IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, July 2025
Impact Factor: 7.67

Leveraging Big Educational Data with Machine Learning for Student Success Prediction and Intervention

Kajol Khetan
VP of Marketing, Morgan Publishing House Inc.
kazolkhetan123@gmail.com

Abstract: *An Education provides the tools for success, boosts confidence, and prepares one for life's challenges. Colleges and universities are adapting their pedagogical approaches to take advantage of new technologies, such as artificial intelligence. One of the most important measures of educational success is students' performance in the classroom. A paradigm shift in education has occurred, innovation in technology, especially in the area of AI, has led to this. Applying OULAD, a dataset that includes over 32,000 student records supplemented with demographic information, academic, behavioral, and assessment data, this study delves into the application of machine learning approaches based on artificial intelligence to predict student performance and provide prompt intervention. A comprehensive methodology was employed, beginning with exploratory data analysis through visualizations, followed by data preprocessing, feature selection using the Pearson correlation coefficient, for numerical features, min-max normalization, and one-hot encoding for descriptive variables. The Light Gradient Boosting Machine (LightGBM) was chosen as the best model because of how well it handled big structured datasets compared to the others. The model achieved high predictive accuracy (92.23%), precision (94.40%), recall (93.21%), and F1-score (96.24%), outperforming other models as Logistic Regression, Random Forest, and Support Vector Machines. Results from ROC curves, precision-recall curves, and confusion matrices were used to further confirm the performance, demonstrating the model's robustness and potential to effectively support data-driven educational interventions.*

Keywords: Educational, student success factors, machine learning, deep learning, academic performance prediction, deep learning, educational data analytics, feature selection

I. INTRODUCTION

The field of education generates and stores vast quantities of data. Spending countless hours at home and in classrooms is a hallmark of the conventional educational model [1]. There is a wealth of data produced by students' engagement with course materials. Data on student engagement with online learning platforms and education management systems is collected by these systems [2][3]. A more comprehensive view of the learning process may be achieved through proper analysis of this data. Moreover, it has the potential to disclose valuable and, maybe, hidden relationships [4][5] the relationship between initial training level and subject performance in school, whether gender, attendance, or instructor has a role on students' ability to become proficient in a certain field, which classes' pupils perform the best. The development of big data analytics entails analyzing large datasets with a variety of data kinds in order to find hidden patterns [6], relationships, customer preferences, market dynamics, and more insightful information [7]. Big data analytics is widely used in corporate settings to forecast consumer trends and behaviors, but its incorporation into educational settings is yet largely unexplored. Students, teachers, educational researchers, course designers, educational institutions, and education administrators are the six main parties involved in education [8][9].

DOI: 10.48175/IJARSCT-28469

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, July 2025
Impact Factor: 7.67

The environment in which contemporary schools' function is complex and highly competitive. Consequently, most institutions are now dealing with concerns such as performance analysis, providing high-quality education, creating methods for evaluating student performance, and anticipating future needs. Universities use student intervention plans to help students who are struggling academically [10]. Predicting student performance at the entrance level and in later phases aids universities in efficiently creating and modifying intervention plans, which benefit both management and teachers. By using data-driven learning systems to give students quick, comprehensive feedback on how they are interacting with the material, big data analytics in education has the potential to completely transform the way that students learn [11][12]. Education has been completely transformed by recent developments in ML and AI [13][14]. Predicting student performance using AI and ML has become more popular as a data-driven, promising approach to identifying and resolving academic challenges [15][16]. The ability to analyze large amounts of data, including attendance patterns, academic records, and social and emotional factors, allows for the creation of a thorough academic profile for each student [17]. The growing need to enhance educational outcomes through data-driven decision-making is what spurred this study. As educational institutions are dealing with increasing student enrollments and diversified learning environments, it becomes rather difficult to pinpoint the students who might fall behind or drop out. Using ML on big datasets, such as the OULAD, educate and administrators are able to unearth patterns of student behavior, performance, and demographics, which are otherwise hidden to the unaided observation.

A. Research Contribution

The objective of this research is to construct accurate predictive models that may be used to support early interventions, personalized learning strategies, and improve academic planning. In addition, feature selection is implemented in order to use only the most significant features, resulting in higher efficient, interpretable, and scalable predictive systems that could benefit student success. This research makes several key contributions to the field of student success prediction in the education sector:

- This study utilizes the extensive and multidimensional OULAD, incorporating demographic, academic, behavioral, and assessment data in order to construct a strong model for predicting students' performance.
- The study proves that the LightGBM model outperforms conventional classifiers such as RF, SVM, and LR in predicting student outcomes.
- By employing Pearson correlation for feature selection and thorough preprocessing techniques, the study improves model efficiency and prediction accuracy.
- Multiple evaluation metrics the model's full performance may be shown through the usage of (accuracy, precision, recall, F1-score, ROC, and precision-recall curves).
- The research provides takeaways that can be applied in the educational systems to enhance personalized learning experiences, as well as increase student retention and success.

B. Significance of this Study

This research's primary contribution is to education, where it improves the application of ML methods for predicting students' performance. This study offers excellent prospects for improving the results of learning institutions by analyzing their issues and utilizing innovative techniques. The result of this study is of great value because it presents an effective and accurate way for predicting student success and for detecting those students who need extra help in time. The findings are informed by the use of the extensive and rich OULAD dataset, which includes demographic, academic, behavioral and assessment data, and so can be used to inform proactive, data-driven assessment tools to boost pupils' academic achievement. Implementation of the LightGBM model, meticulous feature selection, and comprehensive evaluation procedures, including accuracy, recall, and ROC curves, are what make this study innovative. The combination of DL and knowledge graphs improves predictive performance beyond what traditional ML models can achieve. In contrast to previous work, A method like this can achieve a great deal of accuracy while maintaining a high level of precision and recall and thus to lead to more dependable and actionable predictions. In the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, July 2025
Impact Factor: 7.67

end, this work helps to move towards personalized and effective learning support systems that can better fit the needs of heterogeneous students.

II. LITERATURE REVIEW

The review and analysis of numerous important research studies on Educational Data Prediction using ML were done for the purpose of informing and supporting the development of this work.

Angeioplastis et al. (2025), Researching if EDM methods may be used to improve learning outcomes and predict student success in the higher education sector. They used data from Moodle, a popular LMS, which analyses the academic records of 450 students across nine semesters, to accomplish this purpose. To determine the relationships between courses and forecast grades, KNN, RF, LR, DT, and neural networks were used. In the binary classification tasks, the findings demonstrated that courses where the correlation coefficient is +0.3 or higher greatly increased the predicted accuracy of neural networks and kNN; both models were able to raise F1 scores over 0.8. These findings demonstrate how optimizing instructional tactics through the use of EDM may eventually support individual learning paths. The data-driven strategies provide insights to enhance learning outcomes and enable student achievement [18].

Al-Hammouri et al. (2024) analysis of a higher education institution's dataset is followed by the development of an algorithm for categorizing pupils' performance in the classroom. The challenge of categorizing the issue as a multi-class classification with three types: Dropout, Enrolled, and Graduate. The problem is unbalanced and biased in favor of the Graduate. To increase prediction accuracy for the minority class, SMOTE with Edited Nearest Neighbor (SMOTE-ENN), a data balancing approach, is employed. Three well-known classification models—RF, XGBOOST, and CatBoost are employed. Utilizing SMOTE-ENN greatly enhances classification results, according to the data. Furthermore, the confusion matrix examination showed that XGBOOST had the best accuracy (94.6%) in accurately recognizing every class, outperforming earlier research in the literature. By using these models, dropout rates can be decreased, and precise projections of students' performance can be made [19].

Yang, Feng and Jiang (2024) propose an ML-based student learning behavior analysis and early warning system to monitor data on students' learning behavior in real-time in order to anticipate which kids could have difficulty learning and to provide timely early warnings. The system uses a learning management system and an online learning platform to collect data on attendance, homework submission, class participation, and online learning activities. After data cleaning, feature reduction, standardization, LR, DT, RF, and SVM are used to assess the models' performance throughout training. The findings from the experiments show, classically, good performance of an RF model at 90% accuracy, 88% precision, and 86% recall, further giving an F1 of 87%. In addition, the classification performance is also verified by a confusion matrix and an ROC curve [20].

Shannaq (2024) introduces technology that forecasts probable grades based on historical data from 1,972 student records with 26 characteristics, allowing students to make educated course selections. The research employs the (CRISP-DM) methodology, which ensures a methodical approach to data analysis and model creation. The findings demonstrate that with an accuracy rate of 96.37% vs. Random Tree's 84.10%, the J48 algorithm outperforms Random Tree by 12.27%. Random Tree has a misclassification rate of 15.90%, but J48 has a far lower rate of misclassification of 3.63%. J48 outperforms Random Tree in prediction accuracy, with a (MAE) of 0.0287 and a (RMSE) of 0.1259. J48 performs superior to every other category when measuring ROC areas, accuracy, and TPR, as evidenced by class-specific accuracy statistics [21].

Setiawan, Fatichah and Saikhu (2023) aspire to classify student feedback data into many labels. This work utilizes a Bidirectional Encoder Representation from Transformers (BERT) to extract word vectors from student feedback data. The classification of multi-label student feedback and performance comparison is done in this work using a number of ML techniques, including SVM, KNN, RF, and DT. The experiment was designed using a split of 80% training data and 20% testing data. It evaluated the guardianship information system for 3323 students. The SVM method using a linear kernel achieves the greatest results, with an accuracy of 82% and an F1 value of 90% [22].

Khalifa et al. (2023) provide the ARSITUN ML-based EDM technique for identifying pupils who are at danger. To reduce the likelihood of failure, ARSITUN can be used to conduct an early intervention for the identified children. The

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, July 2025
Impact Factor: 7.67

proposed approach was developed and tested using students' data that were collected from the Tunisian administration system for bachelors and masters called «Salima». It created a new dataset, named GCSD, that concerns 358 students from the faculty of Sciences of Gafsa during the school years 2014-2022. The experimental findings show that the EDM model achieves a 90.44% accuracy rate for computer science bachelor's grade prediction (Tunisian case study) [23].

Sanchez-Pozo et al. (2021) compare and contrast ML methods for academic performance prediction. The traits that were deemed most appropriate for seeing patterns in the academic performance of high school pupils were selected, striking a compromise between interpretability and accuracy. It used six supervised learning algorithms—LGBM, GB, AdaBoost, RF, LR, and KNN to identify patterns. The experimental findings demonstrated that, when weighed against competing classification approaches, the GB Classification algorithm had the greatest accuracy (96.77%) [24].

Gull et al. (2020) Classification and regression trees, LR, KNN, and Gaussian are some of the methods used in linear discriminant analysis. Using past grade data from one undergraduate course, they trained an NB and SVM model to predict how students will do in the same class following semester. Final test performance may be most reliably predicted using linear discrimination analysis, as shown in this experiment. A prediction accuracy of 90.74% was produced by the model, which was applied to 49 out of 54 data [25].

Numerous studies have applied ML in Predicting student achievement with EDM, using diverse datasets and models like k-NN, RF, SVM, NN, XGBoost, and BERT. Techniques such as SMOTE-ENN and CRISP-DM have improved data handling and analysis, achieving high accuracy in predicting outcomes like dropout risk and academic success. However, research gaps remain, including reliance on institution-specific datasets, inconsistent preprocessing practices, limited feature engineering beyond academic data, and a narrow focus on accuracy over model explainability and fairness. Advanced models like boosting and tree-based are underutilized, and real-time dataset integration is rare, limiting the practical application of these predictive systems.

Table I summarizes recent studies on Student Success Prediction and Intervention, highlighting innovative models, datasets used, key findings, and the challenges faced

Table 1: Overview of Recent Studies on educational data prediction using machine learning

Author	Proposed Work	Dataset	Key Findings	Challenges/recommend ation
Angeioplastis et al. (2025)	Improving learning outcomes and predicting student success using Educational Data Mining (EDM) techniques.	Academic records of 450 students from Moodle LMS across nine semesters.	Strong course correlations (+0.3 and above) improve grade prediction accuracy. kNN and Neural Networks achieved F1 scores > 0.8.	-Need for handling multi-class classification. Recommend refining feature selection and incorporating behavioral data for further enhancement.
Al-Hammouri et al. (2024)	Multi-class classification to predict student performance using Random Forest, XGBoost, and CatBoost	Higher education institution data (Graduate, Enrolled, Dropout; imbalanced)	XGBoost achieved the highest accuracy (94.6%) using SMOTE-ENN; improved minority class prediction	Address class imbalance; use advanced resampling like SMOTE-ENN
Yang, Feng and Jiang (2024)	Early warning system using student behavior data (attendance, homework, etc.)	LMS and online platform data	Random Forest performed best (90% accuracy, 88% precision, 86% recall, F1 = 87%)	Implement real-time monitoring and alerts
Shannaq (2024)	Predict student grades using CRISP-DM and J48 algorithm	Historical data from 1,972 students (26	J48 outperformed Random Tree with 96.37% accuracy vs.	Recommend J48 for high-accuracy predictive tasks

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, July 2025
Impact Factor: 7.67

		features)	84.10%; MAE = 0.0287, RMSE = 0.1259	
Setiawan, Fatichah and Saikhu (2023)	Classifying student feedback with multiple labels using BERT and ML classifiers	3,323 student feedback records (80/20 train-test split)	The best accuracy (82%) and F1 score (90%) were achieved using SVM with a linear kernel	Effective use of NLP with transformer models; explore other transformer-based approaches
Khalifa et al. (2023)	ARSITUN: ML-based EDM approach for identifying at-risk students	GCSD dataset from Salima system (358 students, 2014–2022)	EDM model achieved 90.44% accuracy for grade prediction	Suggests use of national systems for early intervention
Sanchez-Pozo et al. (2021)	Compare ML models for academic performance prediction	High school student data (selected optimal features)	Gradient Boosting achieved highest accuracy (96.77%)	Emphasize balance between accuracy and interpretability
Gull et al. (2020)	Predict final exam performance using various classifiers	54 historical student records	Linear Discriminant Analysis gave best accuracy (90.74%)	LDA works well with small, well-prepared datasets

III. RESEARCH METHODOLOGY

The Open University Learning Analytics Dataset (OULAD) is used in a structured data analysis pipeline as part of the research process, which comprises over 32,000 student records enriched with demographic, academic, behavioral, and assessment features.

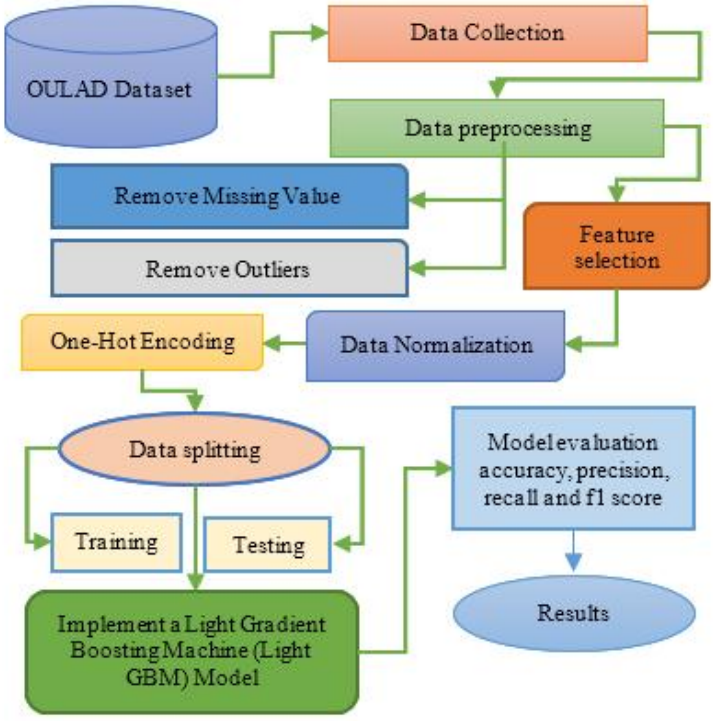

Figure 1: Proposed flowchart for Student Success Prediction

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, July 2025
Impact Factor: 7.67

It started with data exploration by looking into visualizations like heatmaps and bar charts to see what features are correlated with each other and what the target distribution looks like. Then rigorous data preprocessing was done. The most relevant predictors were identified employing statistical methods like the Pearson correlation coefficient and used to feed them to the feature selection process. Using the min-max scaling method, the numerical values were normalised, and the categorical features were encoded using the one-hot encoding method so that they could be used in ML models. The next step is to employ cross-validation to evaluate many models, after which it splits the dataset in half (75/25). Ultimately, a memory-efficient model called LightGBM was selected as the primary model due to its exceptional performance in handling vast amounts of structured data to forecast student achievement. In order to determine how well the model predicted and intervened at different levels of student accomplishment, it used industry-standard metrics including F1-score, recall, accuracy, and precision. Every step of the procedure is shown in Figure 1.
The following section it describes each step in the proposed flowchart for predictive modeling of student success using ML.

A. Data Collection

The dataset used for this study is the Open University Learning Analytics Dataset (OULAD), which is large and well-organized and was created to evaluate how well students do in school. OULAD is comprised of 32,593 student records from a variety of Open University (UK) courses offered across many academic years. The dataset is extremely valuable for educational data mining and predictive modelling since it is quite rich and incorporates a variety of data types, including behavioral, academic, demographic, and assessment-related data. Data visualizations such as bar plots and heatmaps were used to examine attack distribution, feature correlations etc., are given below:

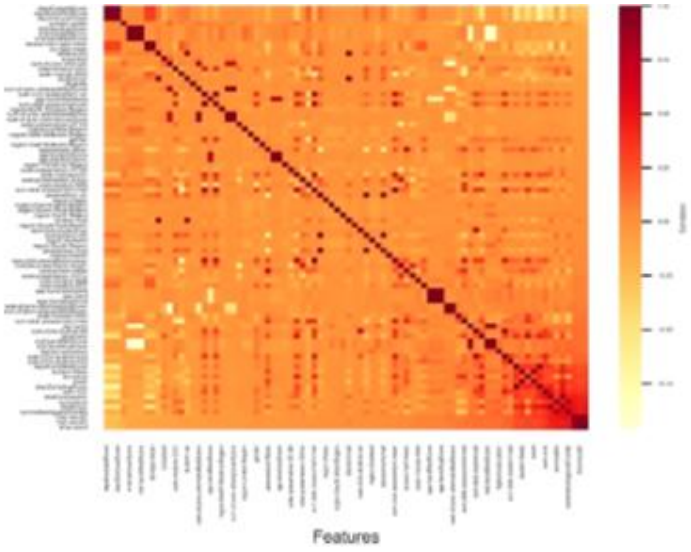

Figure 2 Correlations Heatmap

Figure 2 show a heatmap visualizing the correlation matrix of multiple features, with the color's intensity indicates the direction and degree of the feature pair association. Darker shades (approaching black or deep red) indicate stronger positive correlations, whereas lighter shades (ranging from yellow to white) indicate weaker or negative correlations. The diagonal line of dark squares indicates each feature's perfect correlation with itself. The feature names are listed along both axes, though they are difficult to read due to the resolution and angle. Overall, this visualization helps identify multicollinearity or feature relationships in a dataset, useful for tasks like feature selection or dimensionality.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, July 2025
Impact Factor: 7.67

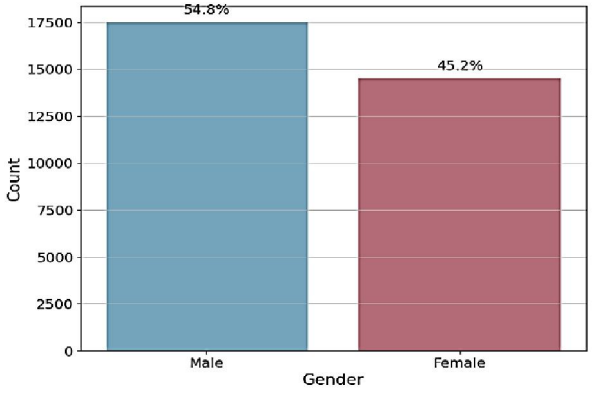

Figure 3 The Distribution of The Target Variable of Student Engagement.

Figure 3 the bar chart displays the gender distribution within a dataset, showing that males constitute a larger portion at 54.8%, while females make up 45.2%. The vertical axis represents the count, with males numbering just above 17,500 and females slightly below 15,000. The bars are colored distinctly blue for males and red for females and labeled with their respective percentages for clarity. This visualization highlights a moderate gender imbalance in the dataset, favoring male representation.

Data Pre-Processing

The data preparation is the process of preparing a dataset for analysis by cleaning, converting, and organizing it. Since this may greatly affect the accuracy and credibility of the investigation's results, it is a crucial action in data science. Moreover, data loss may occur due to a variety of reasons, including incorrect data entry, device malfunctions, lost files, and more. Many statistical and ML techniques are not designed to deal with outliers and missing values. Inadequate handling of missing values might result in inaccurate or biased conclusions. The following steps of pre-processing are as follows:

- **Remove Missing Value:** Missing values occur when there is no recorded data for a certain variable. Data loss may occur for many different reasons, including incorrect data entry, equipment failures, accidentally deleting files, and many more. Every dataset has some missing data. The imputer substituted the median value of that particular attribute for all missing values (NA).
- **Remove Outliers:** In order to prepare the dataset for additional analysis and model training, outliers were managed to guarantee data consistency and integrity.

C. Feature Selection

The process of determining which features (variables) are most pertinent to enhancing model performance is called feature selection. It may increase computing efficiency, improve model interpretability, and decrease the dataset's complexity by choosing key characteristics. The dataset's most pertinent attributes are selected using a variety of mathematical approaches. For numerical characteristics, The Pearson correlation coefficient (Equation (1)) is used to quantify the linear connection between two variables.

$$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2(y_i-\bar{y})^2} \qquad (1)$$

The mean of characteristics $\bar{x}$ and $\bar{y}$, and the correlation coefficient r_{xy} among them.

D. Data Encoding using One-Hot Encoding

The process of categorical data encoding involves converting transformation of categorical data into a numerical representation that ML systems can utilize. Important since many ML algorithms can only process numerical data and not directly deal with categorical data. To make sure the categories were seen by the ML model as independent of one

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, July 2025
Impact Factor: 7.67

another, one-hot encoding was used. This method of categorical data encoding creates an additional binary attribute (either 0 or 1) for each distinct group.

E. Data Normalization

The min–max approach was used to normalize the data, limiting the values to a range of 0 to 1. The performance of the classifiers was optimized, and the impact of outliers was reduced. The following mathematical Equation (2) states that the normalization was completed:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

If X represents the feature's initial value, X' represents its normalised value, X_{min} represents its minimum value, and X_{max} represents its maximum value.

F. Data Splitting

The first step in the model selection process was to utilize a 75/25 split to separate data into two sets: one for training purposes and another for testing purposes. After that, a large number of models were evaluated, and cross-validation was used to compare how well each model performed using the training data.

G. Proposed Light Gradient Boosting Machine (Light GBM) Model

A popular solution for a variety of ML issues, including regression, ranking, and classification, LightGBM is DT-based distributed gradient-boosting system that has shown to be highly effective. In order to create a powerful learning model, this kind of boosting technique combines several weak ML models. Boosting methods raise the weights of incorrectly classified data and lower the weights of well-classified data in the next training cycle, giving misclassified classifiers more weight. An algorithm called LightBGM was created with the GBDT backdrop in mind. This algorithm's ability to perform well via improving memory utilisation is its primary accomplishment. LightBGM is a histogram-based, highly optimized decision-making method that is regarded as XGBoost. This technique enhances the computational memory's effectiveness and optimal operation. The LightBGM's primary job is to reduce the discrepancy between the predicted and actual values, denoting the former with Z. The model is developed using two decision trees.

$$V = F\left(z_t, \hat{z}_t^i\right) + \theta(L_i) = \sum_1^f \left(z_t, \hat{z}_t^i\right) \sum_{t=1}^i \theta(1_t) \quad (3)$$

$$\hat{z}^i = \sum_{k=1}^i L_k(X) = \hat{z}^{-1} + L_i \quad (4)$$

The GB algorithm's objective function is shown in Equation (3). In the equation above, F stands for the loss function and θ for the regularization factor. As seen in Equation (4), the decision tree improvement with the i value is represented by $\hat{z}^i$.

H. Evaluation Metrics

The suggested architecture's efficacy was tested with a battery of performance measures. The actual values and the anticipated results of trained models were contrasted. This comparison was used to determine TN, FN, TP, and FP. The explanation of the following matrix, which includes F1-score, recall, accuracy, and precision, follows:

Accuracy: A balanced dataset, where the proportions of data points in the positive and negative groups are equal, has appropriate accuracy. It is given as Equation (5)-

$$Accuracy = \frac{TP+TN}{TP+Fp+TN+FN} (5)$$

Precision: Precision quantifies how well optimistic predictions work. It is the proportion of TP results to all cases that were anticipated to be positive. How good the classifier is in predicting the positive classes is expressed as Equation (6)-

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall: Recall, often referred to as sensitivity or TPR, quantifies the percentage of TP that the model accurately detects. In mathematical form it is given as Equation (7)-

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, July 2025
Impact Factor: 7.67

$$Recall = \frac{\text{TP}}{TP+FN} \quad (7)$$

F1 score: A measure that strikes a compromise between accuracy and recall is the F1 score, particularly in datasets with class imbalances. Mathematically, it is given as Equation (8)-

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

ROC Curve: At different thresholds, TPR vs. FPR is displayed on the ROC curve. The AUC helps to evaluate how good their model is in separating between classes, summarizing overall performance.

PR Curve: The Precision-Recall curve illustrates how Precision and Recall are traded off at certain levels. In an unbalanced dataset, it is helpful for identifying the positive cases when it matters.

The motivation for choosing these processes was that it allowed us to cope with the difficulties in working with heterogeneous and large educational data. Feature selection and data normalization are important preprocessing to guarantee that the model relies only on the relevant information and that the various data is handled in a consistent way. Since it is really suited for quick and accurate processing of big, structured datasets, it is especially suitable to use LightGBM, which would be well applicable for predicting student outcomes. All in all, this approach strikes a balance between data preparation and model selection in order to provide as much predictive performance as possible and practical utility in the realm of educational interventions.

IV. RESULTS AND DISCUSSION

Following training and testing, the experimental setup and performance of the proposed model are displayed below. In their model, used a 4GB RTC graphics card and an Intel Core i7 11th Gen CPU operating at 2.4 GHz to forecast how well students will do in school. It mostly utilizes Anaconda Spyder as integrated development environment (IDE) and Python as the primary language for doing simulations here. Table II displays the important performance metrics (accuracy, precision, recall, and F1-score) that were used to evaluate the proposed model after it was trained on the OULAD Dataset. Using the OULAD dataset, experimental findings on the suggested Light GBM model's ability to predict educational data have demonstrated strong predictive performance. With this model, the overall effectiveness of the model is high, as the accuracy achieved for it is 92.23%. The model has a precision of 94.40% and a recall of 93.21%, suggesting it is very good at correctly figuring out positive cases and also picking up a large proportion of TP. Further, the model's power and potential to conduct educational data analysis tasks are further demonstrated by the F1 score of 96.24, which balances accuracy and recall.

Table 2: Experiment Results of Proposed Models for of Educational Data Prediction using Machine Learning on OULAD Dataset

Performance Matrix	Light GBM Model
Accuracy	92.23
Precision	94.40
Recall	93.21
F1-score	96.24

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, July 2025
Impact Factor: 7.67

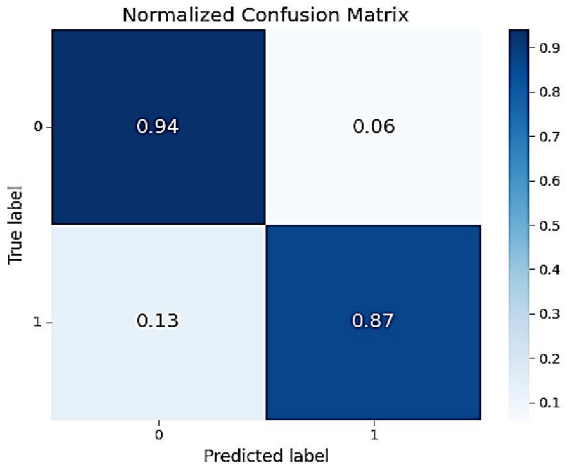

Figure 4: Confusion matrix for Light GBM Model

Figure 4 displays the normalized confusion matrix for the LightGBM model that was trained on the OULAD dataset, which clearly delivers its classification performance. For the cases where the model's prediction was True Label 1, it was wrong in 6% of the cases (Class 1D) and right in 94% (Class 0) of the cases. Class 1 (True Label 1) was predicted correctly in 87% of the cases, that is, it identified positive cases strongly, but misclassified 13% as Class 0. The matrix concludes that, overall, the model classifies Class 0 students with higher accuracy, and performs much more solid, with slightly lower performance, in terms of classification of Class 1 students and presents a good balance between true positives and TN on the OULAD dataset.

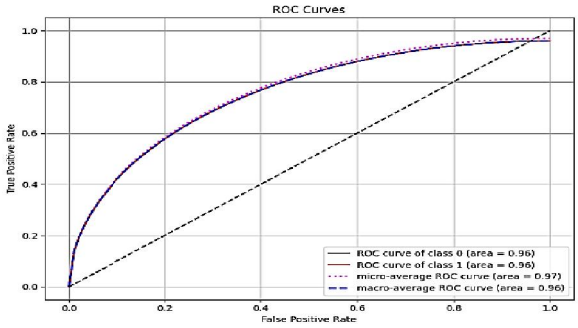

Figure 5 ROC Curves for LightGBM Model

Results from using the LightGBM model on the OULAD dataset are displayed in Figure 5 as ROC curves, highlighting its strong ability to distinguish between student success classes. Both Class 0 and Class 1 achieved an AUC of 0.96, while the micro-average and macro-average ROC curves scored 0.97 and 0.96, respectively. The curves' closeness to the top-left corner indicates high predictive performance and confirms the model's strong classification capability on the OULAD dataset.

Figure 6 illustrates the Precision-Recall (PR) curves for the LightGBM model predicting student performance on the OULAD dataset. The PR curve for Class 0 (students likely to succeed) shows very high precision across recall levels, with an area of 0.985, indicating strong identification with minimal FP. The area of the PR curve for Class 1 (students likely to struggle) is slightly lower (0.902), indicating that there is higher precision–recall trade-off as recall increases. The predicted result for the micro-average PR curve, on both classes, is an area of 0.970, which shows that the model produces good overall prediction results for student success and challenges on the dataset of the OULAD.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, July 2025
Impact Factor: 7.67

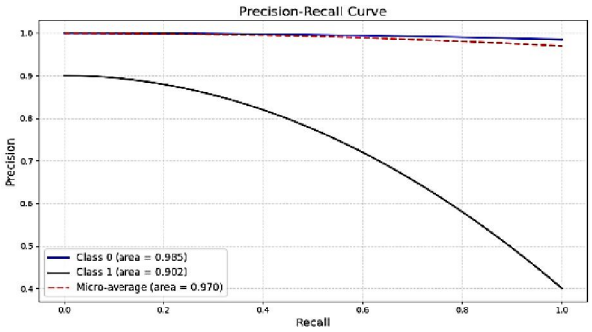

Figure 6 Precision-Recall analysis of the LightGBM model

The results show that the LightGBM model is very efficient in prediction of student success using the OULAD dataset, which provides benefits to educational data analytics. It has a solid foundation in terms of accuracy, precision, recall, and F1-score, as a result, to consistently and fairly identify pupils who are doing well and those who are at risk, and is able to make interventions in time. High ROC and Precision–Recall are indicative of strong discrimination with little FP, a requirement for reducing real-world misclassification given in educational settings. Finally, these results substantiate the use of advanced ML techniques such as LightGBM to boost personalized learning, increase students' retention, and enhance educational institutions' data-driven decision-making.

A. Comparative Analysis

The accuracy of this method was compared to other existing models to validate the usefulness of the proposed Light GBM model. Table III displays a comparison of the accuracy of several ML models used to forecast the students' performance using the OULAD dataset. Among all the models that were assessed, Light GBM has the best prediction ability, since it delivers the maximum accuracy of 92.23%. RF and SVM come with 88.3% and 87.71% accuracy, respectively and are still not as good as Light GBM. It seems that LR is not as efficient for this job as it has a significantly lower accuracy of 72.1%. The results of this thesis show that Light GBM is the promising model to predict student success, in this specific context.

Table 3: Accuracy Comparison of different Student Success Prediction using the OULAD Dataset

Models	Accuracy
Light GBM	92.23
RF[26]	88.3
SVM[27]	87.71
LR[28]	72.1%

The LightGBM model with the highest accuracy of 92.23% proved to be the most efficient model that can handle large-scale high-dimensional data. Its key advantage is that it uses the histogram-based algorithms, which makes the training faster and also reduces the memory usage. In addition, LightGBM supports parallel and GPU learning and is very scalable for educational datasets like OULAD.

V. CONCLUSION AND FUTURE STUDY

In the area of education, predicting student success is a crucial matter. Still, many have turned to ML methods to bolster the dependability and precision of student performance forecasts. Their study proposes a time-saving method for predicting students' grades by combining five machine learning strategies, such as data analysis, pre-processing methods, and LightGBM for grade classification. This study's findings show that ML is a potential prediction tool for student success using large-scale educational data, provided by the OULAD dataset, and more specifically, Light Gradient Boosting Machine (LightGBM). To train the data, the model used a structured pipeline, which included data exploration, preprocessing, sketching, encoding, normalization, and model evaluation, which produced better predictions as compared to traditional models such as RF, SVM, and LR. The model accurately represents the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, July 2025
Impact Factor: 7.67

complicated link between the many demographic, academic, and behavioral factors with good precision, recall, and F1-score values, and it achieves an accuracy of 92.23%. Overall, these results reveal the usefulness of using big educational data and state-of-the-art analytics to find indicators of risk earlier and deliver appropriate interventions in a timely manner. In this sense, this approach can provide possible benefits to improve educational outcomes and assist with personalized learning in higher education environments.

The study has some limitations despite its promising results. On the one hand, it bases itself on historical data from a single institution, which may hinder the external validity of the results to other educational settings. Secondly, the behavioral and psychological aspects which influence student success, for example motivation, stress, or external factors, etc., are not represented in the dataset. Third, although the accuracy of the LightGBM model was very high, interpretation of such complex models is an issue, and this can prevent practical use in academic settings. Future work will extend this work by including further data sources including real-time learning activity logs, socio-emotional indicators and qualitative feedback to increase prediction accuracy and context awareness. Furthermore, combining XAI techniques with the model will make the model more transparent and help educators understand what are the underlying factors that allow students to perform well. This framework can also be expanded across different institutions and educational systems that can further validate the scalability and effectiveness of the proposed approach.

REFERENCES

[1] N. V. M. Bindu and S. Singamsetty, "Enhancing Student Engagement and Outcomes through an Innovative Pedagogy for Teaching Big Data Analytics in Undergraduate Level," *Int. J. Comput. Math. Ideas*, vol. 16, no. 1, pp. 2000–2011, 2024.

[2] S. Tanwar, "Machine Learning," in *Computational Science and Its Applications*, 2024. doi: 10.1201/9781003347484-2.

[3] S. B. Kotsiantis, "Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades," *Artif. Intell. Rev.*, 2012, doi: 10.1007/s10462-011-9234-x.

[4] A. K. Polinati, "Revolutionizing Information Management: AI-Driven Decision Support Systems for Dynamic Business Environments," *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 35s, pp. 322–335, Apr. 2025, doi: 10.52783/jisem.v10i35s.6010.

[5] P. Choudhary and P. Potdar, "Robotics in STEM Education: Enhancing Engagement, Skills, and Future Readiness," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 3, no. 1, p. 11, 2023.

[6] S. Murri, M. Bhoyar, G. P. Selvarajan, and M. Malaga, "Transforming Decision-Making with Big Data Analytics: Advanced Approaches to Real-Time Insights, Predictive Modeling, and Scalable Data Integration," *Int. J. Commun. Networks Inf. Secur.*, vol. 16, no. 5, pp. 506–519, 2024.

[7] N. Abuzinadah *et al.*, "Role of Convolutional Features and Machine Learning for Predicting Student Academic Performance from MOODLE Data," *PLoS One*, vol. 18, no. 11, Nov. 2023, doi: 10.1371/journal.pone.0293061.

[8] S. Pandya, "Innovative blockchain solutions for enhanced security and verifiability of academic credentials," *IJSRA*, vol. 06, no. 01, pp. 347–357, 2022.

[9] R. Q. Majumder, "Machine Learning for Predictive Analytics: Trends and Future Directions," *Int. J. Innov. Sci. Res. Technol.*, vol. 10, no. 4, pp. 3557–3564, 2025.

[10] N. Prajapati, "The Role of Machine Learning in Big Data Analytics: Tools, Techniques, and Applications," *ESP J. Eng. Technol. Adv.*, vol. 5, no. 2, pp. 16–22, 2025, doi: 10.56472/25832646/JETA-V5I2P103.

[11] N. Aslam, I. U. Khan, L. H. Alamri, and R. S. Almuslim, "An Improved Early Student's Performance Prediction Using Deep Learning," *Int. J. Emerg. Technol. Learn.*, 2021, doi: 10.3991/ijet.v16i12.20699.

[12] R. Tarafdar, "AI-Supported Emotional Conflict Resolution: Technical Approaches and Implementation Strategies," *Int. J. Sci. Technol.*, vol. 16, no. 1, 2025.

[13] R. Kumar, "Leveraging LLMs for Continuous Data Streams_ Methods and Applications," *ICIDA*, 2025.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, July 2025 Impact Factor: 7.67

[14] P. Choudhary, R. Choudhary, and S. Garaga, "Enhancing Training by Incorporating ChatGPT in Learning Modules: An Exploration of Benefits, Challenges, and Best Practices," *Int. J. Innov. Sci. Res. Technol.*, vol. 9, no. 11, 2024.

[15] T. R. Noviandy, A. Maulana, T. Bin Emran, G. M. Idroes, and R. Idroes, "QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine Learning Algorithms," *Heca J. Appl. Sci.*, 2023, doi: 10.60084/hjas.v1i1.12.

[16] A. Maulana *et al.*, "Machine Learning Approach for Diabetes Detection Using Fine-Tuned XGBoost Algorithm," *Infolitika J. Data Sci.*, 2023, doi: 10.60084/ijds.v1i1.72.

[17] T. R. Noviandy, A. Maulana, G. M. Idroes, I. Irvanizam, M. Subianto, and R. Idroes, "QSAR-Based Stacked Ensemble Classifier for Hepatitis C NS5B Inhibitor Prediction," in *Proceeding - 2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering: Sustainable Development for Smart Innovation System, COSITE 2023*, 2023. doi: 10.1109/COSITE60233.2023.10250039.

[18] A. Angeioplastis, J. Aliprantis, M. Konstantakis, and A. Tsimpiris, "Predicting Student Performance and Enhancing Learning Outcomes: A Data-Driven Approach Using Educational Data Mining Techniques," *Computers*, vol. 14, no. 3, p. 83, Feb. 2025, doi: 10.3390/computers14030083.

[19] M. F. Al-Hammouri, Z. A. A. Hammouri, I. T. Almalkawi, and A. Lafee, "Optimizing Multi-Class Classification in Educational Data with Ensemble Learning and Data Balancing Techniques," in *2024 Fifth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, 2024, pp. 12–17. doi: 10.1109/IDSTA62194.2024.10746987.

[20] D. Yang, Y. Feng, and J. Jiang, "Student Learning Behavior Analysis and Early Warning System Based on Machine Learning," in *2024 International Conference on Language Technology and Digital Humanities (LTDH)*, 2024, pp. 171–176. doi: 10.1109/LTDH64262.2024.00042.

[21] B. Shannaq, "The Role of AI in University Course Registration in the Middle East: AI and Machine Learning Approaches to Improve Academic Performance," in *2024 2nd International Conference on Computing and Data Analytics (ICCDA)*, 2024, pp. 1–6. doi: 10.1109/ICCDA64887.2024.10867316.

[22] H. Setiawan, C. Fatichah, and A. Saikhu, "Multilabel Classification of Student Feedback Data Using BERT and Machine Learning Methods," in *2023 14th International Conference on Information and Communication Technology and System, ICTS 2023*, 2023. doi: 10.1109/ICTS58770.2023.10330849.

[23] A. Khalifa, F. BenSaid, Y. H. Kacem, and Z. Jridi, "At-Risk Students Identification based on Machine Learning Approach: A Case Study of Computer Science Bachelor Student in Tunisia," in *2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, 2023, pp. 1–8. doi: 10.1109/AICCSA59173.2023.10479243.

[24] N. N. Sanchez-Pozo, J. S. Mejia-Ordonez, D. C. Chamorro, D. Mayorca-Torres, and D. H. Peluffo-Ordonez, "Predicting High School Students' Academic Performance: A Comparative Study of Supervised Machine Learning Techniques," in *Future of Educational Innovation Workshop Series - Machine Learning-Driven Digital Technologies for Educational Innovation Workshop 2021*, 2021. doi: 10.1109/IEEECONF53024.2021.9733756.

[25] H. Gull, M. Saqib, S. Z. Iqbal, and S. Saeed, "Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning," in *2020 IEEE International Conference for Innovation in Technology, INOCON 2020*, 2020. doi: 10.1109/INOCON50539.2020.9298266.

[26] Y. N. Al Husaini, W. Al Kishri, M. A. Al Husaini, M. Al Bahri, and M. Abrar, "Predicting Student Academic Success Using Deep Learning: A Multi-Factor Approach to Performance Prediction," *J. Logist. Informatics Serv. Sci.*, vol. 12, no. 1, pp. 263–283, 2025, doi: 10.33168/JLISS.2025.0114.

[27] K. Wang, "Optimized Ensemble Deep Learning for Predictive Analysis of Student Achievement," *PLoS One*, vol. 19, no. 8 August, pp. 1–19, 2024, doi: 10.1371/journal.pone.0309141.

[28] H. A. Althibyani, "Predicting student success in MOOCs: a comprehensive analysis using machine learning models," *PeerJ Comput. Sci.*, vol. 10, p. e2221, 2024, doi: 10.7717/peerj-cs.2221.