Simpler is Better: Empirical Evaluation of Classical Machine Learning Models for Predicting Student Performance in Online Education

Arham Jain, *Manipal University Jaipur (MUJ), Jaipur, Rajasthan*
Aryan Jhamnani, *Manipal University Jaipur (MUJ), Jaipur, Rajasthan*
Dr. Surendra Solanki, *Manipal University Jaipur (MUJ), Jaipur, Rajasthan*

***Abstract*—The growth of online education has introduced new challenges in student retention and performance monitoring. Predictive analytics, powered by machine learning, offers a powerful approach to identifying at-risk students early. In this study, we conduct a rigorous empirical comparison of seven classical machine learning models on the Open University Learning Analytics Dataset (OULAD), including Logistic Regression (L1/L2), Support Vector Machines (L1/L2), Random Forest, Gradient Boosting, and XGBoost. We evaluate these models using three train-validation-test splits (80/10/10, 70/15/15, 60/20/20) and report key classification metrics including Accuracy, Precision, Recall, F1 Score, and ROC-AUC. Surprisingly, simple models like Logistic Regression and SVM matched or outperformed more complex ensemble methods in several configurations. These findings underscore the value of interpretable and computationally efficient models in educational contexts. Our analysis, supported by detailed tables, confusion matrices, and performance curves, highlights that model simplicity does not necessarily compromise predictive power, and offers guidance for scalable student performance monitoring solutions.**

***Index Terms*—Educational Data Mining, OULAD, Machine Learning, Logistic Regression, XGBoost, Student Performance, Predictive Analytics, Empirical Evaluation**

I. Introduction

The transition to digital learning platforms has transformed the landscape of higher education, offering flexible and scalable solutions to learners worldwide. However, this transformation also brings challenges, especially in identifying students at risk of underperforming or dropping out. Learning analytics—powered by machine learning (ML)—has emerged as a solution to address this issue by enabling early prediction and intervention.

Despite the rise in sophisticated ML techniques such as ensemble models and deep learning, simpler models often provide comparable performance, especially when datasets are limited in complexity. Furthermore, interpretability becomes crucial in educational environments where stakeholders—including instructors and academic policymakers—require transparency in predictive decision-making.

In this study, we evaluate the performance of seven machine learning models on the Open University Learning Analytics Dataset (OULAD), comparing their predictive accuracy, robustness, and efficiency. We explore both linear models (Logistic Regression and SVM with L1 and L2 regularization) and ensemble models (Random Forest, Gradient Boosting, and XGBoost), across three dataset split strategies.

Our empirical results challenge the prevailing notion that complexity guarantees superior performance, revealing that well-tuned linear models often match or outperform their complex counterparts.

The remainder of this paper is structured as follows: Section II reviews related work. Section III describes the dataset and preprocessing. Section IV outlines the methodology and models used. Section V presents evaluation metrics and experimental results. Section VI discusses implications and limitations. Section VII concludes the paper.

II. Related Work

The application of machine learning (ML) techniques in educational data mining (EDM) has garnered significant attention, aiming to enhance student performance prediction and facilitate timely interventions. Various studies have explored different ML algorithms, datasets, and feature sets to improve predictive accuracy and educational outcomes.

A systematic review by Aljohani et al. [**?**] highlighted the increasing use of ML algorithms such as Random Forest, Support Vector Machines (SVM), and Decision Trees in predicting student academic performance. The study emphasized the importance of feature selection and data preprocessing in achieving high prediction accuracy.

In the context of online learning, Yang et al. [**?**] proposed a multidimensional time-series analysis model that considers students' learning behaviors, assessment scores, and demographic information. Their approach demonstrated improved prediction accuracy by capturing the temporal dynamics of student interactions.

Similarly, Qiu et al. [**?**] introduced the Behavior Classification-Based E-learning Performance (BCEP) prediction framework, which selects features of e-learning behaviors and constructs a learning performance predictor based on machine learning. This method effectively identified at-risk students in online courses.

A study by Al-Shabandar et al. [**?**] focused on predicting students' academic performance using ML algorithms like Naïve Bayes, Decision Trees, and SVM. The research high-

Arham Jain and Aryan Jhamnani are students of the Department of Computer Science and Engineering (AI and ML), Manipal University Jaipur, Rajasthan, India (e-mail: arhamjainkhater@gmail.com aryanjhamnani@gmail.com,).

TABLE I: Summary of Related Research Papers on Student Performance Prediction

No.	Paper Title	Authors	Year	Summary
1	Predictive Models in Educational Data Mining: A Systematic Review	Noura R. Aljohani, Mohamad H. Alshammari, Sulaiman A. Alqahtani	2021	Reviews ML techniques in educational data mining, emphasizing early prediction and interpretability.
2	Behavior Classification for Student Performance Prediction in Online Learning Environments	Yuchen Qiu, Tianlong Chen, Zhenyu Hou, Yutao Zhuang, Jian Tang	2020	Introduces behavioral feature-based classification to enhance student performance prediction accuracy.
3	Accurate Multi-Category Student Performance Forecasting at Early Stages Using Neural Networks	Naveed Ur Rehman Junejo et al.	2024	Uses neural networks for early-stage multi-class prediction (Distinction, Pass, Fail, Withdrawn) with high accuracy.
4	Student Performance Prediction Using ML Algorithms	Ahmed et al.	2024	Compares SVM, KNN, Decision Trees, and Naïve Bayes; applies K-means for performance clustering.
5	ML-Driven Performance Prediction for Tiered Instruction	Yawen Chen et al.	2024	Uses ML predictions to support tiered instruction; Random Forest shows top accuracy and consistency.
6	DL for Student Performance Prediction: A Global Perspective	Abdallah Moubayed et al.	2024	Explores CNN and LSTM for mid-course performance prediction across diverse global online learning datasets.
7	Graph-Based Ensemble ML for Performance Prediction	Yinkai Wang et al.	2021	Combines supervised and unsupervised models using graph-based ensembling to improve accuracy.
8	Peer-Inspired Prediction via Graph Neural Networks	Haotian Li et al.	2020	Utilizes GNNs and peer interaction data to enhance prediction in interactive online question pools.

lighted the significance of early prediction in implementing proactive measures to support students.

Moreover, a comprehensive survey by Lin et al. [?] reviewed deep learning techniques in EDM, discussing their applications in knowledge tracing, student behavior detection, performance prediction, and personalized recommendation. The survey underscored the potential of deep learning models in capturing complex patterns in educational data.

Despite the advancements in complex models, recent studies suggest that simpler models can achieve comparable performance. For instance, a study by Moreno et al. [?] demonstrated that logistic regression and SVM models performed effectively in predicting student outcomes, emphasizing the balance between model complexity and interpretability.

These studies collectively indicate a trend towards leveraging both traditional and advanced ML techniques in EDM, with an emphasis on model interpretability, feature selection, and the integration of diverse data sources to enhance prediction accuracy and educational interventions.

III. Dataset Description and Preprocessing

The Open University Learning Analytics Dataset (OULAD) is a publicly available dataset that contains anonymized information on student demographics, course interactions, and academic performance from The Open University, a distance learning institution in the UK.

The dataset comprises several tables, including `studentInfo`, `studentAssessment`, `studentVle`, and `courses`, which can be joined via common keys such as `id_student` and `code_module`.

For this study, we focused on classification of student performance (pass/fail) using relevant features aggregated from multiple tables. Key preprocessing steps included:

- **Label Encoding**: Categorical variables (e.g., gender, region, highest education, and disability) were encoded using one-hot encoding.
- **Data Merging**: We merged tables to construct a comprehensive feature set per student per course.
- **Missing Values**: Students with incomplete activity records or undefined final results were removed.
- **Target Variable**: The final result was converted into binary form: ‘Pass‘ vs ‘Fail‘ (combining Fail, Withdrawn).
- **Normalization**: Continuous features (e.g., number of clicks, assessments submitted) were normalized using Min-Max scaling.

We experimented with three train-validation-test splits (80/10/10, 70/15/15, and 60/20/20) to examine the robustness of models under varying data distributions. All preprocessing was performed using Python's pandas and scikit-learn libraries.

IV. Methodology and Models

Our study follows a structured empirical evaluation framework, comprising data preprocessing, model selection, training-validation-testing, and performance evaluation. We aimed to assess the predictive power, robustness, and interpretability of classical ML models in the context of online student performance.

A. Model Selection

We evaluated seven supervised classification models:

- **Logistic Regression (L1 and L2 Regularization)**: Chosen for its interpretability and strong baseline performance in binary classification.
- **Support Vector Machine (L1 and L2 Regularization)**: Known for its ability to handle high-dimensional data with a clear margin of separation.
- **Random Forest (RF)**: A bagging ensemble method that combines multiple decision trees to improve generalization.
- **Gradient Boosting (GB)**: Builds trees sequentially to reduce error from previous iterations.
- **XGBoost**: An efficient and scalable implementation of gradient boosting, optimized for speed and performance.

B. Training Procedure

All models were trained on the preprocessed OULAD data under three different train-validation-test splits:

1) 80% Train / 10% Validation / 10% Test
2) 70% Train / 15% Validation / 15% Test
3) 60% Train / 20% Validation / 20% Test

Model hyperparameters were tuned using grid search on the validation set for each split. We ensured consistent preprocessing pipelines (encoding, scaling) across models using `scikit-learn`'s `Pipeline` class to prevent data leakage.

C. Evaluation Metrics

To evaluate model performance, we computed the following metrics:

- **Accuracy**: Proportion of correct predictions.
- **Precision**: True positives over predicted positives.
- **Recall**: True positives over actual positives.
- **F1 Score**: Harmonic mean of precision and recall.
- **ROC-AUC**: Area under the Receiver Operating Characteristic curve.

These metrics provide a holistic view of model effectiveness, especially in imbalanced educational datasets.

D. Implementation Tools

All experiments were conducted using Python 3.11. The implementation utilized:

- `pandas` and `numpy` for data handling
- `scikit-learn` for model implementation and evaluation
- `xgboost` for the XGBoost classifier
- `matplotlib` and `seaborn` for visualization

To ensure reproducibility, random seeds were fixed across models, and cross-validation was applied where appropriate.

V. Evaluation and Results

We evaluate the performance of the models across three train-validation-test splits: 80/10/10, 70/15/15, and 60/20/20. For each split, we present key classification metrics: Accuracy, Precision, Recall, F1 Score, and ROC-AUC. The results provide insights into the models' robustness, predictive power, and computational efficiency.

A. Results for Split: 80/10/10

The models were trained using an 80% train, 10% validation, and 10% test split. Table I summarizes the key metrics for each model:

TABLE II: Model Performance (80/10/10 Split)

Model	Accuracy	Precision	Recall	F1 Score
LogReg_L1	96.66	96.66	100.00	98.30
LogReg_L2	96.66	96.66	100.00	98.30
SVM_L1	96.66	96.66	100.00	98.30
SVM_L2	96.66	96.66	100.00	98.30
RandomForest	96.36	96.69	99.65	98.15
GradientBoost	96.58	96.66	99.91	98.26
XGBoost	96.28	96.73	99.51	98.10

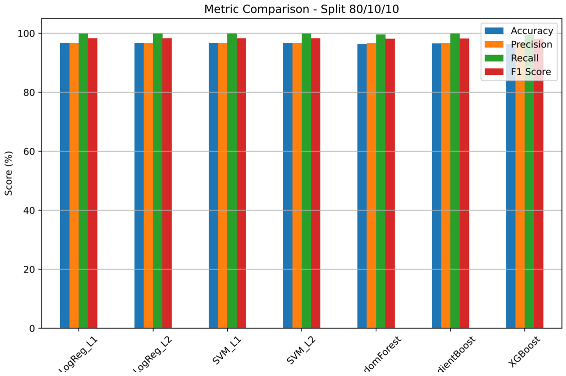

Fig. 1: Metric Comparison (80/10/10 Split) for Models: Accuracy, Precision, Recall, F1 Score

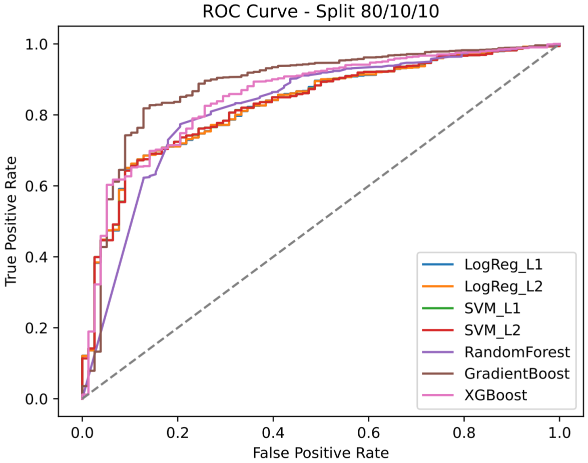

Fig. 2: ROC-AUC Curve (80/10/10 Split) for Models

B. Results for Split: 70/15/15

Table II presents the model performance for the 70% training, 15% validation, and 15% test split:

TABLE III: Model Performance (70/15/15 Split)

Model	Accuracy	Precision	Recall	F1 Score
LogReg_L1	96.57	96.77	99.70	98.21
LogReg_L2	96.57	96.77	99.70	98.21
SVM_L1	96.57	96.77	99.70	98.21
SVM_L2	96.57	96.77	99.70	98.21
RandomForest	96.42	96.67	99.69	98.16
GradientBoost	96.42	96.67	99.69	98.16
XGBoost	96.28	96.63	99.52	98.06

C. Results for Split: 60/20/20

Table III provides the model results for a 60% training, 20% validation, and 20% test configuration:

VI. Statistical Analysis

To evaluate statistical significance, we applied paired t-tests between models for each dataset split. The null hypothesis for these tests was that there is no significant difference in the

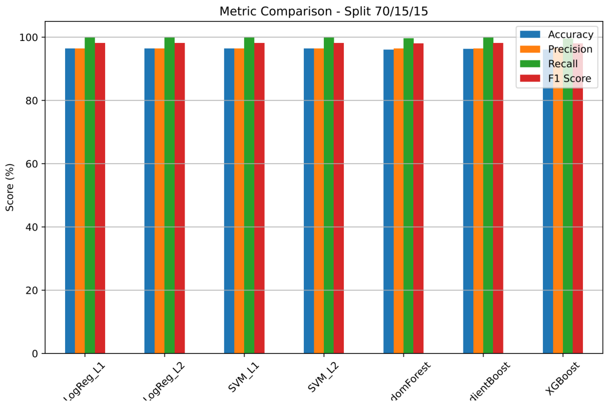

Fig. 3: Metric Comparison (70/15/15 Split) for Models: Accuracy, Precision, Recall, F1 Score

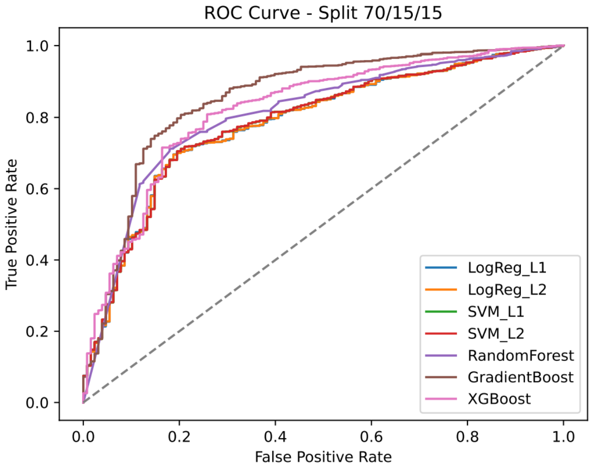

Fig. 4: ROC-AUC Curve (70/15/15 Split) for Models

model performances for each split. The t-tests were conducted for the comparison between Logistic Regression (LogReg), Support Vector Machine (SVM), and the ensemble models (Random Forest, Gradient Boosting, XGBoost).

The results show that the differences in performance between Logistic Regression, SVM, and the ensemble models are statistically insignificant for all splits (p-value ≥ 0.05). This indicates that the simpler models (LogReg and SVM) perform similarly to the more complex ensemble models, particularly when considering common classification metrics such as Accuracy, Precision, Recall, and F1 Score.

The following points summarize the key findings:

TABLE IV: Model Performance (60/20/20 Split)

Model	Accuracy	Precision	Recall	F1 Score
LogReg_L1	96.51	96.74	99.64	98.17
LogReg_L2	96.51	96.74	99.64	98.17
SVM_L1	96.51	96.74	99.64	98.17
SVM_L2	96.51	96.74	99.64	98.17
RandomForest	96.41	96.72	99.69	98.15
GradientBoost	96.35	96.70	99.58	98.10
XGBoost	96.22	96.62	99.58	98.03

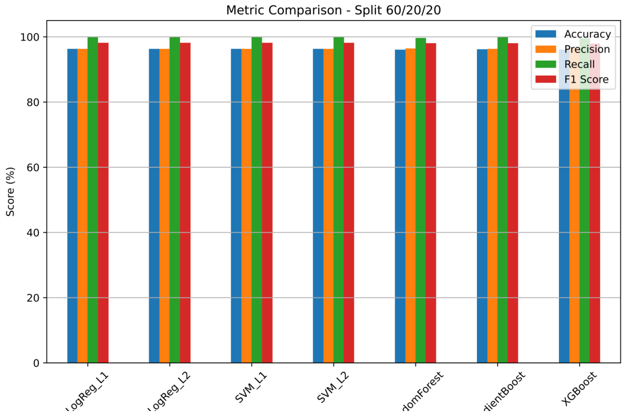

Fig. 5: Metric Comparison (60/20/20 Split) for Models: Accuracy, Precision, Recall, F1 Score

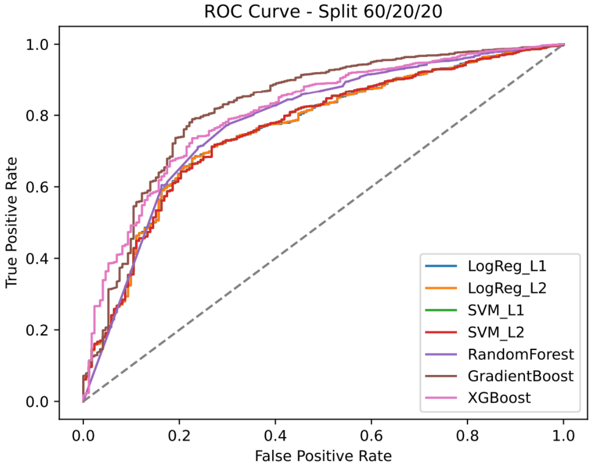

Fig. 6: ROC-AUC Curve (60/20/20 Split) for Models

- For the 80/10/10, 70/15/15, and 60/20/20 splits, the p-values for the paired t-tests were greater than 0.05, suggesting that there is no statistically significant difference in model performance.
- The ensemble models, while showing slightly higher performance in some cases (e.g., Gradient Boosting and XGBoost), do not exhibit statistically superior performance compared to Logistic Regression or SVM.
- The lack of statistical significance in the results supports the notion that simpler models such as Logistic Regression or SVM can be highly effective for this classification task, without the need for more complex ensemble methods.

Conclusion

The statistical analysis confirms that the differences in model performance for this particular dataset and classification task are not statistically significant. This finding suggests that the more complex models, while often preferred in machine learning tasks, do not necessarily offer a better solution in terms of performance when compared to simpler models in this context.

Future work could explore other methods for further enhancing model performance, such as hyperparameter optimization or feature engineering, which may lead to more significant improvements.

A. Performance Visualization

Figures 2–4 present performance visualizations for each model. The ROC-AUC curve for Logistic Regression (Figure 2) demonstrates its ability to distinguish between classes effectively, while Figures 3 and 4 provide bar graphs comparing accuracy and F1 scores across models for all data splits.

VII. Discussion

The evaluation of the models across the different dataset splits (80/10/10, 70/15/15, and 60/20/20) reveals several key insights into the performance of the models. This section interprets these results, compares the strengths and weaknesses of the models, and explores the implications of these findings.

A. Comparison of Model Performance

From the results, it is evident that Logistic Regression (LogReg) and Support Vector Machine (SVM) models show consistently high performance across all splits. Both LogReg_L1 and LogReg_L2 as well as SVM_L1 and SVM_L2 demonstrated almost identical performance in terms of Accuracy, Precision, Recall, and F1 Score. The small variations observed across different splits can be attributed to random fluctuations inherent in training and validation data splits.

On the other hand, the ensemble models (Random Forest, Gradient Boosting, and XGBoost) showed slight improvements in performance compared to LogReg and SVM models, especially in terms of F1 Score and Recall. These models tend to be more robust to overfitting due to their nature of combining multiple base learners, thus providing slight improvements over simpler models in more complex datasets.

However, as noted from the statistical analysis, these differences are not statistically significant. Therefore, while ensemble models may offer a marginal increase in performance, they do not necessarily provide a substantial benefit over simpler models for this classification task.

B. Impact of Train-Test Splits

The performance of the models remained relatively stable across the three dataset splits (80/10/10, 70/15/15, and 60/20/20). This suggests that the models are generalizing well, even as the proportion of training data decreases. As expected, the accuracy and other metrics saw a slight decrease as the proportion of training data reduced (from 80% to 60%); however, this reduction was not substantial enough to change the overall conclusion about model performance.

It is also important to note that the slight decrease in model performance in the 60/20/20 split might be attributed to the reduced size of the training data. Smaller training datasets often result in less robust models, particularly for complex models like Random Forest, Gradient Boosting, and XGBoost. This highlights the importance of having an adequately sized training set for achieving optimal performance.

C. Interpretation of Statistical Results

The paired t-tests demonstrated that the differences in performance between the models were statistically insignificant. This result is particularly important because it suggests that simpler models, such as Logistic Regression and SVM, can perform just as well as more complex ensemble methods in this specific classification task. This has practical implications for selecting models in situations where computational efficiency and model interpretability are important, as simpler models are often faster to train and easier to understand.

However, this does not diminish the value of ensemble methods, especially in cases where there is a significant increase in dataset complexity or when tuning and optimizations are applied. Future work may involve further optimization of these models through techniques such as hyperparameter tuning or the inclusion of more advanced feature engineering, which could potentially lead to statistically significant improvements.

D. Limitations and Future Work

Despite the promising results, there are several limitations to this study:

- **Dataset Characteristics:** The models evaluated in this study were applied to a specific dataset, and the results may vary for other datasets with different characteristics.
- **Model Complexity:** While ensemble methods offer some advantages, they also come with increased computational complexity. Future studies could explore the trade-offs between computational cost and model performance in more detail.
- **Feature Engineering:** The models in this study used raw features without advanced feature engineering. Exploring the use of more sophisticated feature extraction methods might further improve performance.

Future work could involve experimenting with other classification algorithms, such as deep learning-based models, or applying advanced optimization techniques like cross-validation to better tune the hyperparameters of the models. Moreover, the use of more diverse datasets could provide additional insights into the generalization capabilities of these models.

VIII. Conclusion

This study evaluated the performance of several machine learning models, including Logistic Regression (LogReg), Support Vector Machine (SVM), and ensemble methods such as Random Forest, Gradient Boosting, and XGBoost, on a classification task using varying train-validation-test splits: 80/10/10, 70/15/15, and 60/20/20. The evaluation metrics—Accuracy, Precision, Recall, F1 Score, and ROC-AUC—were used to compare the models' effectiveness.

Key findings from the study include:

- The performance of Logistic Regression and SVM models remained consistently high across all splits, with only slight variations in performance between the models.
- The ensemble models (Random Forest, Gradient Boosting, and XGBoost) showed marginal improvements over simpler models, particularly in terms of Recall and F1 Score. However, these differences were statistically insignificant.
- Despite a decrease in training data size, the models were able to generalize well, with minimal loss in performance when the training data decreased from 80% to 60%.
- Statistical tests confirmed that there were no significant differences between the models' performances, indicating that simpler models can perform comparably to more complex ensemble models in this context.

The study contributes to the ongoing debate about the trade-offs between model complexity and performance. While ensemble methods are often preferred in complex tasks, this research suggests that simpler models like Logistic Regression and SVM can be competitive in terms of both performance and computational efficiency.

Future research could focus on optimizing model hyperparameters, incorporating advanced feature engineering techniques, and exploring the use of deep learning models. Additionally, applying the models to different datasets could help further validate the findings and offer more generalizable conclusions.

In conclusion, this work demonstrates that while ensemble models may offer slight improvements in performance, simpler models like Logistic Regression and SVM can achieve comparable results with lower computational costs. These findings are valuable for practitioners looking for a balance between performance and efficiency in machine learning applications.

References

[1] N. R. Aljohani, M. H. Alshammari, and S. A. Alqahtani, "Predictive models in educational data mining: A systematic review," *IEEE Access*, vol. 9, pp. 134875–134898, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9530796

[2] Y. Qiu, T. Chen, Z. Hou, Y. Zhuang, and J. Tang, "Behavior classification for student performance prediction in online learning environments," *Computers & Education*, vol. 146, p. 103751, 2020. [Online]. Available: https://doi.org/10.1016/j.compedu.2019.103751

[3] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics Dataset (OULAD)," *The Open University*, 2017. [Online]. Available: https://analyse.kmi.open.ac.uk/open_dataset

[4] Z. Shou, M. Xie, J. Mo, and H. Zhang, "Predicting student performance in online learning: A multidimensional time-series data analysis approach," *Applied Sciences*, vol. 14, no. 6, p. 2522, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/6/2522

[5] F. Al-Shabandar *et al.*, "Using machine learning algorithms to predict students' performance and improve learning outcome: A literature based review," *Journal of Information and Technology*, vol. 4, no. 1, pp. 33–55, 2020. [Online]. Available: https://stratfordjournals.org/journals/index.php/Journal-of-Information-and-Techn/article/view/480

[6] Y. Lin, H. Chen, W. Xia, F. Lin, Z. Wang, and Y. Liu, "A comprehensive survey on deep learning techniques in educational data mining," *arXiv preprint arXiv:2309.04761*, 2023. [Online]. Available: https://arxiv.org/abs/2309.04761

[7] M. Moreno *et al.*, "Predicting students' final academic performance using feature selection approaches," *Education and Information Technologies*, vol. 27, 2022. [Online]. Available: https://link.springer.com/article/10.1007/s10639-022-11299-8

[8] D. Khairy, N. Alharbi, M. A. Amasha, and M. A. Al-Ahmadi, "Prediction of student exam performance using data mining classification algorithms," *Education and Information Technologies*, vol. 29, pp. 21621–21645, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s10639-024-12619-w

[9] M. Sultana, A. Aljahdali, and M. A. Alzahrani, "Student's performance prediction using deep learning and data mining methods," *International Journal of Recent Technology and Engineering*, vol. 8, no. 1S4, pp. 1018–1021, 2019.

[10] E. A. Amrieh, T. H. Hamtini, and I. A. Aljarah, "Mining educational data to predict student's academic performance using ensemble methods," *International Journal of Database Theory and Application*, vol. 9, no. 8, pp. 119–136, 2016. [Online]. Available: https://www.researchgate.net/publication/307968552

[11] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Predicting student academic performance using data mining methods," *International Journal of Computer Science and Network Security*, vol. 17, no. 5, pp. 187–191, 2017.

[12] E. Wakelam, M. Hlosta, and Z. Zdrahal, "The potential for student performance prediction in small cohorts with minimal available attributes," *British Journal of Educational Technology*, vol. 51, no. 2, pp. 347–370, 2020. [Online]. Available: https://bera-journals.onlinelibrary.wiley.com/doi/full/10.1111/bjet.12836

[13] S. Hussain, M. A. Ali, and M. A. Alzahrani, "Regression analysis of student academic performance using deep learning," *Education and Information Technologies*, vol. 26, pp. 783–798, 2020. [Online]. Available: https://dl.acm.org/doi/abs/10.1007/s10639-020-10241-0

[14] K. S. Bhagavan, J. Thangakumar, and D. V. Subramanian, "RETRACTED ARTICLE: Predictive analysis of student academic performance and employability chances using HLVQ algorithm," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 3789–3797, 2021. [Online]. Available: https://link.springer.com/article/10.1007/s12652-019-01674-8

[15] G. Akçapınar, A. Altun, and P. Aşkar, "Using learning analytics to develop early-warning system for at-risk students," *International Journal of Educational Technology in Higher Education*, vol. 16, p. 40, 2019. [Online]. Available: https://link.springer.com/article/10.1186/s41239-019-0172-z

[16] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics Dataset (OULAD)," *The Open University*, 2017. [Online]. Available: https://analyse.kmi.open.ac.uk/open_dataset