International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

Students Performance Prediction in Virtual Learning Environment Using a Deep Learning System

Thilagavathi.S[1*,] Dr.Anuratha.V[2] and Dr.Elamparithi.M[3]
Department Of Computer Science, Kamalam College of Arts and Science, Anthiyur,
Bharathiar University, Coimbatore, Tamil Nadu, India.
E-Mail: [1*]thilagasubbu@gmail.com, [2]profanuratha@gmail.com, [3]profelamparithi@gmail.com

Abstract: These days, virtual learning environments (VLEs) are essential and widely used worldwide for information exchange. While VLE benefits remote learning, in-person lectures are more challenging to maintain student engagement than in-person lectures, contributing to the high dropout rates among students. Student's learning curves are impacted by their lack of active participation in academic activities. Therefore, the VLE needs to give more attention to severe academic achievement. This paper proposes a novel enhanced activation function-centered recurrent neural network (EARNN) with a K-means Synthetic Minority Over-sampling Technique (KMSMOTE) to predict students' performance in VLE. The four primary steps of the suggested system are data gathering, preprocessing, balancing, and classification. First, the proposed system collected the data from the Open University Learning Analytics Dataset (OULAD) dataset. Next, the system performs preprocessing on the collected dataset to improve its quality. After that, the data balancing is done using KMSMOTE. Finally, the classification of student performance is done by EARNN, which combines demographic, assessment, and click stream features as input. The outcomes demonstrated that the suggested work performs superior to the existing techniques.

Keywords: *Virtual Learning Environment, E-Learning, students' academic performance, Preprocessing, Dataset Balancing, Classification, and Deep Learning.*

1. INTRODUCTION

A growing number of people are becoming interested in distance education due to the advancements in digital resource discovery and the popularization of Internet access. The key advantages of this education are its accessibility (students can attend a course from any place in the world) and adaptability (students can fit their study into their routine) [1]. Higher education has been impacted by the new digital environment [2]. Higher education (HE) institutions use digital technologies like the VLE, which supports students in their attempts to meet specific HE goals and improve the student experience [3]. It dramatically impacts a student's capacity to keep their mental health in check [4]. As the largest university in the UK today, the Open University alone has graduated 2 million students from 157 countries since its founding. It was a pioneer in the field of online education. Numerous esteemed academic institutions, including Stanford University, Harvard University, and Massachusetts Institute of Technology, have introduced massively open online learning environments with course offerings [5]. The most significant obstacle facing the education sector in the era of online learning was the efficient and trustworthy assessment of students' performance in virtual learning environments. The assessment procedure becomes complicated when individuals use printed materials, the Internet, and other helpful resources to obtain information while taking the test [6].

Because of this, evaluating or generating predictions is a crucial component of learning, and everyone involved—students, teachers, and colleagues—should exercise caution when doing so. A student may still be able to succeed even if they do not perform as well as others during the learning process [7]. Machine learning (ML) approaches could be helpful in analyzing a student's recent academic performance in an online learning environment, where data is collected daily [8]. Machine learning systems use data to learn, look for patterns and forecast results. Due to increasing data quantities, less expensive storage, and reliable computational systems, Deep Learning (DL) techniques have replaced pattern recognition algorithms as the go-to machine [9]. DL models can accurately analyze more extensive and more complicated data sets automatically, rapidly, and without taking unanticipated risks [10]. This motivates us to propose a student's performance prediction in an e-learning platform using a DL system. The main contributions of the paper are outlined as follows:

- The study employs KMSMOTE to handle unbalanced data sets, which prevents the issue of new sample marginalization by carefully choosing samples from the minority class.

- To anticipate students' final performance, the system suggests the EARNN model, which uses students' click stream, assessment, and demographic features to increase prediction accuracy.

The remaining parts of the paperwork are organized as follows: Part 2 examines current research on the subject of the study proposal. The suggested methodology's brief explanations are given in Section 3. The simulation outcomes and discussion for the suggested system are explained in Section 4, and the proposed work is concluded in Section 5.

2. RELATED WORK

Monika Hooda *et al.* [11] presented a Fully Connected Network (FCN) to analyze student's achievement in higher education. The data was collected initially from the OULAD database, and data orientation was performed to orient the data into a specific file format. Then, preprocessing methods were applied to improve the data quality. Finally, FCN was utilized for predicting student performance, which achieved a maximum of 84% accuracy, which was better than previous models. **Feiyue Qiu *et al.* [12]** suggested ML methodologies for student performance prediction in e-learning. The system performed data cleaning and standardization as the preprocessing steps, and then feature selection was carried out using variance filtering. The selected features were given to Naive Bayes (NB), K-Nearest Neighbor (KNN), and SoftMax for identifying the students' performance. The system was tested on the OULAD dataset and attained 97.40% accuracy. **Alberto Rivas *et al*. [13]** recommended using an artificial neural network (ANN) in virtual learning environments for a student's academic performance prediction system. The network was trained with the normalized data to identify the behavior patterns. After that, the original data set was examined using a real-world study, which comprised 120 students studying for a master's degree via a VLE. The objective was to raise the pass rate by reducing the factors that caused students to fail their tests. In the end, the pre-trained ANN was employed to forecast whether or not a particular student's parameter changes would enable them to pass their tests. The system outperformed the previous techniques by achieving 0.782% recall, 0.781% f-measure, and 0.782% precision.

Ghassen Ben Brahim *et al*. [14] proposed a student's academic performance system using ML techniques based on online engagement behaviors. When the data was first preprocessed, entries that showed any discrepancy were eliminated. Then, features such as timing statistics, activity type, and peripheral activity count were extracted. In the end, the student performance was predicted using Multilayer Perceptron (MLP), Logistic Regression (LR), NB, Random Forest (RF), and Support Vector Machine (SVM). The RF classifier performed better than others by achieving an accuracy of 97.4%. **Yutong Liu *et al.* [15]** proffered a student performance detection network using clickstream data and an ML model. The dataset was initially stripped of feature sets representing the number of clicks on 12 learning sites at weekly and monthly intervals. The model was then trained using ML techniques for student performance prediction, such as gradient boosting trees, LR, KNN, RF, one-dimensional convolutional neural networks, and long short-term memory. The system attained 90.25% accuracy, which was superior to the traditional approaches,

The works mentioned above provide prominent results but have the following limitations. Some researchers use ML-based approaches to predict the student's performance. They work well and provide satisfactory outcomes, but they are unsuitable for the time series data and could have performed better when the amount of data is significant. Some authors use DL-based approaches to overcome these issues, which handle the time series data more effectively and provide higher prediction performance. However, the imbalanced dataset is the central issue in most existing works, making the training model biased to one class. Also, the existing classifiers face the gradient vanishing issue due to the presence of traditional activation functions in their structure, including sigmoid and tanh, which will affect the prediction performance of the classifier. To overcome the above issues, this paper develops a novel DL approach, namely, EARNN model to predict the student's academic performance with KMSMOTE based data balancing technique, and this will improve the prediction rate of the system with minimal classification loss.

3. PROPOSED METHODOLOGY

Figure 1 shows the proposed methodologies' workflow. It mainly involves four phases: data collection, preprocessing, balancing, and classification. Initially, the student's data is collected from the OLAUD dataset, and then preprocessing operations such as missing value removal and numerical data conversion are done to get higher accurate results. Then, the dataset imbalance issue is solved by applying KMSMOTE. Then, the features such as demographic, assessment, and click stream of the students were extracted and given to the EARNN for student performance prediction, in which the vanishing gradient issue in conventional RNN was rectified by applying the ReLU activation function.

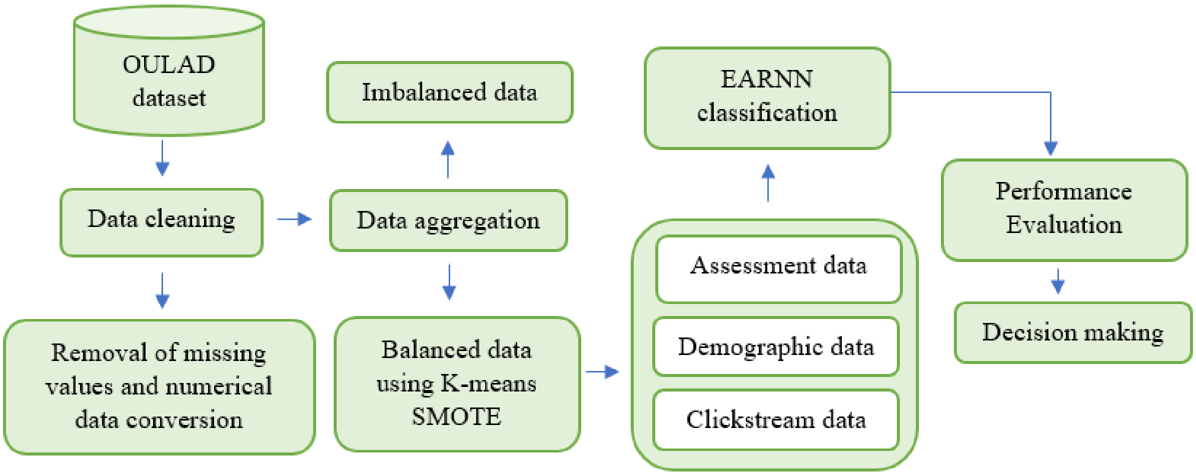


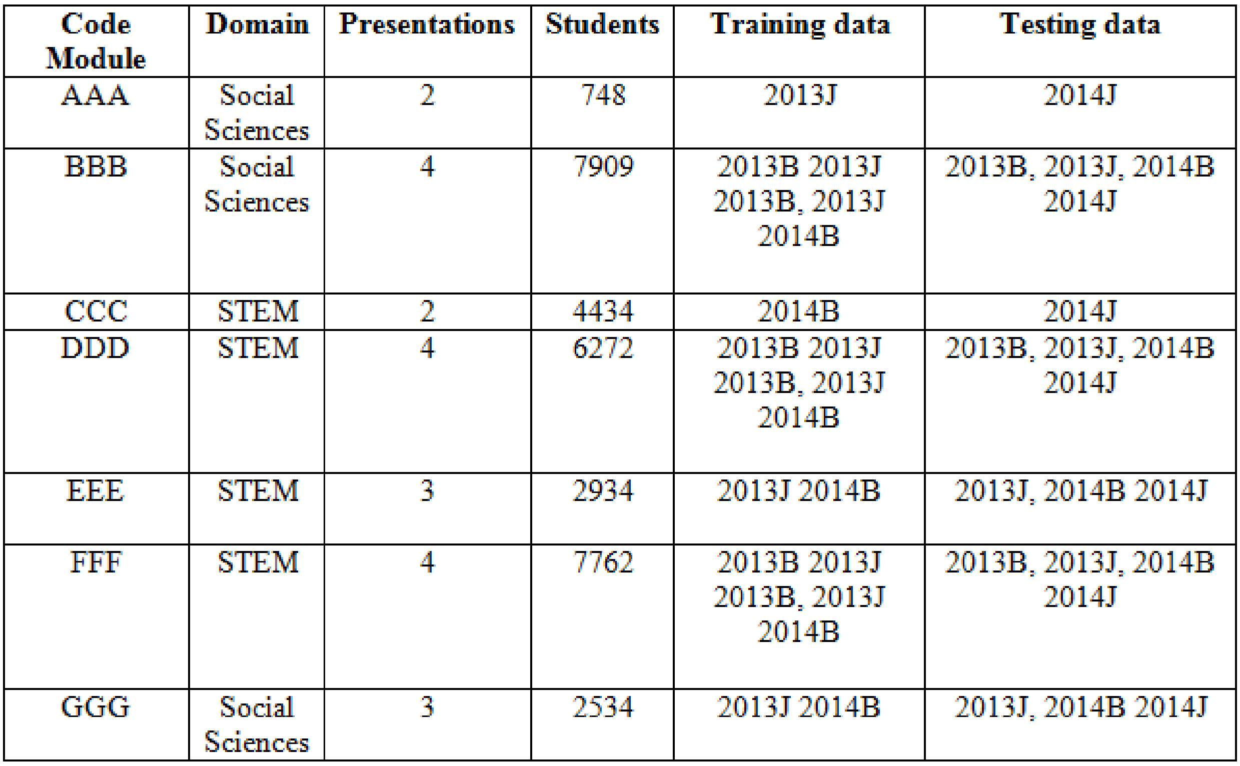
Figure 1: Workflow of the proposed methodology

3.1 Data Collection

At first, the OLAUD was used to gather the data, which is accessible via https://analyse.kmi.open.ac.uk/open_dataset. This dataset is part of the Open University Online Learning platform, which off-campus students employ to access course materials, participate in forum discussions, submit assignments, receive assessments, and more. It consists of the examination results, mutual information, course information, and demographic data of 32,593 respondents for a maximum of 9 months in 2014 and 2015. It consists of seven distinct courses, each introduced at least twice and beginning in a different month of the year. The student gets access to the main page and its subpages, can engage with various digital content, and can even take quizzes. The two categories of assessments are the tutor-marked assessment (TMA) and the computer-marked assessment (CMA). A student's course results can fall into one of the following four groups: distinction, pass, fail, or withdrawal. The student's age, gender, marital status, skills, computer literacy level, location, support, caregiver, nationality, ethnicity, daily commute, and other details are included in the demographic data derived from that. Information about the course covers topics such as high school, universities, training, pedagogy, formal and informal education, training mode, kind of material provided, and more.

Students' interactions with the VLE are indicated by mutual information, and these interactions are tracked in the daily click counts for each course—assessment details, including the quantity, type, and weight of assessments needed for every module. Every module typically consists of evaluations, followed by the final test. The time-series approach performs well when handling variable-length data because certain aspects of the course, including its length and the kind and quantity of assignments, change from semester to semester. Clickstream data is a type of time series data. The activities of the students are documented in the log file, and the clickstream data is used to determine the amount of time spent on each activity based on the students' clicks. The forum variable containing the students' conversation points to a location where students can upload questions and receive answers. This dataset is used in several ways to train and test a single course after the chosen course has ended, reducing the method's significance. Our suggested method can successfully train and validate the material from history courses and produce encouraging outcomes for the current course. Furthermore, the suggested effort created a novel model to determine participants' results according to their click stream, assessment stream, and demographics. It ignored the "withdrawal" instances and combined the "distinction" and "pass" labels into a single "pass" label. The summary of the collected dataset is demonstrated in Table 1

Table 1: Dataset Summary



Code Module	Domain	Presentations	Students	Training data	Testing data
AAA	Social Sciences	2	748	2013J	2014J
BBB	Social Sciences	4	7909	2013B 2013J 2013B, 2013J 2014B	2013B, 2013J, 2014B 2014J
CCC	STEM	2	4434	2014B	2014J
DDD	STEM	4	6272	2013B 2013J 2013B, 2013J 2014B	2013B, 2013J, 2014B 2014J
EEE	STEM	3	2934	2013J 2014B	2013J, 2014B 2014J
FFF	STEM	4	7762	2013B 2013J 2013B, 2013J 2014B	2013B, 2013J, 2014B 2014J
GGG	Social Sciences	3	2534	2013J 2014B	2013J, 2014B 2014J

3.2 Data Preprocessing

Given the nature of OULAD, preprocessing the dataset is a highly crucial task. Herein, the data preprocessing includes the removal of missing values and conversion from the input dataset. These are explained as follows:

Step 1: Missing values removal

At first, the suggested method included each student's incomplete evaluation of the assessment table and gave it a score of zero. Additionally, the VLE data was arranged weekly, indicating the total number of clicks for every category of VLE activity within a given week. Even though some students couldn't use the VLE for a few weeks, the system added missing weeks' worth of interaction data and gave them a zero.

Step 2: Numerical data conversion

After that, numerical data conversion is performed on the collected dataset. Typically, a single student's demographic data, such as their gender, most significant level of education, and place of residence, is included in the OULAD dataset. The suggested method transformed student data into one-hot encodings when most of these attributes are unordered. We employ one hot encoding method to express category variables as numerical values.

3.3 Dataset Balancing

Training a model is facilitated by balancing the dataset since it keeps the model from becoming biased in favor of one class. This work applies the KMSMOTE technique to handle the dataset imbalance problem. The oversampling method known as SMOTE creates synthetic samples for the minority class. Beginning with a random minority class sample, synthetic instances are added throughout the line segment that links the sample to the nearest random neighbor of the minority class. However, because the minority class sample was randomly chosen, it has drawbacks such as higher computational complexity and decreased interpretability. In order to address these issues, this study employs a K-means (KM) algorithm to minimize imbalances between and within classes by selecting minority classes as optimally as possible while preventing the creation of noisy samples. Hence, it is known as KMSMOTE. Three steps make up KMSMOTE: oversampling, filtering, and clustering. The input space is partitioned into k groups in the clustering step using k-means clustering. After the filtering stage, clusters with a high percentage of minority class samples are kept for oversampling. After that, the number of synthetic samples to

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

be produced is divided, and more samples are added to clusters with sparse minority sample distributions. In each cluster selected during the oversampling phase, SMOTE is ultimately implemented to achieve the required ratio of minority to majority occurrences.

3.4 Classification

Finally, the classification is performed using EARNN. The connections between the computational units in an RNN neural network class form a directed circle. RNNs, in contrast to feed forward networks, are capable of processing any order of inputs using their internal memory. An RNN's computing units each have a configurable weight and an actual valued activation that varies over time. The exact weights are applied recursively over a graph-like structure to construct RNNs. Every hidden layer's output is subjected to a tanh activation function, with the outcome being passed on to hidden layers that come after. The RNN model is an effective tool for handling data sequences by assuming that the current time step depends on earlier time steps. However, it has certain drawbacks, such as the vanishing gradient problem, which can make the network less able to weigh long-term dependencies. Therefore, the suggested system solves the gradient vanishing and exploding problems using the ReLU activation function rather than the tanh activation. This improvisation in conventional RNN is termed EARNN. The structure of the EARNN is shown in Figure 2.

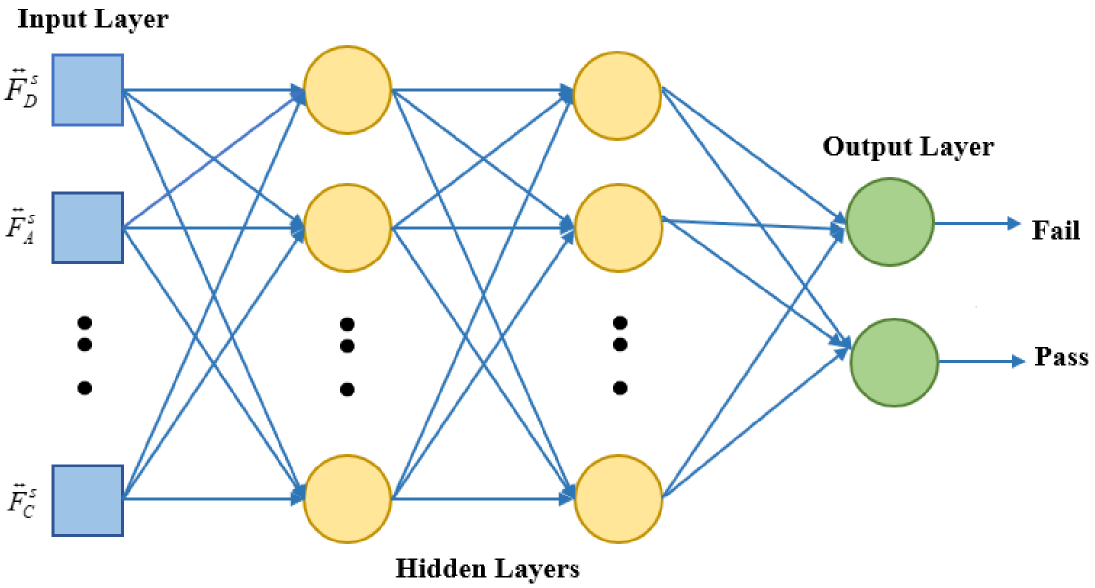

Input Layer
$\overleftrightarrow{F}_D^s$
$\overleftrightarrow{F}_A^s$
$\overleftrightarrow{F}_C^s$
Output Layer
Fail
Pass
Hidden Layers

Figure 2: Structure of the EARNN

The input layer, hidden layer, and output layer are the three layers that make up the suggested EARNN model. These are explained as follows:

Step 1: Input layer

In order to determine participant outcomes based on their clickstream, assessment stream, and demographics, the proposed study created a new deep network. The balanced preprocessed dataset takes some features as inputs, such as the demographics, assessment stream, and click stream. Specifically, for a student s, his/her balanced preprocessed demographics, week-wise interaction stream information, and assessment-wise assessment stream information are denoted as $\widetilde{D}_f$, $\widetilde{C}_f$ and $\widetilde{A}_f$. The demographic information of the student $\widetilde{D}_f$ is converted into a demographical feature vector $\overleftrightarrow{F}_D^s$ in the demographics' module. The week-wise (click stream) features $\overleftrightarrow{F}_C^s$ and the assessment-wise features $\overleftrightarrow{F}_A^s$ are extracted using the click stream and assessment stream. Then the system fuses $\overleftrightarrow{F}_D^s$, $\overleftrightarrow{F}_A^s$, and $\overleftrightarrow{F}_C^s$ to make the final feature vector $\overleftrightarrow{F}_s$ that indicates the student's historical features. These features are given as an input to the RNN.

Step 2: Hidden layer

After initializing the input, it is passed to the hidden layer. The total sum of the input node values multiplied by their given weights determines the hidden layer node values. It is expressed as follows:

$$\overset{\boxtimes}{h}_t = \mu^*\left(\overset{\boxtimes}{W}_{F_s h}\overset{\boxtimes}{F}_{s,t} + \overset{\boxtimes}{W}_{hh}\overset{\boxtimes}{h}_{t-1} + \overset{\boxtimes}{B}_h\right) \quad (1)$$

Where, $\overset{\boxtimes}{h}_t$ indicates a hidden state at time t and acts as “memory” of the network, $\overset{\boxtimes}{W}$ refers to the input feature vectors’ $\left(\overset{\boxtimes}{F}_s\right)$ weight value, $\overset{\boxtimes}{B}_h$ indicates the hidden units’ bias vector, and μ^* denotes the ReLU activation function. ReLU is a piecewise linear or non-linear function that, in the case of a positive input, will output zero; otherwise, it will output the input directly. It is expressed as follows:

$$\mu^* = \max\left(0, \overset{\boxtimes}{F}_s\right) \quad (2)$$

Step 3: Output layer

Finally, the output of the RNN is computed as follows:

$$\overline{\overline{OR}}_t = \overset{\boxtimes}{W}_{h\overline{\overline{OR}}}\overset{\boxtimes}{h}_t + \overset{\boxtimes}{B}_{\overline{\overline{OR}}} \quad (3)$$

Where, $\overset{\boxtimes}{B}_{\overline{\overline{OR}}}$ indicates the bias vector of the output layer and $\overline{\overline{OR}}_t$ indicates the network’s output and it produces the binary classification output, 0 is allocated to not pass (fail) and 1 is allocated to pass.

4. RESULTS AND DISCUSSION

In this section, the outcomes of the proposed student’s performance prediction in the E-learning platform are evaluated using existing methods regarding some performance measures. The proposed study is implemented in PYTHON with Windows 10 OS. The outcomes of the proposed EARNN are investigated against the conventional RNN, MLP, ANN, and RF in terms of accuracy, precision, and recall. Data from 5 to 39 weeks is used to create the test; models can predict performance more accurately with more weeks. Due to the increased quantity of input data, each model’s predictive power has dramatically improved as the number of weeks has increased.

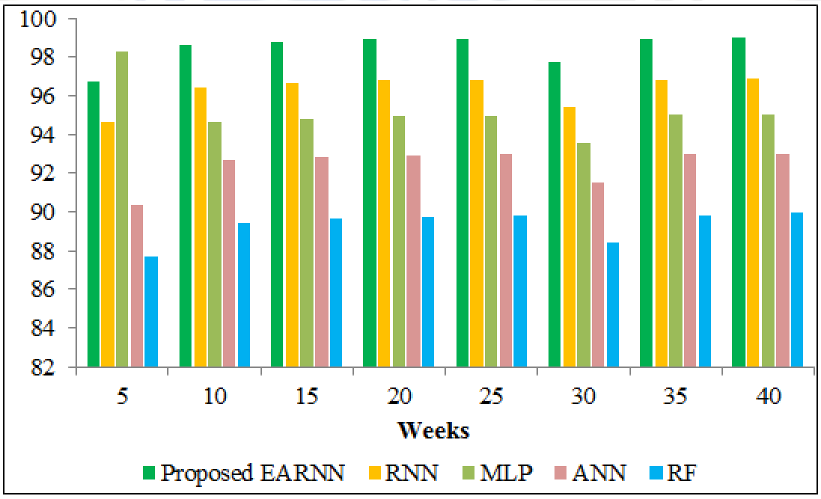


Figure 3: Accuracy analysis

Figure 3 indicates the accuracy of the classifiers for academic performance prediction. Accuracy is the most popular and widely used quality evaluation criterion for predicting system results. The proposed one achieves 96.72% accuracy for week 5, which is higher than the existing methods because the existing RNN, MLP, ANN, and RF offer accuracy of 94.67%, 98.28%, 90.31%, and 87.71% for the same week 5. Similarly, for the remaining weeks (10 to 40), the proposed one achieves maximum accuracy of 98.64%, 98.78%, 98.91%, 98.93%, 97.76%, 98.93%, and 98.99%, which are also better than the existing methods. As a result, the data clearly show that when student engagement and score information improve, the classifier’s performance improves.

Table 1:Results analysis of the proposed model

Metrics	**Weeks**	**Proposed EARNN**	**RNN**	**MLP**	**ANN**	**RF**
Precision	5	96.83	94.72	98.32	90.43	87.82
	10	98.72	96.53	94.75	92.71	89.56
	15	98.81	96.74	94.89	92.91	89.74
	20	98.97	96.85	94.99	92.99	89.82
	25	98.99	96.89	95.03	93.01	89.79
	30	97.86	95.53	93.67	91.52	88.56
	35	98.99	96.92	95.03	93.03	89.96
	40	99.02	96.93	95.12	93.12	89.99
Recall	5	96.64	94.57	98.18	90.21	87.61
	10	98.53	96.32	94.57	92.57	89.32
	15	98.65	96.55	94.72	92.74	89.57
	20	98.85	96.68	94.83	92.84	89.63
	25	98.83	96.72	94.85	92.87	89.68
	30	97.66	95.32	93.46	91.38	88.32
	35	98.84	96.74	94.89	92.89	89.72
	40	98.89	96.78	94.98	92.97	89.85

Table 1 demonstrates the outcomes of the methods regarding precision and recall metrics for weeks 5 to 40. Metrics like precision and recall are also commonly employed to assess the predictive model. Precision in this task refers to the percentage of students who accurately identify as failing out of those classified as failing by the model. Simultaneously, the recall shows the proportion of failed test participants' at-risk samples that our model predicted. The proposed work achieves higher precision and recall than the existing methods. For example, the proposed one achieves maximum high precision and recall of 99.02% and 98.89% for week 40, which is better than the existing methods. The proposed one initially preprocesses the dataset, and KMSMOTE efficiently handles the unbalancing issues of the dataset. In addition, the prediction was enhanced with the ReLU activation function. These improvisations increase the system's performance. Thus, the analysis demonstrated that the system achieved outstanding outcomes over all four existing models.

5. CONCLUSION

The study proposes a student's performance prediction system for an E-learning platform using a novel deep learning model. The proposed work is divided into four sections: data collection, preprocessing, balancing, and classification. The framework uses the OULAD dataset to validate the system's effectiveness. The outcomes of the proposed work are compared to the existing RNN, MLP, ANN, and RF methods with the help of accuracy, precision, and recall metrics for weeks 5 to 40. The experimental outcomes proved that the proposed system is superior to other models. For example, the proposed one achieved 96.72% accuracy, 96.83% precision, and 96.64% recall for week 5, which was better than the existing methods. Likewise, the proposed one achieves better outcomes for the rest of the weeks (10-40). Stated differently, compared to existing models, the suggested EARNN model can forecast the student's ultimate performance earlier and with greater accuracy.

COMPLIANCE WITH ETHICAL STANDARDS

Conflict of Interest: I, Thilagavathi.S, declares no conflicts of Interest to disclose.

Ethical Approval: This article does not contain any studies with human participants or animals performed by any of the authors.

REFERENCES

[1] Esteban, A., Romero, C., & Zafra, A. (2021). Assignments as influential factors to improve the prediction of student performance in online courses. *Applied Sciences*, *11*(21), 10145.

[2] Hatahet, T., Mohamed, A. A. R., Malekigorji, M., & Kerry, E. K. (2021). Remote learning in transnational education: Relationship between virtual learning

engagement and student academic performance in BSc Pharmaceutical Biotechnology. *Pharmacy*, *10*(1), 4.

[3] Lacka, E., Wong, T. C., & Haddoud, M. Y. (2021). Can digital technologies improve students' efficiency? Exploring the role of Virtual Learning Environment and Social Media use in Higher Education. *Computers & Education*, *163*, 104099.

[4] Caprara, L., & Caprara, C. (2022). Effects of virtual learning environments: A scoping review of literature. *Education and information technologies*, 1-40.

[5] Hidalgo, A. C., Ger, P. M., & Valentin, L. D. L. F. (2022). Using Meta-Learning to predict student performance in virtual learning environments. *Applied Intelligence*, 1-14.

[6] Alnassar, F., Blackwell, T., Homayounvala, E., & Yee-king, M. (2021, April). How well a student performed? a machine learning approach to classify students' performance in a virtual learning environment.

[7] Elfakki, A. O., Sghaier, S., & Alotaibi, A. A. (2023). An Intelligent Tool Based on Fuzzy Logic and a 3D Virtual Learning Environment for Disabled Student Academic Performance Assessment. *Applied Sciences*, *13*(8), 4865.

[8] Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., ... & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *Ieee Access*, *9*, 7519-7539.

[9] Nazempour, R., & Darabi, H. (2023). Personalized Learning in Virtual Learning Environments Using Students' Behavior Analysis. *Education Sciences*, *13*(5), 457.

[10] Albreiki, B., Zaki, N., &Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. *Education Sciences*, *11*(9), 552.

[11] Hooda, M., Rana, C., Dahiya, O., Shet, J. P., & Singh, B. K. (2022). Integrating LA and EDM for improving students Success in higher Education using FCN algorithm. *Mathematical Problems in Engineering*, *2022*.

[12] Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., ... & Chen, P. K. (2022). Predicting students' performance in e-learning using learning process and behavior data. *Scientific Reports*, *12*(1), 453.

[13] Rivas, A., Gonzalez-Briones, A., Hernandez, G., Prieto, J., & Chamoso, P. (2021). Artificial neural network analysis of the academic performance of students in virtual learning environments. *Neurocomputing*, *423*, 713-720.

[14] Brahim, G. B. (2022). Predicting student performance from online engagement activities using novel statistical features. *Arabian Journal for Science and Engineering*, *47*(8), 10225-10243.

[15] Liu, Y., Fan, S., Xu, S., Sajjanhar, A., Yeom, S., & Wei, Y. (2022). Predicting Student performance using clickstream data and machine learning. *Education Sciences*, *13*(1), 17.