Explainable Week-6 Early Warning and Intervention Planning for Student Outcome Prediction

1st Rakshit Jain
Department of Computer Science
Chandigarh University
Mohali, 140413, Punjab, India
jainrakshit209@gmail.com

2nd Meenu Gupta
Department of Computer Science
Chandigarh University
Mohali, 140413, Punjab, India
meenu.e9406@cumail.in

3rd Rakesh Kumar
Department of Computer Science
Chandigarh University
Mohali, 140413, Punjab, India
rakesh77kumar@gmail.com

***Abstract*— The importance of quickly identifying at-risk students for online learning is paramount when providing support through timely interventions. This paper outlines an explainable predictive analytics system for predicting a three-class outcome, Pass, Distinction, and At Risk, for students enrolled in an organisation via the Open University Learning Analytics Dataset based solely on course activity during the first 6 weeks of the course. An educator-actionable definition of At Risk merges Fail and Withdrawn students because they will require similar intervention at approximately the same time. A classifier built using XGBoost with 16 behavioral, assessment and demographic features achieved 67.90% accuracy on the stratified holdout test dataset. The full model has a 67.94% mean accuracy when evaluated through 5-fold cross-validation and has improved accuracy from the 9 feature baseline of 66.13% to the full 16 feature model. SHAP explanations enable per-student interpretability and support the automatic generation of personalized learning recommendations based on feature attributions. The final outcome of this work indicates that timely and targeted interventions by educators can occur through competitive and transparent early predictions being made before the midpoint of the course.**

Keywords— Learning analytics, early warning systems, student outcome prediction, explainable AI, SHAP, XGBoost.

I. INTRODUCTION

Online and distance learning scenarios create huge amounts of interaction data through Virtual Learning Environments (VLEs), offering new possibilities for predictive analytics to support data-driven educational decisions [1]. One of the most important applications is the early detection of students at risk of academic failure, because early interventions can significantly improve students' retention, engagement and learning outcomes [2].

Despite the advances achieved in learning analytics, a major drawback still remains: most predictive models are not genuinely early. In fact, many approaches require data collected during a significant part of the course duration, which limits their real-world applicability, since instructors must act within the first weeks of a course. At this early point in time, only sparse and incomplete evidence of the student's activities and learning behaviour is available, which makes the predictions inherently uncertain. Additionally, a large number of existing models are not easily interpretable, i.e., they do not explain why a given student has been identified as at risk, which limits their usefulness for teachers who need to act on such predictions [3]. Student engagement pattern instability in the early stages of a course is caused by the adaptation to a new learning environment, resulting in highly variable behavioral signals and further decreasing the reliability of prediction [4].

Such implications of designing and implementing learning analytics tools, which are based on unreliable predictions, may be profound indeed. In the absence of reliable early warning information, teachers miss the critical window for effective interventions. Students who become disengaged in the beginning of the course are more prone to drop out or fail the course. However, many of the existing frameworks warn of the challenges occurring later in the course, or the interpretations of system predictions are of little help in supporting teachers for effective actions [5]. This highlights the critical importance of developing systems that can make predictions in an explainable manner [6]. Recent research indicated that trust and the success of the system in real educational settings also depend on the explainability of learning analytics systems [7].

These difficulties emphasize the need for future research into the development of prediction tools that, besides being capable of making reliable predictions based on a limited set of data in the initial stages of data analysis, must also make sense to instructors.

In this regard, this paper offers an early warning system that not only has the ability to interpret its predictions but is also capable of making predictions about student success in the first few weeks of their studies, using three classes (Distinction, Pass, At Risk), in which At Risk refers to students who fail or withdraw from their courses. Specifically, the paper develops an early warning prediction scheme at Week-6 using the Open University Learning Analytics Dataset by comparing the results achieved with the use of two different feature sets. This approach to predictions is coupled with explainability through the combination of gradient-boosted trees and SHAP, along with a rule-based recommendation system, where the effectiveness of both prediction and rules is assessed through cross-validation and significance testing.

Through its potential for early prediction and interpretation, with practical recommendations, this model solves the problem of the lack of predictability and application in the context of education.

The rest of the paper is organized as follows. Chapter II presents the literature review, Chapter III the proposed methodology, Chapter IV the experimental results and analysis, and Chapter V the conclusions and future work.

II. LITERATURE REVIEW

Due to the existence of such well-structured datasets, it has become possible to perform more meaningful evaluations of the predictive models. In particular, Kuzilek et al. [8] presented the OULAD(Open University Learning Analytics

Dataset) dataset with complete information about demographics, assessment results, and VLE activity of 22 courses and over 32,000 students. This dataset has become the benchmark used in all further works as the most reliable tool for reproduction and fair comparison of different algorithms for education predictions due to its rich structure including demographic characteristics and behaviors. It has been shown by Grinsztajn et al. [9] that models based on trees work better than the neural networks in predicting performance from the table-like input of typical educational datasets and thus serve as an argument in favor of using ensembles for solving educational tasks. Guyon and Elisseeff [10] proved that feature selection helps improve the results by decreasing dimensions and redundancy.

In general, the field of educational data mining has witnessed significant evolution over time. Romero and Ventura [11] found evidence of further development of the EDM domain and increased application of machine learning techniques over a span of ten years, showing the transition towards more complex and practical predictive methods. The work of Albreiki et al. [12] represents another step forward in the field of explainable prediction in that the researchers have developed a tool for automatic generation of remedial actions based on explainable ML and rule-based reasoning.

Predictions based on initial data not being enough is a topic that has also been extensively studied in the literature. Hlosta et al. [13] introduced the concept of Ouroboros to demonstrate that useful early warning signals can be generated from incomplete behavioral data without any prior information related to the history of any specific course, a significant step towards robust early prediction tools. Conijn et al. [14], predictions conducted in 17 blended Moodle courses have shown that variations in course structure play a critical role in model generalization.

Privacy and ethics remain critical topics within any Learning Analytics system implementation, and even more so when it comes to prediction, which may have an impact on the learners' success rate. For instance, Drachsler and Greller [15] have developed the DELICATE checklist, which consists of eight stages addressing concerns regarding data ownership, transparency, consent, and responsibility, which become necessary for early warning systems due to the high possibility of false prediction, along with the associated ethical responsibility in the use of learner's data. Despite remarkable advancements achieved by the community in all the aforementioned topics, it becomes clear that modern systems cannot manage to combine the ability of early prediction and the interpretability of such predictions.

III. METHODOLOGY

A. Dataset Description:

This study uses the Open University Learning Analytics Dataset (OULAD) which is a benchmark, publicly available dataset first introduced by Kuzilek et al. [8] and is now available through the Open University's institutional repository. The dataset consists of anonymized data of students, including information from the virtual learning environment interaction logs, evaluations, demographics, among others. The classification problem in a three-class form is adopted to define the prediction task. The three classes are derived from the four outcome variables namely, Distinction, Pass, Fail, and Withdrawn.

- Distinction: Students who get a distinction grade
- Pass: Students who get a pass grade
- At Risk: Students who fail or withdraw

Since the Failure and Withdrawal classes reflect students not completing the course and hence require similar intervention measures, they have been merged into an At-Risk class. In order to ensure that the predictions can be made in time for immediate intervention, all features have been calculated based on data up to Day 42.

B. Proposed System

The design is a Week-6 Early Warning Pipeline based on the following three sources of data from OULAD: student attributes, assessment information, and VLE usage logs. The behavioral and assessment information will be limited to the first 42 days of study to mimic the early intervention scenario. From the aforementioned data sources, two types of table-based feature representations, namely, the 9 feature version and the 16 feature version, which will be elaborated in Section III-C.

To determine which class a student belongs to, either "Pass", "Distinction" or "At Risk", XGBoost classifier is used.. To improve the interpretability of the model, the SHAP values are calculated for the trained full feature model, where feature importance rankings and attribution values for each student are included. Finally, the recommended guidance will be generated using a rule-based recommendation engine, where the main drivers are mapped to the suggested intervention language. educators.

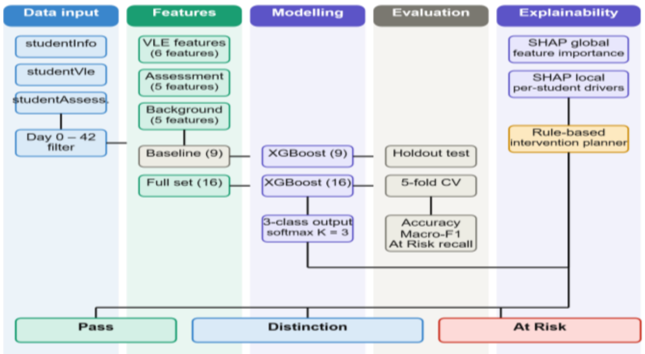

Fig. 1. Proposed Week-6 early warning framework using student data, assessment records, and VLE interactions for prediction, interpretation, and intervention.

C. Week-6 Feature Engineering

All features are based solely on data up until and including Day 42. Aggregations at the student level are performed through concatenation of the Virtual Learning Environment log file and assessments tables into a flat table, augmented with additional background variables from the student information table. Let (i) denote a student and let (t) denote the day index in the VLE logs.

1) VLE engagement features (from studentVle.csv)

Filter the activity log to the early window, and compute the following engagement measures:

- total_clicks: number of all clicks made by a student in the early period.
- days_active: number of unique days with some activity in the VLE.

- resources_accessed: number of distinct resource codes used by a student in the early period.
- avg_daily_clicks: total_clicks divided by (days_active + 1) to compute the average intensity of student interactions. The increment of 1 protects against division by zero when computing avg_daily_clicks for non-active students.
- click_consistency: days_active divided by 42, capturing how consistently the learner is engaging in the early period.

To measure engagement change, the early window is split in half (Day 0 to Day 21 and Day 22 to Day 42), and the following engagement trend metric is defined using $C_{i,1}$ and $C_{i,2}$ - numbers of clicks in each of the two periods:

$$engagement\_trend = \frac{C_{i,2}-C_{i,1}}{C_{i,1}+C_{i,2}+1}, \qquad (1)$$

and clipped to $[-1,1]$for numerical stability. This bounded indicator distinguishes learners whose engagement is growing from those whose engagement is declining.

2) Assessment performance features

Early assessment performance metrics are built from records with submission timestamps <= Day 42:

- avg_score: average of early assessment scores,
- score_std: standard deviation of early assessment scores,
- num_assessments: number of early assessments.

Finally, two additional metrics are defined:

- core_x_attempts = avg_score × num_assessments, which captures both performance and frequency, and
- submission_regularity = num_assessments / 3, an attempt at measuring early submission frequency.

The denominator of 3 represents the average number of assessments that are sponsored in the first six weeks of all OULAD course presentations and normalizes the frequency of participation to a standard scale.

3)Background attributes (from studentInfo.csv)

The selected background attributes are those selected from the student information table. The categorical variables are then ordinally encoded, where age band is coded into age_enc while the highest education is coded into edu_enc. The attributes used include num_of_prev attempts and disability coded into disability_enc.

4) Missing values and final feature sets

Missing engagement metrics due to students who had not participated in the early Virtual Learning Environment are imputed by zero. This signifies the lack of early observations and guarantees that students who show low levels of engagement (and are therefore most likely to need interventions) will be included in the study, rather than excluded.

Two feature sets are considered. The basic one consists of nine features chosen to construct a compact model for comparison purposes. The other one includes all sixteen engineered features and is used in the main model and SHAP analysis. All feature descriptions are presented in Table I.

Table I. Week-6 engineered features used in the baseline (9) and full (16) models.

Feature	**Definition(from implementation)**	**Source**	**Baseline?**
total_clicks	Sum of sum_click for $(t \le 42)$	studentVle	Yes
days_active	Number of unique active days (date) for (t \le 42)	studentVle	Yes
avg_daily_clicks	(total_clicks)/ (days_active + 1)	studentVle	Yes
click_consistency	days_active / 42	studentVle	No
resources_accessed	Numberof unique id_site for $(t \le 42)$	studentVle	Yes
engagement_trend	$\left((C_{late} - C_{early}) / (C_{late} + C_{early} + 1)\right), whereearly = days0 - 21, late = 22 - 42; clippedto([-1,1])$	studentVle	No
avg_score	Mean of early assessment score (restricted to date_submitted ≤ 42 if available)	studentAssessment	Yes
score_std	Std. dev. of early assessment scores	studentAssessment	Yes
num_assessments	Count of early assessment records	studentAssessment	Yes
score_x_attempts	avg_score × num_assessments	studentAssessment	No
submission_regularity	num_assessments / 3 (normalized participation proxy)	studentAssessment	No
forum_posts (proxy)	Proxy feature derived from engagement (e.g., scaled resources_accessed) to approximate interaction level; OULAD does not contain explicit forum activity	Derived (studentVle)	No
age_enc	Ordinal encoding of age_band	studentInfo	Yes
edu_enc	Ordinal encoding of highest_education	studentInfo	No
num_of_prev_attempts	Number of previous attempts	studentInfo	Yes
disability_enc	1 if disability == "Y", else 0	studentInfo	No

Since there are no explicit logs on how often how often a user participates in forums with in dataset, forum_posts derived from the resources_accessed feature using a scaling factor. Although this is not the best measure of actual forum participation, at least one measure of engagement is preserved

D. Classification Model: XGBoost for Three-Class Prediction

Three-class classification is done by using the same Week-6 feature set but utilizing the XGBoost algorithm, which belongs to the category of gradient-boosted trees and can work well on table data with nonlinear relations. The predictions made by this model consist of the labels Pass, Distinction, and At Risk using a multi-class softmax (K=3) objective function. In order to measure the influence of each additional feature, two separate models are constructed according to the same method, namely the baseline model and the full one, containing 9 and 16 features correspondingly. Hyperparameters for each algorithm are selected by random search and kept unchanged further.

E. Model Evaluation

Evaluation is done on both sets through stratified holdout and five-fold cross validation. The stability of the model is analyzed based on the average and standard deviation of accuracies per fold, while a paired t-test is performed to check if the full features set yields a statistical significance compared to the base features set. Holdout set performances are assessed using the accuracy, balanced accuracy, macro F1 score, and the At Risk Recall.

F. Model Explanation

The SHAP method is used to interpret the trained XGBoost model. SHAP explains models on a global scale by ranking feature importance according to their mean absolute SHAP values and on a local level by identifying how much each feature influenced particular students' predictions for the At-Risk class. As a result, it becomes possible to connect students' behaviour such as low participation or poor early assessment performance to the model output

G. Intervention Strategy

In order to turn predictions into actionable insights and guide further action, rule-based recommendations were applied. In case of a low engagement prediction, recommendations include regular studying, higher interaction with material, and consistent studying. For cases of poor assessment performance, recommendations include revision, completing of assessments in time, and asking for help.

IV RESULTS & ANALYSIS

The prediction process for identifying at-risk students at week 6 (42 days) is evaluated using two different feature sets: 1 - The Baseline with 9 features and 2 - The Full Model with 16 features. The predictions are evaluated using both stratified holdout testing (using a single time slice) and five-fold cross-validation to provide the best possible estimate of performance of the predictions on unseen data. The performance is evaluated by accuracy, balanced accuracy, macro-F1 score, and At Risk Recall.

A. Test Performance

The complete model results in a model accuracy of 67.90%, which is an increase of 1.54% compared to the base model, which had an accuracy of 66.35%. There is also an increase in balanced accuracy by 2.55 and an increase in macro-F1 by 0.0226. This indicates that there is improved performance in all classes instead of improvements because of class imbalances. The recall of at-risk students is increased from 0.7693 to 0.7808.

Table II. Baseline vs full model performance on the holdout test set (Week-6).

Model	Features	Accuracy	Balanced Acc.	Macro-F1	At Risk Recall
Majority baseline	–	55.13%	–	–	–
Baseline XGBoost	9	66.35%	54.36%	0.5641	0.7693
Full XGBoost(ours)	16	67.90%	56.91%	0.5867	0.7808

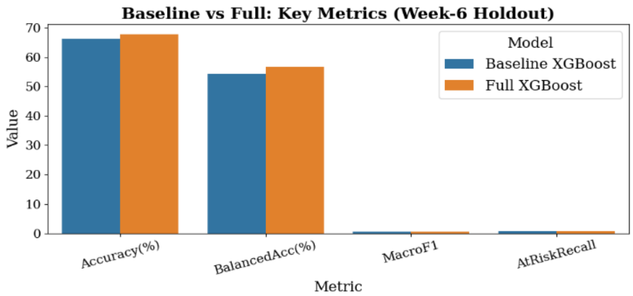

Fig. 2. Holdout metric comparison between baseline and full models.

Table III. Algorithm Comparison at Week 6 (Holdout Test Set)

Model	**Accuracy (%)**	**Balanced Accuracy (%)**	**Macro-F1**	**At-Risk Recall**
Logistic Regression	64.27	49.17	0.5037	0.7760
Random Forest	67.55	52.95	0.5451	0.7831
XGBoost Full (proposed)	**67.90**	**56.91**	**0.5867**	**0.7808**

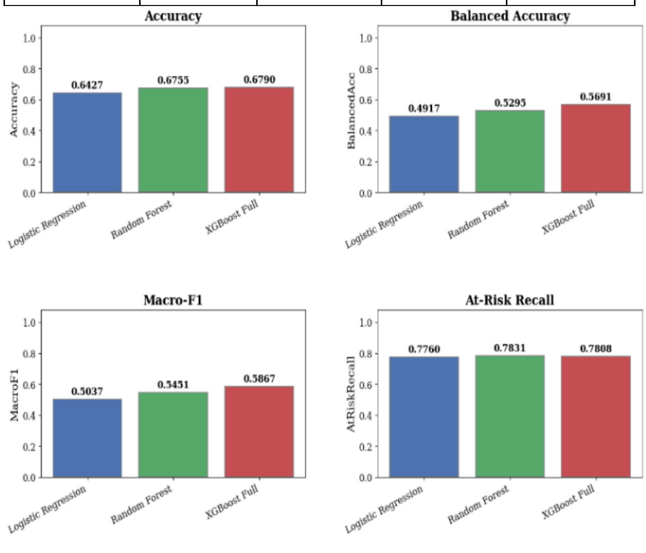

Fig. 3. Algorithm Comparison: Accuracy, Balanced Accuracy, Macro-F1, and At-Risk Recall for Logistic Regression, Random Forest, and XGBoost on Week-6 OULAD holdout test set. XGBoost achieves the highest performance across all metrics.

XGBoost is superior to Logistic Regression and Random Forest in this analysis. The difference between the two methods is 3.63%, where XGBoost outperforms Logistic

Regression by achieving an accuracy rate of 67.90% against 64.27%). The reason for this difference is that Logistic Regression uses linear relationships, while the student engagement pattern is more often not a linear relationship; therefore the tree structure of XGBoost is able to pick up on these inherent complexities to provide an edge. Random Forest uses bagging to combine separate trees. In contrast, XGBoost uses gradient boosting to fix mistakes from earlier trees sequentially. This approach provides stronger regularization and better performance on tabular data. It supports earlier findings that boosting methods often do better than bagging methods.

B. Error Analysis

The confusion matrices(Fig 4)show that misclassifications of the At Risk class decreased when all features were included. The performance of the Pass class remained steady with the addition of all features. The most difficult class, which is Distinction, generally has a lower recall than the other classes. This is expected in early predictions, as distinguishing Distinction from Pass usually needs stronger evidence that is often gathered later in the course.

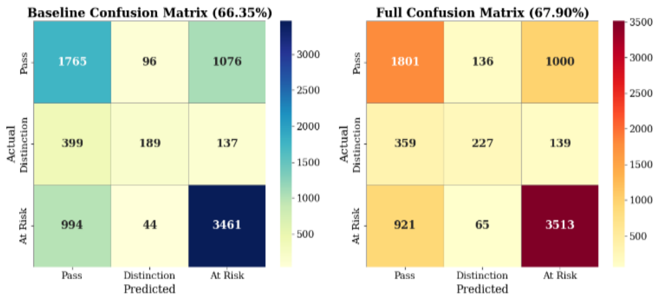

Fig. 4. Confusion matrices for baseline (left) and full (right) Week-6 models on the holdout test set (rows: true class; columns: predicted class). Class order: Pass (0), Distinction (1), At Risk (2).

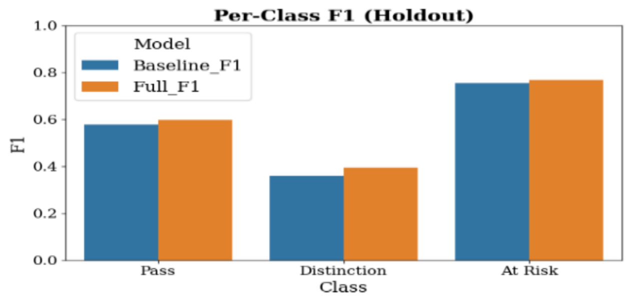

Fig. 5. Per-class F1-score comparison for baseline and full models on the holdout test set.

C. Cross-Validation Results

Figure 6 depicts the comparison between the performance of the baseline and the full models for the 5-fold cross-validation process. The baseline model reaches an average accuracy of 66.13% with only a 0.49% spread, while the full model achieves an average accuracy of 67.94%, with a slightly smaller spread of 0.32%. It is evident that the inclusion of all the features brings about a consistent increase in accuracy without variance across folds. From the results of the paired t-test, the p-value of 0.0018 confirms that the difference in performance from the full feature set is statistically.

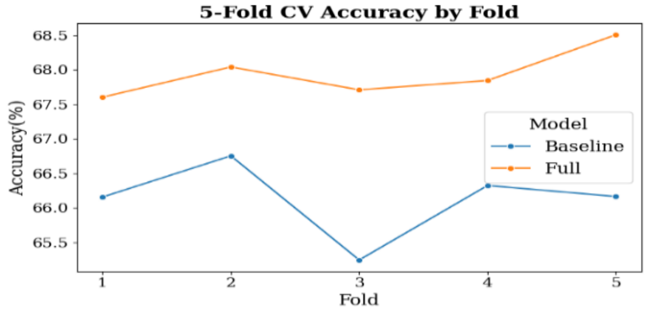

Fig. 6. Five-fold cross-validation accuracy per fold for baseline and full models.

D. Explainability Results (SHAP)

Figure 7 shows the global feature importance based on mean absolute contribution values. The analysis indicates that early assessment performance and engagement indicators are the most influential factors for predictions. Specifically, the average assessment score, number of active days, resource usage, and engagement trend greatly impact the model's decisions. The global measure of importance is determined by the mean (average) of the absolute values of the contributions of the various features/variables. The individual measure of importance indicates how specific features/variables have affected individual student's predicted outcome. Explanations based on the above provide insight into the variables associated with the classifications of At Risk, Pass, Distinction. classifications.

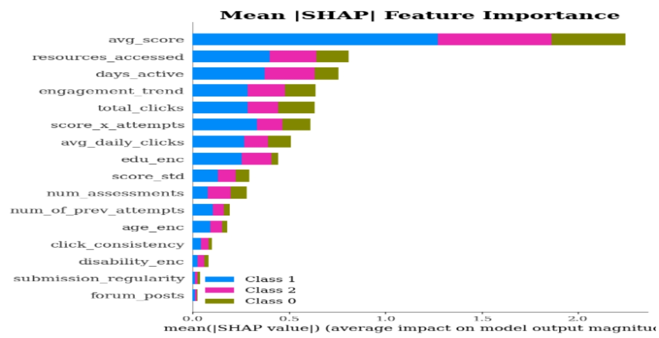

Fig. 7. Global feature importance for the Week-6 full model using mean absolute values $(mean(|\phi|))$. Labels are encoded as Class 0 = Pass, Class 1 = Distinction, and Class 2 = At Risk.

E. Recommendation Planner Demonstration (Qualitative)

To make Week-6 predictions actionable, the recommendation planner outlined in Section IV.F is demonstrated. The planner uses SHAP feature attributions to identify the main risk factors for each student and provides targeted intervention guidance. Table IV lists some examples. For At Risk predictions, the planner suggests proactive support actions (like tutor check-ins and structured study plans), while for Distinction predictions, it offers low-intensity "maintain" guidance.

Table IV. Example recommendation outputs using predicted class and planner uses predicted class + top SHAP driver

Student (anonymized)	Predicted	Actual	Recommendation (summary)
Student 1	At Risk	At Risk	Intervention needed: schedule tutor/TA check-in; prioritize deadlines; follow a daily study plan and

			review core materials
Student 2	At Risk	Pass	High support need: schedule tutor check-in; provide guided resources; increase daily VLE activity and practice key concepts
Student 3	Distinction	Distinction	On track: maintain engagement and keep submission schedule

These examples show how the system connects predicted risk to support-seeking actions and provides targeted, low-cost interventions ready for early implementation.

F. Comparison with prior work

Table V. Comparison with Prior Work

Study	**Prediction Stage**	**Task**	**Key Idea**
Kuzilek et al. [8]	—	Dataset	Introduced OULAD dataset
Hlosta et al. [13]	Late-stage	Binary	Early warning without historical data
Conijn et al. [14]	Mid–late	Binary	LMS features enable prediction
Wolff et al. [4]	Early (Week 3–4)	Binary	Clickstream supports early prediction
Proposed	**Week 6**	**3-class**	Early + explainable prediction with SHAP

Direct comparison with previous studies is difficult because earlier work reports performance using different evaluation metrics, such as AUC, precision, and recall, under various experimental conditions. Additionally, existing approaches differ in prediction timing, class formulation, and dataset subsets. Many studies focus on binary classification or use later-stage data, which makes direct numerical comparison with early three-class prediction settings hard. While some binary approaches report higher performance, these results are not directly comparable due to differences in prediction window, outcome definitions, and data selection. The proposed system achieves an accuracy of 67.90% at Week 6 on a three-class prediction task and also provides per-student interpretability. This mix of early prediction and explanation sets this approach apart from existing work.

V CONCLUSION AND FUTURE WORK

The work created an early warning system to determine which students would succeed in the OULAD, using data till week 6. The full 16-feature XGBoost model produced better results than the 9-feature behavioural baseline on the holdout data (67.90% vs. 66.35%) and consistently outperforming the baseline with 5-fold cross-validation (67.94% vs. 66.13%, paired t-test, p=0.0018).

In addition to allowing for the prediction of outcomes, this work provides insight into how the different features of the model generated output through the use of SHAP explanations, and also created a simple recommendation tool that allows for the generation of actionable recommendations for interventions

Future work will focus on improving distinction classification through cost sensitive learning/threshold tuning, the use of instructor recommendations in our recommendation system, and apply our analysis to OULAD program courses for additional applicability. Lastly, issues related to fairness and privacy in evaluating the effect of early intervention on students' performance will be considered.

VI REFERENCES

[1] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," J. Educ. Data Min., vol. 1, no. 1, pp. 3–17, 2009.

[2] G. Siemens and R. S. J. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in Proc. LAK, 2012, pp. 252–254..

[3] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Trans. Syst., Man, Cybern. C, vol. 40, no. 6, pp. 601–618, 2010.

[4] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek, "Improving retention: Predicting at-risk students by analysing clicking behaviour in a VLE," in Proc. LAK, 2013, pp. 145–149.

[5] Y. Jang, S. Choi, H. Jung, and H. Kim, "Practical early prediction of students' performance using machine learning and eXplainable AI," Educ. Inf. Technol., vol. 27, pp. 12855–12889, 2022.

[6] D. A. K. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, "Student retention using educational data mining and predictive analytics: A systematic literature review," IEEE Access, vol. 10, pp. 72480–72503, 2022.

[7] M. Nagy and R. Molontay, "Interpretable dropout prediction: Towards XAI-based personalized intervention," Int. J. Artif. Intell. Educ., 2023.

[8] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics Dataset," Sci. Data, vol. 4, no. 170171, 2017.

[9] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?" in Proc. NeurIPS, 2022.

[10] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, 2003.

[11] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," WIREs Data Min. Knowl. Discov., vol. 10, no. e1355, 2020.

[12] B. Albreiki, T. Habuza, and N. Zaki, "Framework for automatically suggesting remedial actions to help students at risk based on explainable ML and rule-based models," Int. J. Educ. Technol. High. Educ., vol. 19, no. 49, 2022..

[13] M. Hlosta, Z. Zdrahal, and J. Zendulka, "Ouroboros: Early identification of at-risk students without models based on legacy data," in Proc. LAK, 2017, pp. 6–15..

[14] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, "Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS," IEEE Trans. Learn. Technol., vol. 10, no. 1, pp. 17–29, 2017.

[15] H. Drachsler and W. Greller, "Privacy and analytics: It's a DELICATE issue — a checklist for trusted learning analytics," in Proc. 6th Int. Conf. Learning Analytics and Knowledge (LAK), 2016, pp. 89–98..