Enhancing Student Performance Prediction through LSTM-based Deep Learning Models with Unbalanced Data Handling using Oversampling Approach

Edi Ismanto[1], Hadhrami Ab Ghani[2], Nor Hidayati Binti Abdul Aziz[3+], Nurul Izrin Md Saleh[4], Noverta Effendy[5]

[1,5] Department of Informatics, Universitas Muhammadiyah Riau, Pekanbaru, Indonesia
[2] Faculty of Data Science, Universiti Malaysia Kelantan, Kota Bharu, Malaysia
[3+] Faculty of Engineering and Technology, Multimedia Universiti, Melaka, Malaysia
[4] Faculty of Information Technology and Communication, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
edi.ismanto@umri.ac.id

Abstract. Accurate prediction of student performance is crucial in learning analytics to prevent course failures and improve academic outcomes. However, publicly accessible educational data often contains noise and imbalanced data distributions, requiring effective handling techniques. In this study, we propose a novel approach that combines the Synthetic Minority Over-sampling Technique (SMOTE) with Long Short-Term Memory (LSTM) and Feed-Forward Neural Network (FFNN) models for performance prediction in virtual learning environments (VLEs). Our experimental results show that utilizing the SMOTE technique significantly improves the accuracy of predicting student withdrawals, with the LSTM model achieving the highest accuracy of 94.90% in the 25th week of data testing. These findings indicate the effectiveness of the SMOTE technique in addressing data imbalance issues in VLE datasets and the potential of our proposed deep learning models in accurately predicting student performance. The implications of our study are significant for learning analytics and educational institutions, as accurate prediction of student performance can inform early interventions and personalized support. Future research could explore the generalizability of our approach in diverse educational contexts and the integration of additional features for further improving prediction accuracy. Hence, our study contributes to the field of learning analytics by proposing a novel approach that combines SMOTE with deep learning models for student performance prediction in VLEs. Our findings highlight the potential of our approach in addressing data imbalance challenges and accurately predicting student performance, with implications for enhancing student success in educational settings.

Keywords: Imbalanced Data, Deep Learning, Predictive Modelling, Student Performance

M. I. Zainudin and H. Rahmat (eds.), *Proceedings of the 4th International Conference on Communication, Language, Education and Social Sciences (CLESS 2023)*, Advances in Social Science, Education and Humanities Research 819, https://doi.org/10.2991/978-2-38476-196-8_18

1 Introduction

Student performance prediction is currently regarded as one of the biggest problems facing education institutions due to a lack of trustworthy models. The lack of research on the proper metrics to use in evaluating student performance and the incompatibility of the current models with institutional frameworks make this issue very important to study. Predicting student performance is one effort in anticipating student failures in taking courses [1], [2]. Numerous publicly available data in the area of education can be examined with the goal of improving student academic performance. The public data still has a lot of noise that needs to be removed before it can be used. Additionally, data imbalances must be corrected in order for the prediction results to match the desired outcome [3]–[5]. This happens when one of the classes is underrepresented throughout the entire dataset. The dataset must be balanced because an imbalance can result in instances being incorrectly classified during the prediction phase [6], [7]. The equal distribution of instances for each class is taken for granted when applying prediction to educational datasets.

Unbalanced data can be handled in a variety of ways, including data level (preprocessing) and algorithm-based methods [8], [9]. The Synthetic Minority Over Sampling Technique (SMOTE), which creates synthetic samples between minority samples and their neighbors, is one of the most well-known preprocessing techniques [10]–[12]. The first goal of this study is to use the SMOTE technique to address the issue of data imbalance in a virtual learning environment (VLE). The second objective is to use deep learning techniques to forecast which students will complete a course or drop it. Afterward, the model that is being tested is then evaluated.

2 Related Works

The SMOTE method, which was used in research by [12], generates synthetic data for predictive models with the highest accuracy and robustness. According to the study's [13] results, AP SMOTE had the highest yield when SMOTE and AP SMOTE were applied to an unbalanced data set. The study's findings showed that class data when students graduated that weren't balanced could be classified with greater accuracy, precision, and sensitivity when the SMOTE method was used [4], [11], [14], [15].

The SMOTE-based random forest algorithm has the highest accuracy, according to the study's findings [16]. Research [17] used the SMOTE preprocessing method to classify student academic achievement and compared the performance of two classifiers, C4.5 and k-Nearest Neighbor (KNN). When it comes to accuracy, recall, and precision values, the C4.5 Decision Tree method outperforms other prediction techniques.

According to research [18]–[21], deep learning is effective at resolving prediction issues. The numerical results it was concluded that the LSTM technique outperforms state-of-the-art methods in terms of achieving higher precision, recall, f-measure, and less time consumption [22]. The problem of spatiotemporal data correlation can be successfully resolved by LSTM, which improves prediction outcomes [23].

The main hypothesis of the study is that the SMOTE algorithm can solve the problem of unbalanced VLE data. The second hypothesis is whether the deep learning approach is superior to other approaches for resolving issues predicting student performance.

3 Research Method

The study's methodology for putting the suggested system into practice will be covered in this section. The steps of the proposed methodology are shown in Fig. 1. This section begins with a description of the data sets that were used in the study. Utilizing readily available data from a public institution, preprocessing, classification, and evaluation procedures were carried out. The sections below go over the specifics of each procedure.

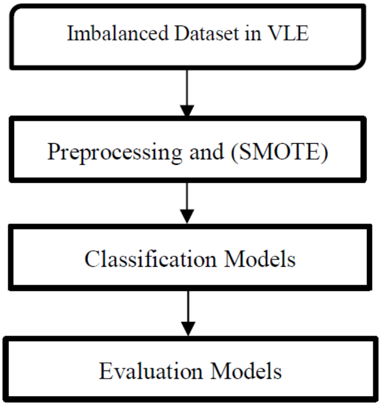


Fig. 1. The proposed research method and algorithm

3.1 Data Preprocessing

This study makes use of the OULAD Student dataset, which records demographic information about students as well as user behavior on the Virtual Learning Platform. OULAD dataset contains data about courses, students, and their interactions with Virtual Learning Environment (VLE) for seven selected courses (called modules) [24]. Course presentations begin in February and October, and they are denoted by the letters "B" and "J," respectively. The dataset consists of tables connected by distinctive identifiers. A data frame with 12 variables and 32593 rows. Fig. 2 displays the quantity of virtual Learning Environment (VLE) pages that students frequently access. The top seven items with the most clicks are the homepage, subpage, resource, content, forums, URL, and quiz.

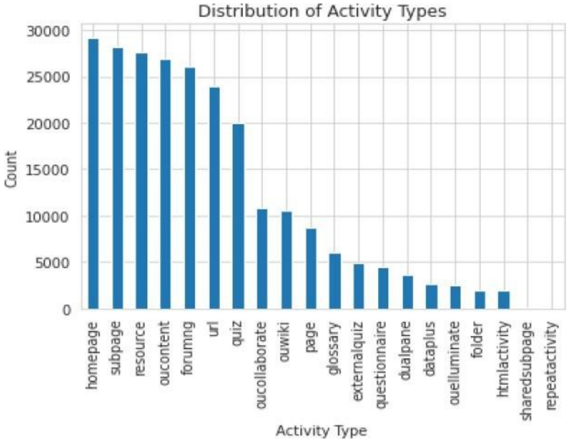
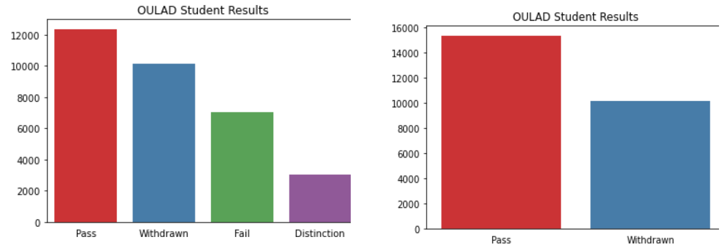


Fig. 2. Pages that students frequently access.

When it comes to problems with data quality, NaN entries show that the module presentation did not include these activities, which skews the data results due to their unavailability. Since not all activity types were present in all of the modules, we have chosen to handle this by inputting these NaNs as zero. Another issue with data quality is the fact that different modules and module presentations have different average activity lengths (even for activities of the same type). For instance, an 11-question quiz in one module presentation might require fewer clicks than a 20-question quiz, because of its shorter length. Because of these factors, the number of clicks that the student made may not be the best metric for measuring how well they interacted with the ma- terial.

This course has four different possible results: Pass, Withdraw, Fail, and Distinction. Hence, records from three categories are considered (Pass, Withdraw, and Distinction) as displayed in Fig. 3 (a). The Pass and Distinction will be combined as the Pass category as shown in Fig. 3 (b). This issue is simplified to a binary classification in order to predict whether or not students will drop the course



(a) (b)

Fig 3. (a) Results of student courses and (b) Two prediction targets

Finding the students who can withdraw from a course during the first few weeks of its run is crucial. Consequently, it is essential to conduct the student analysis on a weekly basis. The reported date of material access needs to be mapped to the appropriate weeks.

3.2 SMOTE Method

The Synthetic Minority Oversampling Technique (SMOTE) algorithm is used to deal with the dataset's unbalanced data. This approach is based on rebalancing the dataset during initial training by combining over- and under-sampling [4-5],[16],[24]. The SMOTE method locates the k nearest neighbors of a class with small data regarding a specific set of small data from the same class, draws a straight line with the neighbor, and generates points until the random points have a balanced ratio [32]. SMOTE aims to increase the proportion of minority classes in the distribution of classes by synthesizing data for oversampling purposes [9]. Eq. (1), is used to generate fresh data for the minority class.

$$y' = y^i + (y^j - y^i) * \gamma \qquad (1)$$

y': is used to store the outcome of the new data. y^i : stands in for the minority class.
y^j: is a randomly chosen value from the minority class's k-nearest neighbors y^i, and
γ : is a value chosen at random from a random vector with a 0–1 range. SMOTE creates fresh synthesis training data for the minority class using linear interpolation. The training data for the synthesis is generated by randomly choosing one or more of the k-nearest neighbors for each sample in the minority class.

3.3 The Deep Learning Model

Neural networks (NN) are classification, regression, and clustering algorithms that were inspired by the human brain. Through the use of parallel processing, NN can be used to resolve difficult and ambiguous issues [26]. A NN's node composition varies depending on the algorithm being used, such as feed-forward and reverse (sequential or convolution) [2],[34]. We use the Feed-Forward NN and LSTM models in the DL model we proposed.

A feed-forward neural network algorithm is also known as a multilayer perceptron. The architecture of a feedforward neural network consists of three layers: input, hidden, and output [29]. According to the research done [15], the MLP classifier with the SMOTE technique performs better than the ML algorithm used. A significant improvement in accuracy can be achieved by basing the model's development on the input of particular variables [8],[10],[18],[20]. LSTM is a subtype of recurrent neural networks (RNN) [33]. In particular, the LSTM can be used to extract temporal patterns from nonlinear time-series data [6],[14],[19],[21],[25],[31]. Due to its superior time series data processing performance, it has been extensively used in many different fields [12],[30]. The memory module of the LSTM recurrent neural network is composed of three multiplication units: the input gate, forget gate, and the output gate. These gates regulate information input, update, and output in turn, giving the network a specific memory function.

3.4 Evaluation Method

We employ cross-validations, also referred to as k-folds, as one of the validation techniques when performing model validation [9],[32]. At this stage, the determination test is also used, and the confusion matrix is used to compute accuracy, recall, and precision [1],[4]. The performance of the Confusion Matrix can be assessed using Research Method

$$Accuracy = \frac{}{TP+T} \tag{2}$$

$$N$$

$$TP+TN+F \quad P+FN \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$F1 - Score = \frac{}{}$$

$$2*Precision*Recall \quad Precision+Recall$$

4 Results and Discussion

In this study, experiments were conducted using the sci-kit-learn library and the Python programming language. The dataset was tested using the Logistic Regression (LR) model on demographic data, which resulted in the creation of the fundamental model. In the training set, there is an imbalance of data: 7605 instances were drawn out of 11550 instances of pass. Predicting the maximum number of students who might withdraw from a course is the main goal of this study. The confusion matrix below demonstrates that 1019 withdrawn instances were correctly predicted by the model. Correct predictions that were withdrawn should increase in order to enhance model performance. The confusion matrix results using the LR are shown in Table 1.

Table 1. The confusion matrix results using the LR

Variable	Actual Withdrawn	Actual Pass
Predicted Withdrawn	1019	1532
Predicted Pass	624	3211

The LR results shown in Table 1 seem to be biased towards the Pass class. The training dataset is unbalanced and has more data points related to the pass category, as can be seen in Table 2. By oversampling the instances for the withdraw category, the SMOTE method is used to solve this issue. The confusion matrix measurement shows an increase in withdrawal cases correctly classified into 1504 records. Then, using the Feed-Forward Neural Network (FFNN) and Long Short-Term Memory (LSTM) models of deep learning, we test the models.

Table 2. The confusion matrix results using the LR and SMOTE methods.

Variable	Actual Withdrawn	Actual Pass
Predicted Withdrawn	1504	1047
Predicted Pass	1248	2587

4.1 Analysis Feeds Forward NN model with SMOTE

The FFNN model was then put to the test using the SMOTE approach. Utilize weekly click stream data from weeks 5, 10, 15, 20, and 25 to evaluate the model. Through oversampling, the SMOTE method addresses problems with data imbalance. The measurement outcomes of the FFNN model using SMOTE for weekly click count data are shown in Table 3.

Table 3. Model evaluation using the typical weekly clickstream.

		Not utilizing cross-validation		10-fold cross-validation	
Models	Weeks	Accuracy	F-Score	Accuracy	F-Score
Feed Forward NN + SMOTE	05	74.76	0.80	75.78	0.76
	10	80.45	0.85	79.98	0.81
	15	84.62	0.88	85.42	0.86
	20	90.10	0.92	89.90	0.90
	25	92.94	0.94	92.92	0.93
Model evaluation using clickstreams and demographic data with materials					
		Not utilizing cross-validation		10-fold cross-validation	
Models	Weeks	Accuracy	F-Score	Accuracy	F-Score
Feed Forward NN + SMOTE	05	77.11	0.84	77.00	0.84
	10	78.24	0.84	79.83	0.86
	15	81.75	0.87	81.32	0.87

	20	84.37	0.89	83.31	0.88
	25	86.14	0.90	85.53	0.89
Model evaluation using clickstreams aggregated by VLE Materials					
		Not utilizing cross-validation		10-fold cross-validation	
Models	Weeks	Accuracy	F-Score	Accuracy	F-Score
Feed Forward NN + SMOTE	05	73.16	0.79	74.86	0.75
	10	77.05	0.82	77.07	0.77
	15	79.23	0.84	79.23	0.79
	20	82.93	0.87	81.75	0.82
	25	84.79	0.88	84.58	0.85

The confusion matrix measurement shows an increase in withdrawal cases correctly classified into 1194 records. Table 4 displays the measurements' findings. It has been demonstrated that the FFNN model can correct for unbalanced dataset conditions using the SMOTE method.

Table 4. Increase in withdrawal cases correctly classified

Variable	Actual Withdrawn	Actual Pass
Predicted Withdrawn	1194	780
Predicted Pass	603	3061

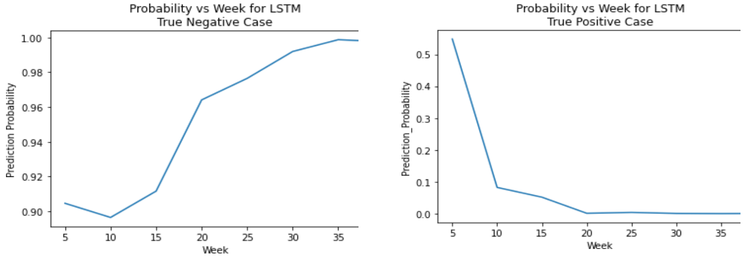
4.2 Analysis of LSTM model with SMOTE

The LSTM model needs training data in a specific format (samples, time steps, and features). The time steps are kept small because this research's primary goal is the early detection of dropouts (5,10,15 weeks). The input to the LSTM model consists of the click streams organized in a multidimensional format. The features will be based on the clickstreams from 20 different VLE materials, and data for each student is kept up to date for the entire 38-week period. Thus, the 3D data expected by LSTM shows that for all students, we have 38 weeks and 20 resources. The combination of the student id, course, and presentation is used to create a unique id that is used to keep track of each course and presentation a student completes. The most recent data is organized into a multidimensional format using the variables (student*weeks*resources). Table 5 displays the findings from measuring LSTMs using data weekly click streams.

Table 5. Model evaluation using the weekly clickstream for LSTM

Models	Weeks	Accuracy	F-Score
Long short-term memory network (LSTM) + SMOTE	05	77.10	0.85
	10	83.47	0.89
	15	87.29	0.91
	20	91.78	0.94
	25	94.90	0.96

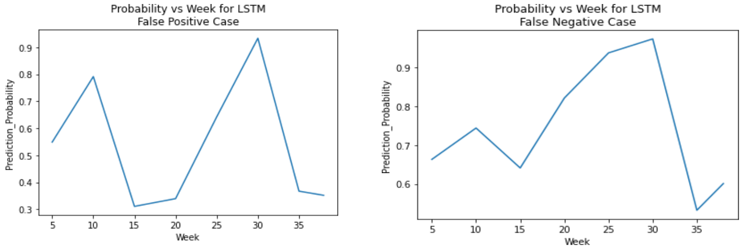
Probability of model prediction over weeks. To comprehend how a student instance is classified, the probabilities generated by the model over the course of the weeks are analyzed. True Negative (TN) case illustration: How the model correctly foresees the passing student. The probabilities and weeks for LSTM True Negative (TN) cases are shown in Figure 4. True Positive (TP) case study: how the model correctly predicted the student who withdrew. There have been noticeable changes to the True Positive (TP) graph, as can be seen.



(a) (b)

Fig 4. Probability vs week for LSTM (a)True Negative (TN) and (b)True Positive (TP) cases

We track the False Positives (FP) and False Negatives (FN) to see how the predicted probability changes over the course of the weeks and how the model decision is affected. The results of adjusting the probabilities for false positives and false negatives are depicted in figure 5.



(a) (b)

Fig 5. Probability vs week for LSTM (a) False Positive (FP) and (b)False Negative (FN) cases

5 Conclusion

The training dataset is imbalanced and contains more information about the pass category. By oversampling the instances for the withdraw category, the SMOTE method. helps to resolve this problem. The dataset was tested using Long Short-Term Memory (LSTM) and Feed-Forward Neural Network (FFNN) models. In the first experiment, the model was tested without using the SMOTE method. And in the second experiment, the model was tested using the SMOTE methodology. Evaluation of the model using data clickstreams, and demographics data combined with materials in VLE. A confusion matrix is used to measure the evaluation of model predictions, and the results are compared. When comparing using the SMOTE method to not using the SMOTE method, the confusion matrix showed an increase in the number of withdrawal cases that were correctly classified. Based on our model testing, we draw the conclusion that the SMOTE method can effectively address the issue of data imbalance. For upcoming research, it is essential to try to include extra features, like the capacity to predict students' graduation and failure rates.

References

1. Anggrawan, A., Hairani, H., & Satria, C. Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE. *International Journal of Information and Education Technology*, 1-7 (2022).
2. Aslam, N., Khan, I. U., Alamri, L. H., & Almuslim, R. S. An Improved Early Student's Performance Prediction Using Deep Learning. *International Journal of Emerging Technologies in Learning*, 108-122 (2021).
3. Bujang, S. D., Selamat, A., Ibrahim, R., Krejcar, O., & Herrera-Viedma, E. Multiclass Prediction Model for Student Grade Prediction Using Machine Learning. *IEEE Access*, 95608-95621 (2021).
4. Byeon, H. Predicting the Depression of the South Korean Elderly using SMOTE and an Imbalanced Binary Dataset. *International Journal of Advanced Computer Science and Applications*, 74-79 (2021).
5. Casuat, C. D. Predicting Students' Employability using Support Vector Machine: A SMOTE-Optimized Machine Learning System. *International Journal of Emerging Trends in Engineering Research*, 2101-2106 (2020).
6. Chen, H. C., Prasetyo, E., Tseng, S. S., Putra, K. T., Prayitno, Kusumawardani, S. S., & Weng, C. E. Week-Wise Student Performance Early Prediction in Virtual Learning Environment Using a Deep Explainable Artificial Intelligence. *Applied Sciences (Switzerland)*, 1-16 (2022).
7. Conijn, R., Van den Beemt, A., & Cuijpers, P. Predicting student performance in a blended MOOC. *Journal of Computer Assisted Learning*, 615-628 (2018).
8. Durga, V. S., & Jeyaprakash, T. Predicting Academic Performance of Deaf Students Using Feed Forward Neural Network and An Improved PSO Algorithm. *Webology*, 112-126 (2021).
9. Flores, V., Heras, S., & Julian, V. Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education. *Electronics (Switzerland)*, 457 (2022).
10. Gupta, S. L., & Mishra, N. Artificial Intelligence and Deep Learning-Based Information Retrieval Framework for Assessing Student Performance. *International Journal of Information Retrieval Research*, 1-27 (2021).
11. Hamoud, A. K., Kamel, M. B., Gaafar, A. S., Alasady, A. S., & Humadi, A. M. A prediction model based machine learning algorithms with feature selection approaches over imbalanced dataset. *Indonesian Journal of Electrical Engineering and Computer Science*, 1105-1116 (2022).
12. Hu, J., Wang, X., Zhang, Y., Zhang, D., Zhang, M., & Xue, J. Time Series Prediction Method Based on Variant LSTM Recurrent Neural Network. *Neural Processing Letters*, 1485-1500 (2020).
13. Jha, N. I., Ghergulescu, I., & Moldovan, A. N. OULAD MOOC dropout and result prediction using ensemble, deep learning and regression techniques. *CSEDU 2019 - Proceedings of the 11th International Conference on Computer Supported Education*, 154- 164 (2019).
14. Jo, Y., Maki, K., & Tomar, G. Time Series Analysis of Clickstream Logs from Online Courses. *arXiv*, 1-13 (2018).
15. Kehinde, A. J., Adeniyi, A. E., Ogundokun, R. O., Gupta, H., & Misra, S. Prediction of Students' Performance with Artificial Neural Network Using Demographic Traits. *Lecture Notes in Electrical Engineering*, 613-624 (2022).
16. Lanie, B. L. Affinity Propagation SMOTE approach for Imbalanced dataset used in Predicting Student at Risk of Low Performance. *International Journal of Advanced Trends in Computer Science and Engineering*, 5066-5070 (2020).
17. Lee, S., & Chung, J. Y. The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences (Switzerland)*, 3039 (2019).

18. Li, S., & Liu, T. Performance Prediction for Higher Education Students Using Deep Learning. *Hindawi Complexity*, 1-10 (2021).
19. Liu, J., Yin, C., Li, Y., Sun, H., & Zhou, H. Deep Learning and Collaborative Filtering-Based Methods for Students' Performance Prediction and Course Recommendation. *Hindawi Wireless Communications and Mobile Computing*, 1-13 (2021).
20. Madeira, B. C., Tasci, T., & Celebi, N. Prediction of Student Performance Using Rough Set Theory And Backpropagation Neural Networks. *European Scientific Journal ESJ*, 1-15 (2021).
21. Magalhaes, E. B., Santos, G. A., Molina, F. C., Da Costa, J. P., De Mendonca, F. L., & De Sousa, R. T. Student Dropout Prediction in MOOC using Machine Learning Algorithms. *2021 Workshop on Communication Networks and Power Systems, WCNPS 2021*, 15-22 (2021).
22. Mohseni, Z., Martins, R. M., Milrad, M., & Masiello, I. Improving classification in imbalanced educational datasets using over-sampling. *ICCE 2020 - 28th International Conference on Computers in Education, Proceedings*, 278-283 (2020).
23. Pan, F., Huang, B., Zhang, C., Zhu, X., & Wu, Z. A survival analysis based volatility and sparsity modeling network for student dropout prediction. *PLoS ONE*, 1-22 (2022).
24. Pujianto, U., Agung Prasetyo, W., & Rakhmat Taufani, A. Students Academic Performance Prediction with k-Nearest Neighbor and C4.5 on SMOTE-balanced data. *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2020*, 348-353 (2020).
25. Ramanathan, K., & Thangavel, B. Minkowski Sommon Feature Map-based Densely Connected Deep Convolution Network with LSTM for academic performance prediction. *Concurrency and Computation: Practice and Experience*, 1-17 (2021).
26. Ranjeeth, S., Latchoumi, T. P., Sivaram, M., Jayanthiladevi, A., Kumar, T. S., Anuar, M. A., . . . Zulkafli, Z. D. Early prediction of student performance in blended learning courses using deep neural networks. *Proceedings - 2019 International Symposium on Educational Technology, ISET 2019*, 39-43 (2019).
27. Salal, Y. K., & Abdullaev, S. M. Deep Learning based Ensemble Approach to Predict Student Academic Performance: Case Study. *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 191-198 (2020).
28. Sun, Y., Li, Z., Li, X., & Zhang, J. Classifier Selection and Ensemble Model for Multi-class Imbalance Learning in Education Grants Prediction. *Applied Artificial Intelligence*, 290-303 (2021).
29. Thaher, T., & Jayousi, R. Prediction of Student's Academic Performance using Feedforward Neural Network Augmented with Stochastic Trainers. *14th IEEE International Conference on Application of Information and Communication Technologies, AICT 2020 - Proceedings*. (2020).
30. Villa-Torrano, C., Bote-Lorenzo, M. L., Asensio-Pérez, J. I., & Gómez-Sánchez, E. Early prediction of students' efficiency during online assessments using a long-short term memory architecture. *CEUR Workshop Proceedings*, 39-46 (2020).
31. Widjaja, A. T., Wang, L., Nghia, T. T., Gunawan, A., & Lim, E.-p. Next-Term Grade Prediction : A Machine Learning Approach. *International Conference on Educational Data Mining (EDM)*, 700-703 (2020).
32. Wu, N. CLMS - Net : Dropout Prediction in MOOCs with Deep Learning. *ACM Association for Computing Machinery*, 17-19 (2019).
33. Xiao, Y., Yin, H., Zhang, Y., Qi, H., Zhang, Y., & Liu, Z. A dual-stage attention-based Conv-LSTM network for spatio-temporal correlation and multivariate time series prediction. *International Journal of Intelligent Systems*, 2036-2057 (2021).
34. Yusof, M. H., & Khalid, I. A. Precision Education Reviews: A Case Study on Predicting Student's Performance using Feed Forward Neural Network. *2021 International Conference of Technology, Science and Administration, ICTSA 2021*, 29-32 (2021).