequently used in model training and evaluation.

### 3.4. Model Training and Configuration

Five classifiers evaluated for the educational predictive system are: Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), XGBoost (XGB), and Support Vector Machine with a Radial Basis function kernel (SVM-RBF). LR, RF, GB, and SVM-RBF classifiers are implemented using scikit-learn, whereas XGB is implemented using the XGBoost library. LR runs for 1000 iterations, and RF consists of 200 decision trees. For GB and XGB, the number of estimators is set to 200 with a learning rate of 0.05. In addition, XGB uses a maximum depth of 4 and a subsampling ratio of 0.8. SVM-RBF uses  $C = 1.0$  and  $\gamma = \text{"scale"}$  with probability estimation enabled. A fixed random state ensures reproducibility. Feature scaling using StandardScaler is applied for LR and SVM-RBF.

### 3.5. Cross-Validation, Performance, Fairness and Statistical Evaluation

Performance and fairness are evaluated using 5-fold stratified cross-validation. Performance metrics include Accuracy, Area Under the ROC Curve (AUC), and F1-score. Fairness metrics are: Demographic Parity Difference (DPD) and Equalized Odds Difference (EOD). DPD measures differences in positive prediction rates between gender groups, whereas EOD measures differences in true-positive and false-positive rates between these groups.

Two statistical tests assess classifier significance. First, the non-parametric Friedman test is applied separately to the fold-level values of each evaluation metric, using a significance threshold of 0.05. If a metric shows a significant Friedman result, then pairwise post-hoc comparisons are subsequently performed using the two-sided Wilcoxon signed-rank test. Holm adjustment is additionally applied to the resulting p-values of the paired tests to control the family-wise error rate.

### 3.6. Model Interpretability Analysis

Model interpretation is conducted using SHAP (Shapley Additive Explanations) for each classifier to visualize the feature contributions. A separate experimental setup is configured to reduce the computational intensity of the analysis. This analysis uses the trained LR, RF, GB, XGB, and SVM-RBF classifiers. The tree-based classifiers are trained with fewer estimators. A random subset of up to 100 test instances is selected for each classifier to reduce computational complexity. Different SHAP explainers are applied based on the model type. LinearExplainer is used for

LR, KernelExplainer for the SVM-RBF using clustered background samples from the training data, and TreeExplainer with approximation enabled is used for tree-based models. Finally, SHAP summary plots are generated for each explainer, showing the SHAP values.

## 4. Results and Discussion

The comparative evaluation of the classifiers is presented in terms of performance, fairness, and interpretability. It is divided into five subsections: comparison of performance and fairness metrics, fairness–performance trade-off analysis, statistical significance analysis, comparison with representative state-of-the-art educational predictive modeling studies, and interpretability analysis. First, cross-validated mean and standard deviation values for all metrics are examined, followed by fairness–performance trade-off analysis. Statistical testing then determines whether the differences among classifiers extend beyond fold-to-fold variation. The present study is then compared with reviewed studies from the literature, followed by SHAP interpretation.

### 4.1. Comparison of Model Metrics

Tables 2 and 3 report classifiers' performance and fairness metrics with two values: (1) the mean across folds, and (2) the Standard Deviation (std. dev.). The mean indicates the average across folds, while the standard deviation indicates the extent of fold-to-fold variation. As shown in Table 2, SVM-RBF, XGB, and GB achieved the highest Accuracy, close to each other, while RF had the lowest (0.7463). GB performed best in AUC (0.8452), approximately equal to XGB and LR, whereas RF showed the lowest (0.8270). XGB achieved the highest F1-score (0.7781), which is close to GB and SVM-RBF; again, LR showed the lowest (0.7410). All classifiers showed low standard deviations, indicating less variation in their performance across the five folds.

**Table 2. Cross-validated mean and standard deviation of predictive performance metrics (Accuracy, AUC, F1 score) for the classifiers**

| Model   | Accuracy<br>(mean $\pm$ std.<br>dev.) | AUC<br>(mean $\pm$<br>std. dev.) | F1-score<br>(mean $\pm$ std.<br>dev.) |
|---------|---------------------------------------|----------------------------------|---------------------------------------|
| LR      | 0.7535 $\pm$ 0.0320                   | 0.8426 $\pm$<br>0.0420           | 0.7410 $\pm$<br>0.0230                |
| RF      | 0.7463 $\pm$ 0.0342                   | 0.8270 $\pm$<br>0.0416           | 0.7567 $\pm$<br>0.0233                |
| GB      | 0.7640 $\pm$ 0.0427                   | 0.8452 $\pm$<br>0.0475           | 0.7766 $\pm$<br>0.0316                |
| XGB     | 0.7649 $\pm$ 0.0421                   | 0.8450 $\pm$<br>0.0475           | 0.7781 $\pm$<br>0.0317                |
| SVM-RBF | 0.7662 $\pm$ 0.0395                   | 0.8355 $\pm$<br>0.0493           | 0.7762 $\pm$<br>0.0297                |

Table 3 shows that the classifier with the lowest mean value of DPD and EOD has the highest fairness. RF has the lowest mean DPD (0.1147), followed closely by SVM-RBF (0.1181). The remaining classifiers also show very similar

DPD values, ranging from 0.1190 to 0.1200. For EOD, RF again performs best with the lowest mean value (0.1610). This indicates that RF has comparatively lower group disparity for the data. The highest disparity is observed in LR, with the highest EOD (0.2306), as it can only handle linear decision boundaries. The standard deviations for DPD are relatively large compared to their mean values for all classifiers, indicating substantial fold-to-fold variability.

Therefore, the minimum and maximum DPD values are also reported to show the extent of variation across the cross-validation partitions. However, EOD exhibits relatively smaller standard deviations. Thus, its mean and standard deviation are sufficient to describe its variation. The DPD ranges further indicate that demographic parity is sensitive to the composition of the cross-validation partitions.

**Table 3. Cross-validated mean and standard deviation of fairness metrics (Demographic parity and equalized odds) for the classifiers**

| Model   | DPD (mean $\pm$ std. dev.) | DPD min | DPD max | EOD (mean $\pm$ std. dev.) |
|---------|----------------------------|---------|---------|----------------------------|
| LR      | 0.1190 $\pm$ 0.1454        | 0.0155  | 0.4027  | 0.2306 $\pm$ 0.0238        |
| RF      | 0.1147 $\pm$ 0.1320        | 0.0046  | 0.3716  | 0.1610 $\pm$ 0.0109        |
| GB      | 0.1200 $\pm$ 0.1446        | 0.0023  | 0.4009  | 0.1776 $\pm$ 0.0140        |
| XGB     | 0.1199 $\pm$ 0.1427        | 0.0038  | 0.3973  | 0.1784 $\pm$ 0.0125        |
| SVM-RBF | 0.1181 $\pm$ 0.1451        | 0.0065  | 0.4024  | 0.1856 $\pm$ 0.0211        |

Fairness results are specific to gender, the sensitive attribute, used to compute DPD and EOD. Therefore, the observed disparities should not be generalized to other demographic attributes or intersectional subgroups. Different fairness patterns may emerge when attributes such as age group or their intersection are considered.

#### 4.2. Fairness-Performance Trade-Off Analysis

The fairness-performance trade-off is analyzed using F1-score and absolute EOD; both have the lowest variation across data folds. Figure 2 illustrates the trade-off between these two metrics based on their cross-validated mean values. From the

plot, all classifiers are shown distributed within a moderate range of F1-scores (0.74-0.78) and absolute EOD (0.16-0.23). RF has the lowest EOD while maintaining a competitive F1-score, thus showing the strongest fairness. The other two tree classifiers, XGB and GB, achieve the highest F1-scores, but exhibit moderately higher fairness gaps. Thus, they achieve better predictive performance, but with increased disparity. SVM-RBF shows good performance with a slightly higher fairness gap than RF. LR shows the worst trade-off, with the lowest F1-score and the highest fairness gap. Thus, the increase in predictive performance is accompanied by some increase in group disparity. However, this pattern is not observed consistently across all classifiers.

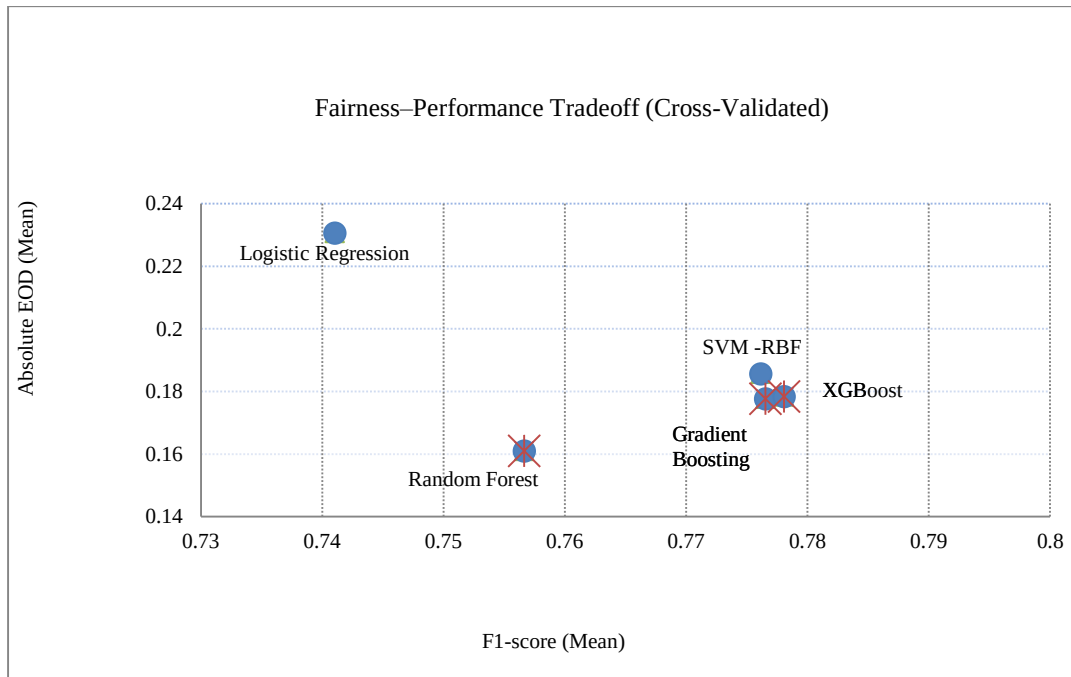


**Fig. 2 Cross-validated fairness-performance trade-off showing the relationship between mean F1 score and mean absolute equalized odds difference for the evaluated classification models**

#### 4.3. Statistical Significance Analysis

Friedman test results in Table 4 show significant differences among the evaluated classifiers for all the performance metrics. Accuracy has a test statistic of 13.4545 with  $p=0.0093$ , while the AUC statistic is 14.2400 with  $p=0.0066$ .

The largest difference is observed for F1-score, with a Friedman statistic of 17.1200 and  $p=0.0018$ . Since all three p-values are below the significance threshold, the classifiers significantly influence predictive performance.

In contrast, the fairness metrics show a different pattern, as DPD achieves the lowest Friedman statistic (3.2000) with the highest p-value (0.5249). Therefore, although the mean DPD values (Table 3) vary numerically, the p-value was above the significance threshold.

Therefore, there is insufficient evidence that the classifiers differ significantly in DPD disparity. However, EOD shows a significant classifier effect, with the test statistic of 16.1600 and  $p=0.0028$ . Thus, the fold-to-fold variation alone does not explain the variation in EOD across classifiers.

**Table 4. Friedman test results for cross-validated performance and fairness metrics**

| Metric   | Friedman Statistic | p-value | Significant |
|----------|--------------------|---------|-------------|
| Accuracy | 13.4545            | 0.0093  | Yes         |
| AUC      | 14.2400            | 0.0066  | Yes         |
| F1-score | 17.1200            | 0.0018  | Yes         |
| DPD      | 3.2000             | 0.5249  | No          |
| EOD      | 16.1600            | 0.0028  | Yes         |

Despite the significant Friedman test, none of the individual pairwise tests reach the significance threshold for any metric. Table 5 presents the post-hoc Wilcoxon signed-rank test results with Holm correction for Accuracy, AUC, F1-score, and EOD. For Accuracy, several comparisons have the lowest raw p-value of 0.0625, including LR with RF and SVM-RBF, and RF with GB, XGB, and SVM-RBF. Similar results are obtained for AUC, F1-score, and EOD, where several comparisons also have a raw p-value of 0.0625. In some F1-score and EOD comparisons,  $W=0$ , showing that the paired fold-level differences are in the same direction. However, the corresponding raw p-value remains above the 0.05 significance level, and the Holm-adjusted p-value is 0.6250. Therefore, none of the classifier pairs differ significantly on any evaluated metric.

**Table 5. Post-hoc Wilcoxon signed-rank test results with Holm correction**

| Metric   | Classifier 1 | Classifier 2 | W | Raw p-value | Holm-adjusted p-value |
|----------|--------------|--------------|---|-------------|-----------------------|
| Accuracy | LR           | RF           | 0 | 0.0625      | 0.6250                |
|          |              | GB           | 2 | 0.1875      | 0.7500                |
|          |              | XGB          | 1 | 0.1250      | 0.6250                |
|          |              | SVM-RBF      | 0 | 0.0625      | 0.6250                |
|          | RF           | GB           | 0 | 0.0625      | 0.6250                |
|          |              | XGB          | 0 | 0.0625      | 0.6250                |
|          |              | SVM-RBF      | 0 | 0.0625      | 0.6250                |
|          | GB           | XGB          | 3 | 0.6250      | 1.0000                |
|          |              | SVM-RBF      | 6 | 0.8125      | 1.0000                |
|          | XGB          | SVM-RBF      | 7 | 1.0000      | 1.0000                |
| AUC      | LR           | RF           | 1 | 0.1250      | 0.7500                |
|          |              | GB           | 5 | 0.6250      | 1.0000                |
|          |              | XGB          | 5 | 0.6250      | 1.0000                |
|          |              | SVM-RBF      | 4 | 0.4375      | 1.0000                |
|          | RF           | GB           | 0 | 0.0625      | 0.6250                |
|          |              | XGB          | 0 | 0.0625      | 0.6250                |
|          |              | SVM-RBF      | 2 | 0.1875      | 0.9375                |
|          | GB           | XGB          | 4 | 0.4375      | 1.0000                |
|          |              | SVM-RBF      | 0 | 0.0625      | 0.6250                |
|          | XGB          | SVM-RBF      | 0 | 0.0625      | 0.6250                |
| F1-score | LR           | RF           | 0 | 0.0625      | 0.6250                |
|          |              | GB           | 0 | 0.0625      | 0.6250                |
|          |              | XGB          | 0 | 0.0625      | 0.6250                |
|          |              | SVM-RBF      | 0 | 0.0625      | 0.6250                |
|          | RF           | GB           | 0 | 0.0625      | 0.6250                |
|          |              | XGB          | 0 | 0.0625      | 0.6250                |
|          |              | SVM-RBF      | 0 | 0.0625      | 0.6250                |
|          | GB           | XGB          | 3 | 0.3125      | 0.9375                |

|     |     |         |   |        |        |
|-----|-----|---------|---|--------|--------|
| EOD | XGB | SVM-RBF | 5 | 0.6250 | 0.9375 |
|     |     | SVM-RBF | 4 | 0.4375 | 0.9375 |
|     | LR  | RF      | 0 | 0.0625 | 0.6250 |
|     |     | GB      | 0 | 0.0625 | 0.6250 |
|     |     | XGB     | 0 | 0.0625 | 0.6250 |
|     |     | SVM-RBF | 0 | 0.0625 | 0.6250 |
|     | RF  | GB      | 0 | 0.0625 | 0.6250 |
|     |     | XGB     | 0 | 0.0625 | 0.6250 |
|     |     | SVM-RBF | 0 | 0.0625 | 0.6250 |
|     | GB  | XGB     | 3 | 0.3125 | 0.9375 |
|     |     | SVM-RBF | 3 | 0.3125 | 0.9375 |
|     | XGB | SVM-RBF | 3 | 0.3125 | 0.9375 |

The Friedman test showed statistical differences among the classifiers for Accuracy, AUC, F1-score, and EOD. However, the Wilcoxon test found no significant differences between individual classifier pairs. This may be due to the limited number of observations from five-fold cross-validation and the multiple-comparison correction. Therefore, although the classifiers showed different numerical performance, no individual classifier was statistically superior to another.

#### 4.4. Comparison with Representative State-of-the-Art Educational Predictive Modeling Studies

Recent educational predictive modeling has moved beyond performance-oriented toward fair and explainable predictions. Table 6 compares the present study with representative recent approaches. However, these studies differ in datasets, prediction tasks, features, pre-processing, and evaluation protocols. The predictive values should not be interpreted as direct performance rankings.

Earlier studies primarily emphasized predictive performance, with subsequent work increasingly incorporating explainability into educational predictive modeling [11-13]. Although these studies improve the interpretability of educational predictions, they do not explicitly evaluate group-based fairness. Later studies are methodologically closer to the present work because they explicitly consider fairness, while also integrating explainability [20-22]. The study using OULAD [22] provides the closest dataset-level comparison. The present study extends this line of evaluation by jointly examining predictive performance, group-based fairness, statistical evidence, and model interpretability. The observed predictive results may be associated with the combined use of demographic and aggregated behavioral engagement features. In addition, model-specific SHAP analysis is stratified by gender to examine whether predictor-feature contributions vary across the sensitive groups. The comparison indicates that the present study's contribution is an integrated evaluation supported by statistical tests.

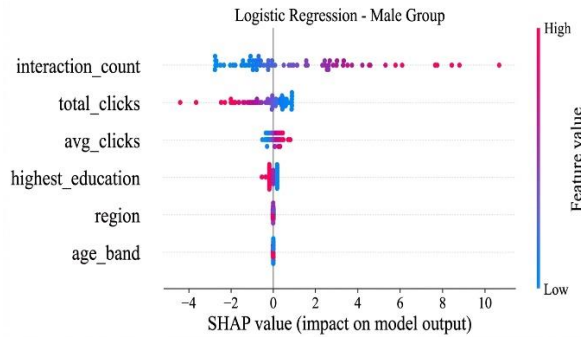
**Table 6. Comparative analysis of the present study and representative existing educational predictive modeling studies across predictive performance, fairness, and explainability**

| Study                   | Primary Focus                            | Predictive Result                      | Fairness | XAI                           | Key Distinction                                           |
|-------------------------|------------------------------------------|----------------------------------------|----------|-------------------------------|-----------------------------------------------------------|
| Badal and Sungkur [11]  | Student grade/engagement prediction      | RF Acc. $\approx$ 0.85                 | No       | Limited                       | Primarily performance-oriented                            |
| Guleria and Sood [12]   | ML-based career counseling               | NB Recall = 0.912; F1 = 0.907          | No       | Yes                           | Combines educational prediction with explainability       |
| Nnadi et al. [13]       | Student adaptability prediction          | RF Acc. = 0.91                         | No       | SHAP and multiple XAI methods | Strong interpretability; fairness not evaluated           |
| Raftopoulos et al. [20] | Educational bias mitigation              | Acc. and fairness vary with mitigation | Yes      | No                            | Compares fairness-mitigation strategies                   |
| Kesgin et al. [21]      | Performance, fairness and explainability | XGB Acc. = 0.789; F1 = 0.803           | DPD, EOD | SHAP, LIME                    | Integrated evaluation using multiple sensitive attributes |

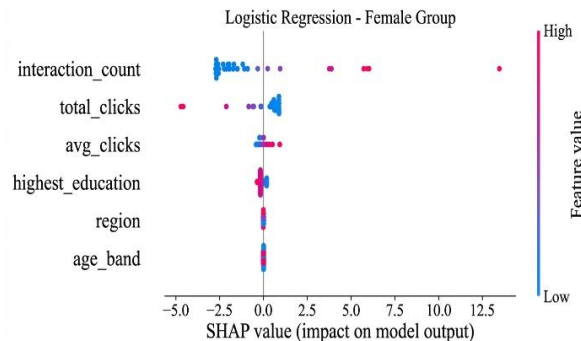
|                             |                                                         |                                          |                                           |                        |                                                                                                           |
|-----------------------------|---------------------------------------------------------|------------------------------------------|-------------------------------------------|------------------------|-----------------------------------------------------------------------------------------------------------|
| Gunasekara and Saarela [22] | Fairness–performance trade-off on OULAD                 | Accuracy, F1, AUC                        | Yes                                       | Not central            | OULAD fairness mitigation and statistical analysis                                                        |
| Current study               | Integrated performance–fairness–XAI evaluation on OULAD | Acc. = 0.7662; AUC = 0.8452; F1 = 0.7781 | DPD, EOD + stability/statistical analysis | Gender-stratified SHAP | Behavioral modeling, Joint diagnostic evaluation of performance, fairness, stability and interpretability |

#### 4.5. Interpretability Analysis

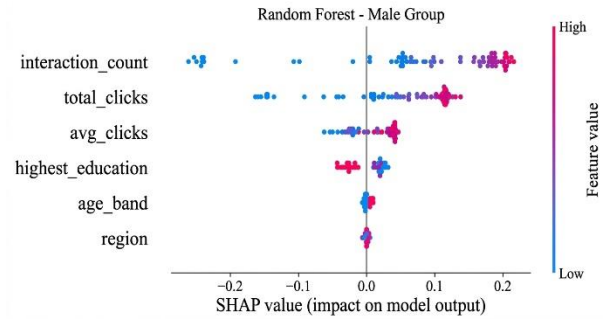
Figure 3(a)-(j) shows the SHAP summary scatter plots of each classifier for the sensitive attribute gender (female and male). Each feature label is shown along with its SHAP values, showing each feature's contribution across all classifiers. The comparison is based on male and female groups among the behavioral and demographic features. In all plots, `interaction_count` shows the strongest influence across all classifiers for both gender groups. However, its SHAP values are more spread out for Logistic Regression, but comparatively narrower for Random Forest and SVM-RBF. The second most influential feature is `total_clicks`, with a similarly varied spread across classifiers. Apart from these two behavioral features, `avg_clicks` is the next most influential feature in Logistic Regression, Random Forest, and Gradient Boosting. Among the demographic features, `highest_education` contributes more than `region` and `age_band`. It is more influential than `avg_clicks` in most classifiers. However, its influence remains lower than that of the dominant behavioral features.



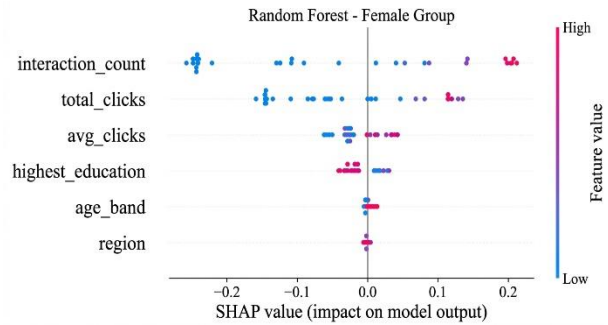
(a) Male group (Logistic Regression)



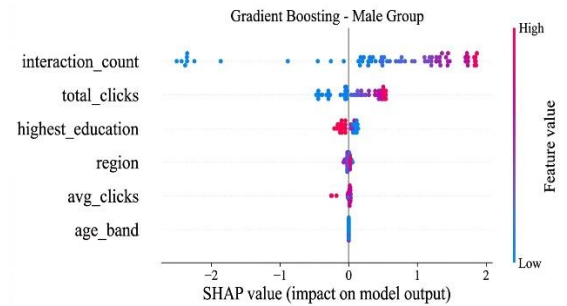
(b) Female group (Logistic Regression)



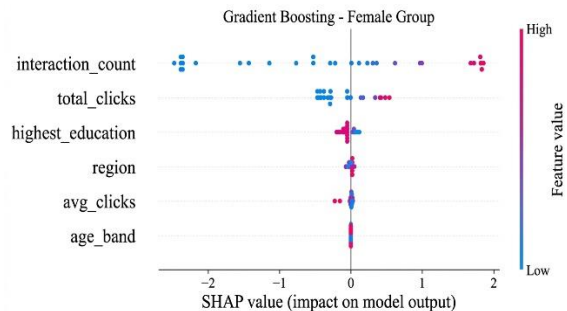
(c) Male group (Random Forest)



(d) Female group (Random Forest)



(e) Male group (Gradient Boosting)



(f) Female group (Gradient Boosting)

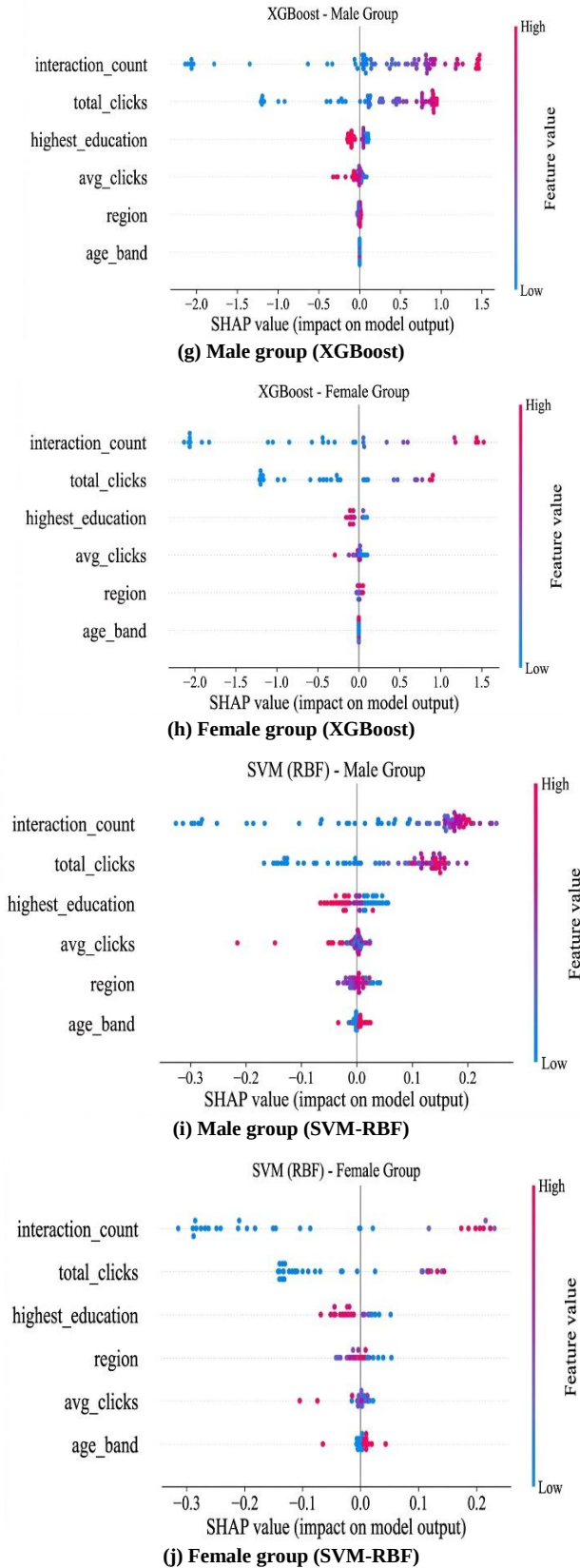


**Fig. 3** SHAP summary scatter plots ((a)-(j)) for the classifiers showing feature contributions to the model output, illustrating feature impact and variability across the sensitive attribute.

The analysis shows that behavioral engagement features contribute more than demographic features. The gender-stratified plots show similar feature rankings for most classifiers across the two gender groups. This suggests that the classifiers rely on similar predictors for male and female students. However, the strength of some feature contributions varies between groups. This group-wise analysis complements the fairness metrics by showing how the models use predictor features across sensitive groups.

## 5. Conclusion

Educational predictive models need to be assessed not only for their predictive performance but also for fairness and interpretability. This study examined these three aspects together using OULAD to develop a fair, interpretable predictive framework. It evaluated five classifiers using five-fold stratified cross-validation. The margin-based (SVM-RBF) and Gradient Boosting (GB and XGB) classifiers showed the highest predictive performance. In contrast, RF showed lower group disparity with the lowest mean DPD and EOD. Beyond the mean and standard deviations, the Friedman test identified significant differences among classifiers for all metrics, except DPD. The Holm-corrected Wilcoxon tests found no significant pairwise differences among the classifiers. Thus, no classifier is statistically superior. In addition, the variation in DPD across folds highlights the importance of fairness stability. SHAP analysis reported that behavioral features had more influence on predictions than demographic features. It also provided insight into each feature's contribution across gender groups.

Despite various modeling aspects, the framework is limited by a single sensitive attribute for fairness assessment. Thus, the reported findings should not be generalized to demographic groups other than gender. The framework has also not been validated across different educational contexts, which again limits the generalizability of the findings. In addition, the focus was on fairness assessment rather than bias-mitigation strategies. In future, the framework can be extended to consider multiple sensitive attributes and intersectional groups (gender  $\times$  age) across different educational contexts. Bias-mitigation strategies can also be investigated when persistent disparities are identified. Finally, the study highlights the importance of integrating multiple modeling aspects when developing accountable and responsible educational predictive systems.

## Conflicts of Interest

The authors declare that they have no conflict of interest.

## Funding Statement

This research received no external funding.

## Acknowledgments

Not applicable.

## References

- [1] Marcela Hernández-de-Menéndez et al., “Learning Analytics: State of the Art,” *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 16, no. 3, pp. 1209-1230, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] René F. Kizilcec, and Hansol Lee, *Algorithmic Fairness in Education*, 1<sup>st</sup> ed., The Ethics of Artificial Intelligence in Education, Routledge, pp. 174-202, 2022. [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Krzysztof Fiok et al., “Explainable Artificial Intelligence for Education and Training,” *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, vol. 19, no. 2, pp. 133-144, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] Hassan Khosravi et al., “Explainable Artificial Intelligence in Education,” *Computers and Education: Artificial Intelligence*, vol. 3, pp. 1-22, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Ali Baniyadi et al., “Fairness in Classroom Assessment: A Systematic Review,” *The Asia-Pacific Education Researcher*, vol. 32, no. 1, pp. 91-109, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Bahar Memarian, and Tenzin Doleck, “Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and Higher Education: A Systematic Review,” *Computers and Education: Artificial Intelligence*, vol. 5, pp. 1-12, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Fan Ouyang et al., “Integration of Artificial Intelligence Performance Prediction and Learning Analytics to Improve Student Learning in Online Engineering Course,” *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, pp. 1-23, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] Nabila Sghir, Amina Adadi, and Mohammed Lahmer, “Recent Advances in Predictive Learning Analytics: A Decade Systematic Review (2012-2022),” *Education and Information Technologies*, vol. 28, no. 7, pp. 8299-8333, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Ashwin Rachha, and Mohammed Seyam, “Explainable AI in Education: Current Trends, Challenges, and Opportunities,” *SoutheastCon 2023*, Orlando, FL, USA, pp. 232-239, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Gomathy Ramaswami et al., “Use of Predictive Analytics within Learning Analytics Dashboards: A Review of Case Studies,” *Technology, Knowledge and Learning*, vol. 28, no. 3, pp. 959-980, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Yudish Teshal Badal, and Roopesh Kevin Sungkur, “Predictive Modelling and Analytics of Students’ Grades using Machine Learning Algorithms,” *Education and Information Technologies*, vol. 28, no. 3, pp. 3027-3057, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Pratiyush Guleria, and Manu Sood, “Explainable AI and Machine Learning: Performance Evaluation and Explainability of Classifiers on Educational data Mining Inspired Career Counseling,” *Education and Information Technologies*, vol. 28, no. 1, pp. 1081-1116, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Leonard Chukwualuka Nnadi et al., “Prediction of Students’ Adaptability using Explainable AI in Educational Machine Learning Models,” *Applied Sciences*, vol. 14, no. 12, pp. 1-26, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Fangyuan Chai et al., “Grading by AI Makes me Feel Fairer? How Different Evaluators Affect College Students’ Perception of Fairness,” *Frontiers in Psychology*, vol. 15, pp. 1-18, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Michael Sailer et al., “The End is the Beginning is the End: The Closed-Loop Learning Analytics Framework,” *Computers in Human Behavior*, vol. 158, pp. 1-19, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [16] Qianhui Liu, Juan D. Pinto, and Luc Paquette, *Applications of Explainable AI (XAI) in Education*, Trust and Inclusion in AI-Mediated Education, Springer, pp. 93-109, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Mutlu Cukurova, “The Interplay of Learning, Analytics and Artificial Intelligence in Education: A Vision for Hybrid Intelligence,” *British Journal of Educational Technology*, vol. 56, no. 2, pp. 469-488, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Samuel Abiodun Oyedotun, Otega Prosper Ejenarhome, and Godfrey Perfectson Oise, “Learning Analytics and Predictive Modeling: Enhancing Student Success through Data-Driven Insights,” *Journal of Science Research and Reviews*, vol. 2, no. 3, pp. 42-51, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [19] Sachini Gunasekara, and Mirka Saarela, “Explainable AI in Education: Techniques and Qualitative Assessment,” *Applied Sciences*, vol. 15, no. 3, pp. 1-16, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] George Raftopoulos, Gregory Davrazos, and Sotiris Kotsiantis, “Evaluating Fairness Strategies in Educational Data Mining: A Comparative Study of Bias Mitigation Techniques,” *Electronics*, vol. 14, no. 9, pp. 1-24, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Kadir Kesgin et al., “Beyond Performance: Explaining and Ensuring Fairness in Student Academic Performance Prediction with Machine Learning,” *Applied Sciences*, vol. 15, no. 15, pp. 1-25, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] Sachini Gunasekara, and Mirka Saarela, “Trade-offs between Fairness and Performance in Educational AI: Analyzing Post-Processing Bias Mitigation on the OULAD,” *Information and Software Technology*, vol. 189, pp. 1-10, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [23] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal, “Open University Learning Analytics Dataset,” *Scientific Data*, vol. 4, no. 1, pp. 1-8, 2017. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]