OULAD MOOC Student Performance Prediction using Machine and Deep Learning Techniques

Wala Torkhani[1*] and Kalthoum Rezgui[1,2]

[1] University of Manouba, ISAMM, University Campus of Manouba, Manouba, Tunisia
[2] University of Tunis, ISG of Tunis, SMART Lab, Tunis Tunisia
*walatorkhani2@gmail.com

Abstract. In online learning, the accurate prediction of student performance is essential for timely interventions and personalized learning experiences. This work leverages the Open University Learning Analytics Dataset (OULAD) to evaluate the effectiveness of various machine learning (ML) and deep learning (DL) techniques in predicting student performance. We implemented a range of models, including traditional ML algorithms like Logistic Regression (LR), Decision Trees (DT), Random Forests (RF), and Support Vector Machines (SVMs), as well as DL models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks. The performance of each model was assessed using various metrics such as accuracy, precision, recall, and F1-score. The experimental results revealed that, among DL models, LSTM prevailed in terms of accuracy and precision, which are 83.41% and 82.20%, respectively. Additionally, the RF and optimized DT models performed well and provided a strong balance between accuracy and recall, making them a solid choice when computational efficiency is a concern.

Keywords: Performance prediction, Learning Analytics, Machine Learning, Deep Learning, OULAD dataset

1 Introduction

During last years, predicting student performance has become increasingly important in the fields of educational data mining (EDM) [1] and learning analytics (LA) [2]. Indeed, with the rise of online and blended learning environments, institutions are generating vast amounts of data on student interactions, behaviors, and academic progress. Collecting this data to predict student success can enable educators and administrators to provide timely interventions, personalize learning experiences, and improve overall educational outcomes.
In this context, the Open University Learning Analytics Dataset (OULAD) [3] provides a rich source of data for exploring predictive models. It includes detailed

S. Krichen et al. (eds.), *Proceedings of International Conference on Decision Aid and Artificial Intelligence (ICODAI 2024)*, Atlantis Highlights in Intelligent Systems 12,
https://doi.org/10.2991/978-94-6463-654-3_18

records of student demographics, assessment results, and interactions with virtual learning environments (VLE), making it an ideal dataset for evaluating the efficacy of various ML and DL techniques in predicting student performance.
To predict student performance in online learning, ML techniques, like Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), and Support Vector Machines (SVMs), have been widely employed due to their ability to model complex relationships within the data [10] [5].
While ML methods have proven effective, recent studies have also explored DL approaches [12] [11]. In fact, the advent of deep learning has introduced more sophisticated models, such as Convolutional Networks (CNNs) and Long Short-Term Memory (LSTM) networks, which are capable of capturing intricate temporal patterns and dependencies that may be missed by traditional ML algorithms.
This paper aims to evaluate the performance of both traditional ML algorithms and advanced DL models in predicting student performance using the OULAD dataset. By comparing these approaches, we seek to identify the strengths and limitations of each model and determine which one is most effective for this task.
The remainder of this paper is organized as follows: Section 2 provides a review of related work in the field of student performance prediction. Section 3 presents the OULAD dataset. Section 4 describes the methodology we proposed for predicting learners' performance and outlines the ML and DL models implemented in this study. Section 5 presents the experimental results and discussion. Finally, Section 6 concludes the paper and suggests directions for future research.

2 Related Work

The prediction of student performance on OULAD dataset using ML and DL techniques has recently attracted considerable attention in educational research, with researchers employing a wide range of methods to enhance predictive accuracy and interpretability. This section reviews key contributions to the field, focusing on the application of traditional ML techniques and advanced DL models.

2.1 ML algorithms in student performance prediction

In recent years, a number of studies have employed ML techniques to analyze the OULAD dataset, leveraging its rich information on student behavior and performance.
In their work, [13] conducted a comprehensive study using various ML algorithms, including LR, DT, and RF, to predict student performance and identify at-risk students. Their findings highlighted the effectiveness of feature selection in enhancing the predictive power of these models and provided insights into the importance of different types of data for accurate predictions.
In [4], the authors focused on analyzing learner performance in digital learning environments using ML methods to predict learners outcomes. The objective is

to develop indicators that alert students and teachers of potential challenges and allowing a proactive support. The proposed methodology encompasses a trace analysis approach, which involves manually and automatically selecting relevant attributes followed by rule extraction to explain learner outcomes. While the DT classifier is mentioned as one of the effective algorithms, the study also highlights that Gaussian NB, KNN and LinearSVC classifiers are also among the top performers based on precision, mean and standard deviation performance metrics.
An additional study conducted by [14] showed that ML algorithms achieved high accuracy rates in predicting student performance, reaching 89.7% for KNN and 97.40% for SVM.
In [16], the authors reviewed several approaches for predicting student outcomes in online courses using ML. They examined various ML techniques, including DT, SVM, and NN, and evaluated their effectiveness in forecasting student outcomes.

2.2 DL models in student performance prediction

With the advent of deep learning, new opportunities have been introduced for modeling the intricate patterns present in educational data.
Artificial Neural Networks (ANNs), with their ability to learn hierarchical representations, have been applied to various predictive tasks in education. For instance, [6] employed ANNs to predict student performance, demonstrating significant improvements in accuracy over traditional ML models, particularly when dealing with large and complex datasets.
In [15], the authors used a transformer encoder model designed for predicting student performance on the OULAD dataset. By analyzing students' sequential log activities, the model captures temporal dependencies and patterns in the data, offering improved predictive accuracy, achieving approximately 83.17%, compared to traditional methods.

Similarly, [11] proposed the DOPP model which leverages neural networks (NN) to analyze various features, such as clickstream data and demographic information, to predict student outcomes in online courses. This deep model outperforms traditional machine learning methods (i.e., LR, SVM), demonstrating its ability to capture complex relationships in the data.

Recently, the Long Short-Term Memory (LSTM) networks, a type of recurrent neural network designed to capture temporal dependencies, have shown promise in educational contexts where sequential data is critical.
In [7], the authors utilized LSTMs to model the progression of student knowledge over time, achieving state-of-the-art results in predicting future performance based on past interactions. This ability to model temporal sequences makes LSTMs particularly suitable for analyzing clickstream data and other time-dependent educational metrics.
Recent research has increasingly focused on hybrid approaches that combine the strengths of both traditional ML and DL methods. These approaches aim to leverage the interpretability of ML models with the powerful feature extraction capabilities of DL models. For instance, [9] developed a hybrid approach

that integrates LR with LSTM networks to predict student dropout in MOOCs, successfully balancing predictive accuracy with model interpretability.

3 OULAD Dataset

This section introduces the Open University Learning Analytics Dataset (OULAD) (available for download here: https://analyse.kmi.open.ac.uk/open_dataset) which is one of the most popular open datasets in educational data mining and learning analytics research [3].
The OULAD dataset encompasses data from seven different courses modules (AAA GGG) and 22 courses offered by The Open University (OU) in the United Kingdom. The data is collected from approximately 32,593 students who were enrolled in these courses between the years 2013 and 2014. More precisely, this dataset is structured across seven data tables (See Figure 1), each representing different aspects of student interaction (clickstream data), demographic information, and assessment outcomes. These tables are available in a comma separated

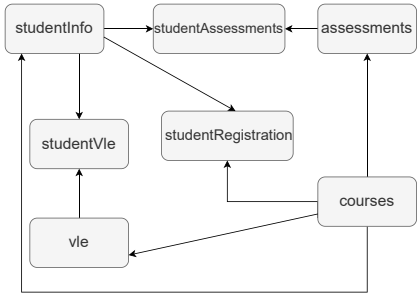

Fig. 1. Detailed OULAD dataset structure.

value (CSV) format, the content of each .csv file is described below. Besides, the corresponding details of these files have been provided in Table 1.

- **Courses.csv:** Contains metadata about each course, such as the module code (*code_module*) identifying the course, the code name of the presentation (*code_presentation*).
- **Assessments.csv:** Provides details on module presentations (*code_module*, *code_presentation*) and various assessments within each course, including the type of assessment (*assessment_type*) (e.g., assignments, exams), and the submission dates (*date*).
- **StudentAssessment.csv:** Contains the results of students' assessments. The *date_submitted* attribute states the date of student submission, measured as the number of days since the start of the module presentation. The

is_banked attribute is a status flag indicating that the assessment result has been transferred from a previous presentation. The *score* column is the student's score in this assessment which ranges from 0 to 100. A score lower than 40 is interpreted as Fail.

- **studentInfo.csv:** Includes demographic information of students, such as age, gender, region of residence, highest education level, disability status, socio-economic background, total number of credits for the module the student is currently studying as well as the student's final result (i.e., withdrawn, fail, pass or distinction).
- **StudentRegistration.csv:** Contains the date when the students registered for (and eventually unregistered from) a module presentation (date_registration, date_unregistration). Students who completed the course have the field date_unregistration empty whereas students who unregistered have Withdrawal as the value of the *final_result* column in the *studentInfo.csv* file.
- **StudentVle.csv:** Captures students' interactions with the VLE resources, such as the number of clicks on different course materials, forum participation, and accessing to learning resources. The *date* attribute expresses the date of student's interaction with the material measured as the number of days since the start of the module presentation. The *sum_click* attribute measures the number of times a student interacts with the material in that day.
- **Vle.csv:** Includes information about the resources available in the VLE.

Table 1. CSVFiles and corresponding details in OULAD dataset.

File ID	Attributes
courses.csv	code_module, code_presentation, length
studentInfo.csv	code_module, code_presentation, id_student, gender, region, highest_education, imd_band, age_band, num_of_prev_attempts, studied_credits, disability, final_result
studentRegistration.csv	code_module, code_presentation, id_student, date_registration, date_unregistration
studentAssessment.csv	id_student, id_assessment, date_submitted, is_banked, score
assessments.csv	code_module, code_presentation, id_assessment, assessment_type, date, weight
vle.csv	id_site, code_module, code_presentation, activity_type, week_from, week_to
studentVle.csv	id_student, id_site, code_module, code_presentation, date, sum_click

4 Methodology

This section focuses on describing the methodology (See Figure 2) that we propose for predicting learners' performance on the OULAD dataset using ML and DL models.

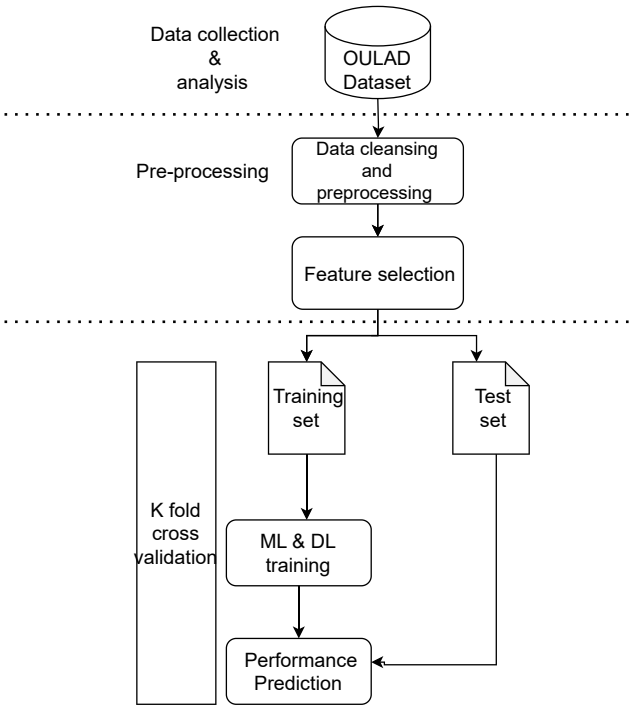

Fig. 2. Architecture of the proposed performance prediction approach.

4.1 Pre-processing

This stage encompasses the data pre-processing and transformation performed prior to building the proposed predictive models. Initially, we filtered the original dataset to focus on the AAA course module and the associated educational data of 748 students who participated in this course. Then, the *studentInfo* table, which contains demographic information, was merged with the *studentRegistration* table to track each student's enrollment status, and with the *studentAssessment* table to incorporate assessment scores [13]. Next, the *studentAssessment*

table was used to compute features related to students' performance on individual assignments and exams. These include total assessment scores, average scores, the number of assessments submitted, and the the number of failed assessments ($assignment\_failed$) for each student.
As the *studentVle* table contains detailed clickstream data recording every interaction a student has with the VLE, this table is aggregated to create features such as the number of unique sites visited, the total number of clicks, the number of days active, and the average number of clicks per day.
The processed dataset is further cleaned by removing invalid entries while missing values were handled. For instance, numeric attributes, like as *score*, assignment_failed, *id_site*, *sum_click*, with missing values were filled with appropriate replacements. In particular, missing score values were set to 0, indicating no score was recorded, while missing $assignment\_failed$ values were filled with the maximum observed value, suggesting a high likelihood of failure.

4.2 Feature selection

As outlined in Section 3, the OULAD dataset consists of seven data tables with 40 unique columns (attributes). Among these attributes, we have selected and added new attributes to learn and predict each learner's final result based on the data associated with these attributes. Indeed, attributes that are not significant for the learning process have been removed, such as *length*, *date_unregistration*, *id_assessment*, *date_submitted*, *is_banked*, $week\_from$, and *week_to*.
Table 2 below provides a summary of the OULAD dataset after the feature extraction step.

4.3 Proposed predictive models

This research work aims to use ML and DL methods to predict the performance of students by analyzing their related features and interaction traces with the learning environment. The main idea is to find the most appropriate prediction model and compare the various performance metrics of each classifier used.

4.4 Machine Learning Models

In this research work, the following classification ML algorithms are used:

- Logistic Regression (LR) is a statistical model used for binary classification tasks. It estimates the probability of a binary outcome (e.g., pass/fail) based on one or more predictor variables.
- Decision Trees (DT) are a type of supervised learning model that splits the data into subsets based on feature values to make predictions.
- Decision Tree Optimization (DT_opt) is an optimized DT model that involves tuning the hyperparameters of the C4.5 algorithm to improve its performance.

- Random Forest (RF) is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (for classification).
- Support Vector Machines (SVMs) is a supervised learning model used primarily for classification tasks. It works by finding the optimal hyperplane that maximizes the margin between different classes in the feature space.
- Naive Bayes (NB) is a probabilistic classifier based on Bayes' theorem. It assumes that the features are conditionally independent given the class label (hence "naive").

4.5 Deep Learning Models

The following DL models were considered for predicting student performance:

- Multilayer perceptron (MLP) is a class of feedforward artificial neural networks consisting of multiple layers of neurons. It typically includes an input layer, one or more hidden layers, and an output layer, where each layer is fully connected to the next.
- Convolutional Neural Networks (CNNs) are primarily used for image processing but can be adapted for structured data by treating data as a grid. They apply convolutional layers to extract features from input data. CNNs are effective at identifying hierarchical patterns and reducing dimensionality.
- Recurrent Neural Networks (RNNs) are designed to handle sequential data by maintaining a memory of previous inputs. Their main advantage lies in their ability to handle sequential and time-series data, capturing dependencies over time.
- Long Short-Term Memory (LSTM) is a type of RNN specifically designed to overcome the limitations of traditional RNNs by maintaining long-term memory through a series of gates that control the flow of information. These networks are used to predict student performance by analyzing sequences of interactions over time, allowing the model to learn temporal dependencies.

5 Experiment and results

In this section, we present the experimental results of the proposed ML and DL models. First, we outline the parameter settings used for each ML and DL model applied in this work. Then, we give the experimental results for these models in predicting students' performance. The implementation and experiments of all proposed models were implemented using the Python 3.0 language and its Scikit-Learn and TenserFlow libraries.

5.1 Parameter settings

Following preprocessing, the dataset was split into two data sets with a ratio of 7:3, i.e., 70% of data was used as training set and the rest 30% of the data was

used as testing set.
The evaluation metrics we selected to assess the performance of the proposed prediction models are accuracy, precision, recall, and F1-score.
In the model development stage, each model was configured with a set of hyper-parameters that influence the learning process and the resulting performance. These hyperparameters were either set to their default values or tuned based on cross-validation to improve the models' performance on the OULAD dataset.
The following table (see Table 3) presents the parameters used for each model, along with their descriptions and values.

5.2 Prediction results

Figure 3 illustrates the comparative performance of proposed ML models across the key performance metrics. It can be seen from this figure that the RF model achieved the highest scores for most metrics, particularly excelling in recall and accuracy with values of 85.69% and 82.73%.
It can, also, be revealed that the DT_opt and LR models performed well, their accuracy are between 78% and 79%. However, DT_opt provides a better recall than LR, making it a solid choice to get accurate predictions. The NB model is the less effective in terms of all performance metrics, but still offers value for rapid and lightweight predictions.

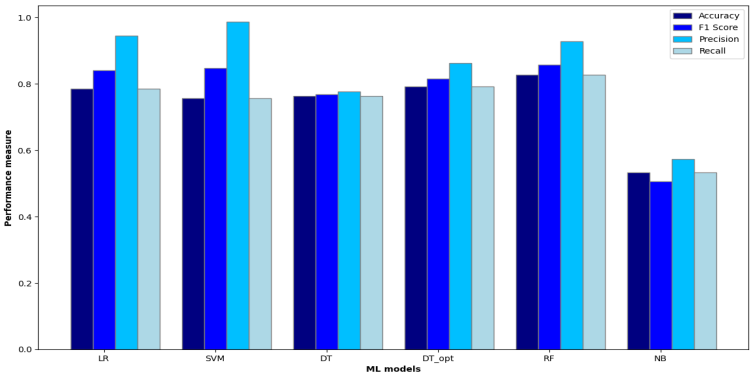

Fig. 3. Comparative performance analysis of ML models.

The experimental results in Figure 4 illustrates the performance of various deep learning models in predicting student performance.
The LSTM model exhibits the highest performance across all metrics, with an accuracy of 83.41%, a precision of 82.20%, and a recall of 81.88%. This indicates that the LSTM model is the most effective in modelling time-series data and sequential nature of the OULAD dataset.

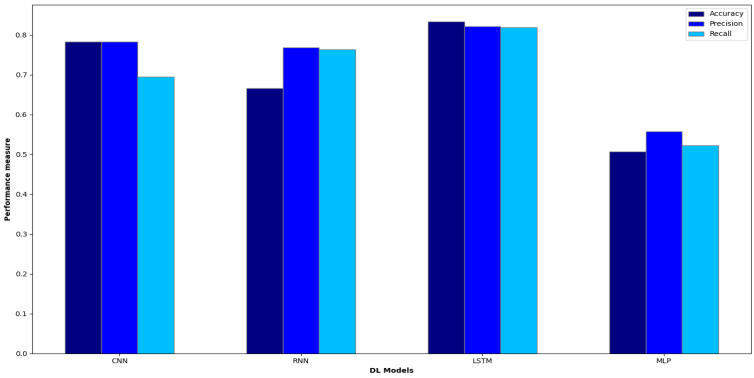

Fig. 4. Comparative performance analysis of DL models.

6 Conclusion

In this study, we explored the effectiveness of various machine learning and deep learning models in predicting student performance using the OULAD dataset. Our goal was to identify the most suitable model for accurately predicting student outcomes, which is crucial for enabling timely interventions and enhancing educational support systems.

The results of our experiments indicate that DL models, particularly LSTM, outperforms other models in accuracy, precision, and recall metrics.

Among traditional ML models, RF and DT_opt demonstrated strong performance, making them viable alternatives, especially when computational efficiency or model interpretability is a priority.

The significance of these findings advances that educational institutions can leverage LSTM models to better predict student performance in online learning and MOOCs. Furthermore, while DL models require more computational resources, their superior performance in handling complex and sequential data justifies their use in scenarios where prediction accuracy is crucial.

Future research could focus on exploring hybrid approaches that combine the strengths of both ML and DL techniques, and expanding the analysis to include additional datasets and educational environments. Moreover, integrating interpretability techniques will be essential to ensure that the predictions made by DL models can be understood and acted upon by educators.

Table 2. Description of the OULAD dataset after feature selection.

Attributes	**Description**
code_module_AAA	Identification code of AAA module course on which the student is registered.
code_module_DDD	Identification code of DDD module course on which the student is registered.
code_presentation	Code name of the presentation
age_band_0-35, age_band_35-55	Different age ranges
assignment_failed	Number of assignments failed by a student
sum_click_x	Sum of the students clicks on online learning materials before the first third of the course module
sum_click_y	Sum of the student's clicks on online learning materials during the remaining period
imd_band_0-10%, imd_band_10-20%, imd_band_20-30%, imd_band_30-40%, imd_band_40-50%, imd_band_50-60%, imd_band_60-70%, imd_band_70-80%, imd_band_80-90%	Different bands of the Index of multiple deprivation
gender_F	Indicator or binary flag for female gender
gender_M	Indicator or binary flag for male gender
id_student	A unique identification number for the student
studied_credits	Total number of credits for the modules
date_registration	Number of days measured relative to the start of the module presentation
num_of_prev_attempts	Number of times the student has attempted the course module
score	Student's score in an assessment
date_diff	Difference in days between the planned date for an assessment and the date submitted
activity_type	Type of activity in the VLE converted into a numerical index
total_days	Total number of noted activity days
actionArray_train and actionArray_test	Students interactions over time with the VLE encoded using three-dimensional arrays representing the number of clicks per student, per day, and per activity type
region_East Anglian, region_East Midlands, region_Ireland, region_South, region_London, region_North, region_Scotland, region_North Western, region_South East, region_South West, region_Wales, region_West Midlands	Variables for different regions

Table 3. Prediction models parameter settings.

Model	*Parameter*	*Description*	*Value*
LR	C	Inverse of regularization strength	0.1
	max_iter	Maximum number of iterations taken for the solvers to converge	100
	penalty	A regularization term added to the loss function to prevent overfitting	l1
	solver	Algorithm to use in the optimization problem	saga
SVM	C	Regularization parameter strength	1.0
	kernel	Kernel type to be used in the algorithm	rbf
DT	max_depth	Maximum depth of the tree	7
	min_samples_split	Minimum number of samples required to split an internal node	5
	criterion	Function to measure the quality of a split	gini
RF	n_estimators	Number of trees in the forest	100
	max_features	Number of features to consider when looking for the best split	auto
	bootstrap	Whether bootstrap samples are used when building trees	true
NB	var_smoothing	Portion of the largest variance of all features added to variances for numerical stability	1e-9
MLP	activation	Function applied to the output of each neuron in a layer	sigmoid
		Number of neurons in each hidden layer	
	hidden_units	Number of time steps that the model processes at once	128
	n_steps	Number of input features that the model processes at	10
	n_features	each time step or input instance	20
	loss	Function used to measure the difference between the model's predictions and the actual target values	binary
	optimizer	Algorithm used to update the model's weights during training in order to minimize the loss function	crossen-tropy Adam
CNN	activation	Function applied to the output of each neuron in a layer	sigmoid
		Number of neurons in each hidden layer	
	hidden_units	Number of time steps that the model processes at once	10
	n_steps	Number of filters in the convolutional layers	20
	filters	Size of the convolutional kernel	64
	kernel_size	Size of the max-pooling window	(3,3)
	pool_size		(2,2)
RNN	units	Number of LSTM units (neurons) in the layer	64
	activation	Function applied to the output of each neuron in a layer	ReLU
LSTM	units	Number of LSTM units (neurons) in the layer	300
	n_steps	Number of time steps or sequential inputs that the model processes	10
	n_features	Number of input features that the model processes at each time step	20
	mask_value	A value that is ignored or masked during training or evaluation	0
	return_sequences	Specifies whether the output of each time step should be returned or just the output of the final time step	False
	dropout	Fraction of the input units to drop for regularization	0.1
	recurrent_dropout	Fraction of the recurrent units to drop for regularization	0.1
	activation	Function applied to the output of each neuron in a layer	sigmoid

References

1. Peña-Ayala, A. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems With Applications*. **41**, 1432-1462 (2014)
2. Leitner, P., Khalil, M. & Ebner, M. Learning analytics in higher education—a literature review. *Learning Analytics: Fundaments, Applications, And Trends: A View Of The Current State Of The Art To Enhance E-learning*. pp. 1-23 (2017)
3. Kuzilek, J., Hlosta, M. & Zdrahal, Z. Open university learning analytics dataset. *Scientific Data*. **4**, 1-8 (2017)
4. Sehaba, K. Learner Performance Prediction Indicators based on Machine Learning. *CSEDU (1)*. pp. 47-57 (2020)
5. Albreiki, B., Zaki, N. & Alashwal, H. A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences*. **11**, 552 (2021)
6. Kotsiantis, S., Pierrakeas, C. & Pintelas, P. Predictiong students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*. **18**, 411-426 (2004)
7. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. & Sohl-Dickstein, J. Deep knowledge tracing. *Advances In Neural Information Processing Systems*. **28** (2015)
8. Crossley, S., Paquette, L., Dascalu, M., McNamara, D. & Baker, R. Combining click-stream data with NLP tools to better understand MOOC completion. *Proceedings Of The Sixth International Conference On Learning Analytics & Knowledge*. pp. 6-14 (2016)
9. Whitehill, J., Williams, J., Lopez, G., Coleman, C. & Reich, J. Beyond prediction: First steps toward automatic intervention in MOOC student stopout. *Available At SSRN 2611750*. (2015)
10. Chen, Y. & Zhai, L. A comparative study on student performance prediction using machine learning. *Education And Information Technologies*. **28**, 12039-12057 (2023)
11. Karimi, H., Huang, J. & Derr, T. A deep model for predicting online course performance. *Association For The Advancement Of Artificial Intelligence*. pp. 1-6 (2020).
12. Alnasyan, B., Basheri, M. & Alassafi, M. The Power of Deep Learning Techniques for Predicting Student Performance in Virtual Learning Environments: A Systematic Literature Review. *Computers And Education: Artificial Intelligence*. pp. 100231 (2024)
13. Gašević, D., Dawson, S., Rogers, T. & Gasevic, D. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet And Higher Education*. **28** pp. 68-84 (2016)
14. Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., Jiang, B. & Chen, P. Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports*. **12**, 453 (2022)
15. Kusumawardani, S. & Alfarozi, S. Transformer encoder model for sequential prediction of student performance based on their log activities. *IEEE Access*. **11** pp. 18960-18971 (2023) .
16. Alhothali, A., Albsisi, M., Assalahi, H. & Aldosemani, T. Predicting student outcomes in online courses using machine learning techniques: A review. *Sustainability*. **14**, 6199 (2022).