A Systematic Literature Review of Student Performance Prediction Techniques in Virtual Learning Environment

S.Thilagavathi [1], Dr.V.Anuratha [2], Dr.M.Elamparithi [2]

[1] Research Scholar, Department of Computer Science, Kamalam College of Arts and Science, Anthiyur, Bharathiar University, Coimbatore, Tamilnadu, India

[2]Associate Professor, Department of Computer Science, Kamalam College of Arts and Science, Anthiyur, Bharathiar University, Coimbatore, Tamilnadu, India

Abstract: The virtual learning environment (VLE) is essential today and widely used globally for information exchange. Compared to in-person lectures, a VLE aids distant learning, although it might be challenging to maintain constant student interest. Academic activities are not actively pursued by students, which has an impact on their learning curves. The primary goal of this review is to impart a thorough knowledge and comprehension of various techniques, including machine learning (ML) and deep learning (DL), which are utilized for predicting student progress and performance and, consequently, how these prediction techniques help to find the most crucial student attribute for prediction. Additionally, this analysis reveals a rising trend in the volume and diversity of this field's research. At the same time, the assessment revealed several problems with research quality that highlight the need for the community to strengthen efforts to validate and replicate work and to describe methods and outcomes in greater detail. It can help teachers, parents, students, and tutors decide on the appropriate learning support for their charges when taking online courses.

Keywords: Student Academic Performance Prediction System, Virtual Learning Environment, E-Learning, Machine Learning, and Deep Learning.

1. INTRODUCTION

Education fosters learning, which includes acquiring information, values, abilities, beliefs, morals, habits, and personal development. Predicting student performance is increasingly important today because of how crucial it is to the growth of nations worldwide and how entirely dependent it is on the educational process that produces a generation able to assume the responsibility of leading this nation. In essence, gender, age, the teaching staff, and the student's earnings affect pupils' academic progress. It is crucial to accurately assess pupils in a VLE so they can get the finest education possible. This may significantly affect impaired pupils' learning and their ambitions to acquire a higher education degree. It has become essential for the educational environment to improve learning systems, particularly e-learning systems. Adaptable strategies to meet student needs have emerged due to ongoing developments in the e-learning process [1]. Organizations and educators have noted some difficulties with e-learning. Among the most important is determining what influences students' performance in online courses [2]. Therefore, effective methods are required to predict students' performance early.

One technique to forecast students' performance that improves the quality control of online training programs is ML. Many ML methods are used to forecast student performance in online courses, including Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), and Logistic Regression (LR). It generates two sorts of output: pass, the learner will complete the course successfully, and fail, the learner might not [3]. Teachers can gain a better grasp of their data by utilizing ML techniques. However, these techniques could perform better in generalizing knowledge and become useless when the data lacks sufficient volume for training and contains irrelevant features [4]. Many studies have lately employed the most well-known approach, DL, to predict students' academic achievement. Compared to other approaches, the DL methodology achieves high-level performance by automatically extracting the features. This survey covers the benefits and drawbacks of ML and DL techniques to forecast students' performance in e-learning systems effectively.

The rest of the paper is structured as follows: Section 2 discusses the background information of student's performance prediction system. Section 3 reviews methods used to predict the student's performance. Sections 4 and 5 give the discussion and conclusion with future studies.

2. BACKGROUND INFORMATION

The typical student performance prediction system is a binary classification task that divides students into two groups of "passed" or "failed" to forecast the likelihood of passing the test in the future. This study compared and analyzed several methods for forecasting student performance. Relevant articles were found, chosen, and critically assessed using various criteria before findings were incorporated. Data gathering, preprocessing, feature extraction, and classification are the primary procedures in the current study. Figure 1 depicts the overall flow of the system for predicting student achievement.

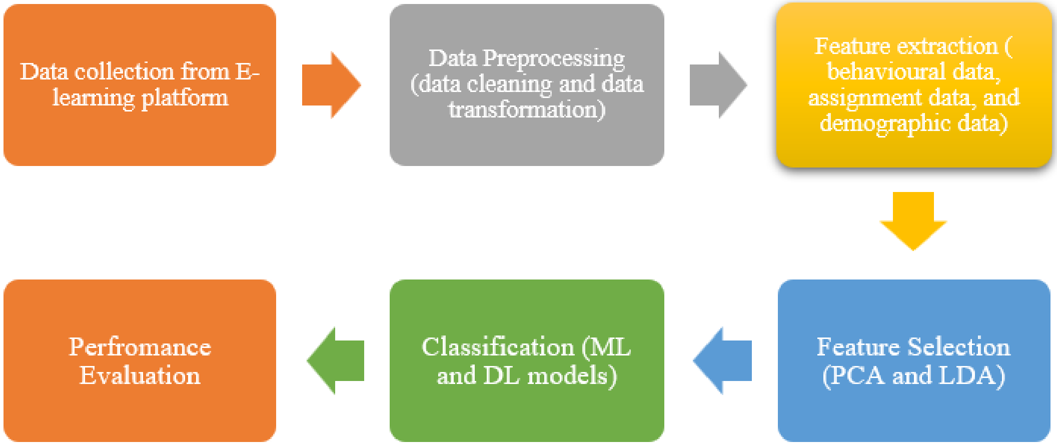
Data collection from E-learning platform
Data Preprocessing (data cleaning and data transformation)
Feature extraction (behavioural data, assignment data, and demographic data)
Feature Selection (PCA and LDA)
Classification (ML and DL models)
Perfromance Evaluation

Figure 1: Flow of the student's performance prediction system

2.1 Data Collection

In related studies, it has proven possible to forecast student performance and dropout rates using several available databases. These datasets can be used as benchmark datasets by researchers to assess how well the model performs compared to other models. The Open University Learning Analytics Dataset (OULAD), Center for Advanced Research via Online Learning (CAROL), HarvardX and MITx dataset (HMedx) and KDD Cup 2015 (KDDcup) are the public datasets in this field [5]. The OULAD dataset is used in most studies. It consists of 32,593 students in total, seven courses, and four semesters. A student's course results can fall into one of four categories: distinction, pass, fail, or withdrawal.

2.2 Data preprocessing

After accurate data has been gathered from publicly accessible sources, the data will be preprocessed. Before using predictive modelling, data preprocessing was required. Through various methods and procedures, data preprocessing comprises dealing with inconsistent data, removing data noise, and imputing missing values [6].

2.3 Feature Extraction

Feature extraction is a technique for eliminating unnecessary data from the original data set and extracting essential features. It increases the effectiveness of suggested strategies, decreases redundant data, increases model correctness, and quickens learning. Each student is given a different VLE type for feature extraction, which is then aggregated into a single value. Additional features were employed [7], including dynamic behavioural, demographic, and assignment features. Autoencoders and PCA, among other methods, are prevalent for feature extraction. It automatically extracts the necessary data for the prediction using DL techniques.

2.4 Classification

One of the most used techniques in prediction tasks is classification. The goal of classification is to establish the category of upcoming data objects using knowledge from the past. Many studies employ ML and DL primarily for classification in order to get the best accuracy results [8]. With the advent of the midterm exams in the ninth week, ML approaches were employed to forecast students at risk based on attendance, quizzes, and assignments. The most precise approach for classifying students as successful or failed and determining performance indicators is ML. Additionally, student engagement patterns successfully captured students' actions and convinced them to improve their performance. The development of DL for predicting student learning performance is still in its infancy. DL is a computer technique to investigate data representation at various levels of abstraction. It consists of numerous processing layers.

3. LITERATURE REVIEW

A student's academic achievement is one of the most crucial aspects of higher education. A thorough review of the literature in the area of learning analytics has been offered by a number of research that looked into studies that used learner behaviour analysis to forecast student performance. Similar to the survey done in this research, some studies presented a review analysis. Table 1 reviews recent works that used ML methods to predict students' performance in e-learning.

Table 1: Review of ML methods in predicting student's academic performance

Author name & Ref.	**ML Algorithms**	**Dataset Used**	**Outcomes**	**Advantages**	**Limitations**
Feiyue Qiu *et al.* [9]	Support vector classifier (SVC), NB, and KNN	OULAD dataset	Accuracy=97.40%	The framework was accurate and stable, ensuring the quality of online learners' learning.	Many algorithms prolonged the classification process and led to specific misclassification outcomes.
Ghassen Ben Brahim [10]	RF, SVM, NB, LR, and Multi-layer perceptron (MLP)	Digital Electronics Education and Design Suit (DEEDS) dataset	Accuracy of RF was 0.957%, SVM was 0.948%, NB was 0.826%, LR was 92.1%, and MLP was 0.957%	It has improved program learning results through better planning and precise adjustments to education management procedures.	Due to the model's intrinsic dependence on the specified features, NB produced the worst results.
Yutong Liu *et al.* [11]	LR, KNN, RF, and gradient boosting trees (GBT)	OULAD dataset	Highest accuracy=90.25%	To assist students who are at risk, this framework includes teaching intervention techniques.	Numerous algorithms could be complicated and hard to understand, making it impossible to predict outcomes.
Khurram Jawad *et*	RF	OULAD dataset	Accuracy=97.8%, Area	RF solves the problem of	With so many trees in RF, the

al. [12]			under characteristics (AUC)= 0.894%, and F1-score= 89.8%	overfitting as output is based on majority voting or averaging.	algorithm might need to be faster and more efficient for making predictions in real time.
Shrouk H. Hessen *et al.* [1]	DT, LR, RF, NB, and KNN	OULAD dataset	Accuracy of DT was 82.23%, LR was 80.59%, RF was 86.73%, NB was 67.45%, and KNN was 82.26%	The combination of these ML algorithms efficiently selected the high-ranked and relevant features, so the prediction rate was high.	Implementing and maintaining many ML algorithms could be costly, and specialized infrastructure for hardware may be required.
Abdulkream A. Alsulami *et al.* [2]	DT, NB, RF, and Ensemble methods (Boosting and Bagging)	Kalboard 360 E-Learning system	Accuracy= 76.88%, precision=0.768%, recall=0.769% , and f-measure=0.768%	It can be used as a proficient approach for the prediction of student performance, and it was better for real-time prediction scenario	Due to the requirement for simultaneous training, storing, and integrating numerous models, it was both time- and resource-intensive.
Mohammed Nasser Alsubaie [13]	SVM	Maharat platform at Taif University student's data	Accuracy=93.2%	With a surface that maximized the margin between them, SVM was used to divide a number of classes in the training set.	Due to the quadratic growth of the SVM's kernel matrix, training SVMs on big data sets was highly laborious.

Most users recently used DL techniques for student academic performance prediction in e-learning systems. An artificial neural network (ANN) was suggested by **Alberto Rivas *et al***. [14] for forecasting students' academic performance in online learning environments. The dataset was initially collected from a group of students who had taken four separate online courses. The acquired dataset was then subjected to normalization to enhance the data's quality. Finally, ANN was employed to predict the academic performance of the student. The experimental findings demonstrated that the system outperformed previous approaches by attaining 0.782% precision and recall and 0.781% f-measure, respectively. **Monika Hooda *et al.*** [15] proposed a system based on an enhanced fully connected network (FCN) to enhance students' academic performance. Data was initially acquired for the system via the OULAD database, compiled from Open University students. The dataset was then cleaned up by doing data preprocessing on the collected data. Following that, the FCN algorithm was used to forecast student performance. The stochastic gradient descent (SGD) approach was used to determine the optimal learning parameters for the FCN algorithm. The system had an accuracy of 84%, which was higher than the existing schemes.

Sadique Ahmad et al. [16] used an iterative model of frustration severity to predict students' performance. Frustration was first separated into its four outer levels. Second, the academic outcome for students was divided into 34 inner layers. The prediction was then iteratively optimized through outer and inner iterations under the guidance of frustration severity layers. During the experiment, the system's accuracy was compared to a dataset of student scores, and the results showed that the system had a greater accuracy of 0.79%. **Xiaoxia Jiao [17]** suggested a student physical performance prediction system using a factorization deep product neural network (DPNN). Initially, the system used an embedding layer that transformed the input higher-dimension features into lower dimensions. Then, the first-order, second-order, and higher-level features were expressed using factorization and DPNN in the concatenation layer. Finally, the prediction was made using the prediction layer, and the system attained 0.87% accuracy and precision, 0.91% f-score and 0.95% recall on the OULAD dataset.

Heyul Chavez *et al.* [3] predicted students' academic performance using ANN. The system used the Open University of the United Kingdom dataset, which contains 32,000 student's data. Then, the collected data was fed into the preprocessing stage to improve prediction accuracy. Finally, ANN was utilized to predict whether the student will pass or fail in the considered learning model. The system attained an accuracy of 93.81%, precision of 94.q5%, recall of 95.13%, and 94.64% of f-score, which was better than the existing models. **Xiao Wen and Hu Juan [18]** recommended a deep neural network (DNN) for predicting students' performance in online learning. Initially, a pre-trained auto-encoder extracted latent features from the input sequences. The features were given to the DNN to predict the student's performance. The system attained an accuracy of 0.84 when tested on the OULAD dataset.

Ming Li *et al.* [19] introduced a multi-topology graph neural network (MTGNN) for student performance prediction. Initially, the system used the OULAD dataset for data collection, and then preprocessing was done on the collected data to improve the prediction quality of the classifier. Then, the graph was constructed for the preprocessed data using similarity learning, and the constructed graph was fed into the MTGNN for classification. The system attained a f-score of 92.59%, recall of 97.60%, and accuracy of 91.95%, which were better than the previous models. **Sri Suning Kusumawardani and Syukron Abu Ishaq Alfarozi [20]** proffered a student's performance prediction model using a transformer encoder, which worked based on the student's log activities. The system attained 83.17% accuracy on the OULAD dataset, which was entirely satisfactory.

4. DISCUSSION

Predicting student performance has been an exciting field of study for many scientists and researchers in cognitive computing who are also interested in education. It has numerous uses for crucial cognitive tasks, including performance forecasting during class activities, written tests, and student quizzes. This comprehensive study examined the ML and DL methods currently used to assess student performance. The findings are conflicting because they come from several authors. The available research demonstrates that ML and DL classification algorithms generate accurate and reliable prediction accuracy.

Additionally, it is clear from the data comparison study that the authors employed both supervised learning and unsupervised learning techniques to forecast the students' performance. However, most studies used minimal amounts of data to train the ML techniques. However, ML algorithms indeed require a vast amount of data in order to function well. Additionally, most of the surveys described above use numerous ML methods for prediction, making the system slow and computationally complex. DL algorithms produce superior results over conventional ML techniques. Most surveys use an ANN model to forecast students' academic performance in virtual learning environments. Due to its parallel features, it can continue the process without any problems, but it uses a lot of processing power. Numerous authors have recently proposed convolutional neural networks (CNN), recurrent neural networks (RNN), and other well-known DL techniques. These techniques produce higher-level performance than the existing ML schemes. The research also showed that a few studies concentrated on class or data balancing. Because it prevents the model from becoming biased towards one class, balancing a dataset makes it easier to train models, which is crucial for achieving good classification performance.

5. CONCLUSION

This study examines the predictive model and recent developments in VLE prediction. One of the most current topics in the E-learning system is predicting student academic achievement. According to earlier studies, several variables impacted students' academic achievement, including student family income, family size, mother's education level, and student learning behaviour. The primary goals of this survey are to enhance participants' behaviour and offer the most outstanding early prediction performance. Overall, this evaluation successfully achieved its goals of raising student performance by identifying at-risk individuals and emphasizing the value of applying both ML and DL models. The results of this study can aid parents, instructors, students, and tutors in deciding on the best educational support for their children. Future system versions will include cutting-edge deep learning techniques, which can significantly increase accuracy by choosing the features based on an improved understanding of the student's performance prediction.

REFERENCES

[1] Hessen, Shrouk H., Hatem M. Abdul-Kader, Ayman E. Khedr, and Rashed K. Salem. "Developing multiagent e-learning system-based machine learning and feature selection techniques." *Computational Intelligence and Neuroscience* 2022 (2022).

[2] Alsulami, Abdulkream A., Abdullah S. AL-Malaise AL-Ghamdi, and Mahmoud Ragab. "Enhancement of E-Learning Student's Performance Based on Ensemble Techniques." *Electronics* 12, no. 6 (2023): 1508.

[3] Chavez, Heyul, Bill Chavez-Arias, Sebastian Contreras-Rosas, Jose María Alvarez-Rodríguez, and Carlos Raymundo. "Artificial neural network model to predict student performance using nonpersonal information." In *Frontiers in Education*, vol. 8, p. 1106679. Frontiers, 2023.

[4] Jhaveri, Rutvij H., A. Revathi, Kadiyala Ramana, Roshani Raut, and Rajesh Kumar Dhanaraj. "A review on machine learning strategies for real-world engineering applications." *Mobile Information Systems* 2022 (2022).

[5] Alhothali, Areej, Maram Albsisi, Hussein Assalahi, and Tahani Aldosemani. "Predicting student outcomes in online courses using machine learning techniques: A review." *Sustainability* 14, no. 10 (2022): 6199.

[6] Li, Shuping, and Taotang Liu. "Performance prediction for higher education students using deep learning." *Complexity* 2021 (2021): 1-10.

[7] Alsariera, Yazan A., Yahia Baashar, Gamal Alkawsi, Abdulsalam Mustafa, Ammar Ahmed Alkahtani, and Nor'ashikin Ali. "Assessment and evaluation of different machine learning algorithms for predicting student performance." *Computational Intelligence and Neuroscience* 2022 (2022).

[8] Ragab, Mahmoud, Ahmed MK Abdel Aal, Ali O. Jifri, and Nahla F. Omran. "Enhancement of predicting students performance model using ensemble approaches and educational data mining techniques." *Wireless Communications and Mobile Computing* 2021 (2021): 1-9.

[9] Qiu, Feiyue, Guodao Zhang, Xin Sheng, Lei Jiang, Lijia Zhu, Qifeng Xiang, Bo Jiang, and Ping-kuo Chen. "Predicting students' performance in e-learning using learning process and behaviour data." *Scientific Reports* 12, no. 1 (2022): 453.

[10] Brahim, Ghassen Ben. "Predicting student performance from online engagement activities using novel statistical features." *Arabian Journal for Science and Engineering* 47, no. 8 (2022): 10225-10243.

[11] Liu, Yutong, Si Fan, Shuxiang Xu, Atul Sajjanhar, Soonja Yeom, and Yuchen Wei. "Predicting Student performance using clickstream data and machine learning." *Education Sciences* 13, no. 1 (2022): 17.

[12] Jawad, Khurram, Muhammad Arif Shah, and Muhammad Tahir. "Students' Academic Performance and Engagement Prediction in a Virtual Learning Environment Using Random Forest with Data Balancing." *Sustainability* 14, no. 22 (2022): 14795.

[13] Alsubaie, Mohammed Nasser. "Predicting student performance using machine learning to enhance the quality assurance of online training via Maharat platform." *Alexandria Engineering Journal* 69 (2023): 323-339.

[14] Rivas, Alberto, Alfonso Gonzalez-Briones, Guillermo Hernandez, Javier Prieto, and Pablo Chamoso.

"Artificial neural network analysis of the academic performance of students in virtual learning environments." *Neurocomputing* 423 (2021): 713-720.

[15] Hooda, Monika, Chhavi Rana, Omdev Dahiya, Jayashree Premkumar Shet, and Bhupesh Kumar Singh. "Integrating LA and EDM for improving students Success in higher Education using FCN algorithm." *Mathematical Problems in Engineering* 2022 (2022).

[16] Ahmad, Sadique, Najib Ben Aoun, Mohammed A. El Affendi, M. Shahid Anwar, Sidra Abbas, and Ahmed A. Latif. "Optimization of Students' Performance Prediction through an Iterative Model of Frustration Severity." *Computational Intelligence and Neuroscience* 2022 (2022).

[17] Jiao, Xiaoxia. "A Factorization Deep Product Neural Network for Student Physical Performance Prediction." *Computational Intelligence and Neuroscience* 2022 (2022).

[18] Wen, Xiao, and Hu Juan. "Early Prediction of Students' Performance Using a Deep Neural Network Based on Online Learning Activity Sequence." *Applied Sciences* 13, no. 15 (2023): 8933.

[19] Li, Ming, Xiangru Wang, Yi Wang, Yuting Chen, and Yixuan Chen. "Study-GNN: a novel pipeline for student performance prediction based on multi-topology graph neural networks." *Sustainability* 14, no. 13 (2022): 7965.

[20] Kusumawardani, Sri Suning, and Syukron Abu Ishaq Alfarozi. "Transformer Encoder Model for Sequential Prediction of Student Performance Based on Their Log Activities." *IEEE Access* 11 (2023): 18960-18971.

DOI: https://doi.org/10.15379/ijmst.v10i2.2961