Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 238 (2024) 888–895

Procedia Computer Science

www.elsevier.com/locate/procedia

The 14th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2024)
April 23-25, 2024, Hasselt, Belgium

A Hybrid Machine Learning Approach for Predicting Student Performance Using Multi-class Educational Datasets

Ghaith Al-Tameemi[a,*], James Xue[a], Israa Hadi Ali[b], Suraj Ajit[a]

[a] *University of Northampton, Northampton, United Kingdom*
[b] *University of Babylon, Babil, Hilla, Iraq*

Abstract

Prediction of students' academic performance has garnered considerable interest, with many institutions seek to enhance students' performance and their quality of education. The integration of both unsupervised and supervised machine learning techniques has demonstrated significant efficacy in predicting student performance. This paper explores the application of different machine learning methods in predicting student academic performance. Initially, Principal Component Analysis (PCA) was utilised to reduce the dataset's dimensionality, thereby improving its visualisation. Subsequently, K-Means clustering was employed to segregate students into distinct groups, reflective of their learning behaviors. Afterwards, the observed clusters were utilised for training classification models to address each student cluster individually. This approach was implemented in a case study involving an undergraduate science course at a North American University (NAU) and the Open University Learning Analytics Dataset (OULAD). Empirical findings indicate that the combined use of Feedforward Dense Network (FDN), Random Forest (RF), and Decision Tree (DT), specifically in their clustered forms, outperforms other classifiers in predicting student academic performance effectively.

Keywords: Hybrid Machine Learning; Principal Component Analysis; K-Means clustering; Student Performance

1. Introduction

The education sector's digital transformation has significantly influenced teaching and learning, leading to the rise of learning analytics. This analytics helps institutions enhance financial stability, improve learning outcomes, and formulate effective policies [1]. Universities focus on hiring qualified staff and enhancing teaching methods to boost student performance, adjusting curricula and methods based on factors influencing student learning [2]. Most institutions now use Learning Management Systems (LMS) for course delivery and student engagement tracking [3]. Analysing LMS data helps in addressing issues like low achievement and high dropout rates [4]. This paper introduces

* Corresponding author
E-mail address: ghaith.al-tameemi@northampton.ac.uk

1877-0509
10.1016/j.procs.2024.06.108

a hybrid machine learning approach to predict student performance using NAU and OULAD datasets. The paper's major contributions include:

- A hybrid approach combining unsupervised and supervised learning for multi-class educational datasets.
- Utilisation of PCA and pairwise correlation for optimal feature selection.
- Enhanced prediction by training clusters with various classifiers.
- Comparative analysis between clustered and non-clustered classifiers.

The rest of this paper is organised as follows. Section 2 covers recent advancements in machine learning techniques to predict student academic performance. Section 3 presents the proposed hybrid machine learning approach for predicting student performance. Section 4 reports the experimental results and discussion. Finally, Section 5 concludes the paper and outlines some future research directions.

2. Related Work

Several research projects were conducted in educational data mining, and it remains a controversial topic in deep learning and machine learning. Many researchers strive to create an automatic system that can predict grades, marks, institution ratings, and institutional recommendations using different methodologies and tools. This section seeks to comprehend existing research works that can aid in identifying research gaps and developing the proposed model approaches.

Educational data mining remains a contentious field in deep learning and machine learning, focusing on predicting academic aspects like grades and institution ratings using various methodologies. The researchers in [5] developed a hybrid method combining k-means clustering and decision trees to categorise students and achieved 86.25% accuracy. They highlighted key factors such as attendance and family income. The authors in [6] employed a deep neural network to predict student performance using OULAD. Their proposed method achieved an accuracy of over 95% for a specific course.

The authors in [7] utilised machine learning algorithms to classify student performance based on demographics and academic attributes, achieving 88.54% accuracy. They emphasised factors like high school GPA and university entrance exam scores. The authors in [8] employed ANNs for predicting student performance, obtaining 91.2% accuracy, but their work had limitations due to lack of external validation. The authors in [9] explored deep learning models like CNNs and LSTM for predicting final grades, outperforming traditional algorithms but requiring more computational resources. The researchers in [10] proposed a model combining feature selection, clustering, and classification, effective for small datasets but potentially lacking generalisability.

3. The Proposed Hybrid Machine Learning Approach

This paper discusses the use of the K-Means clustering and different machine learning classifiers (hybrid classifier) to obtain better accuracy results for predicting students' academic performance. We used publicly available multi-class educational datasets of NAU and OULAD. These datasets consist of several learning activities such as LMS interactions, effort activities, and assessment scores.

3.1. Data Collection

The paper utilised raw data from the OC2 lab, detailing an undergraduate science course at a North American University (NAU). The dataset, with 486 students, included grades and event logs [11]. It encompassed interaction data in the Learning Management System (LMS) across five activities (F1-F5), covering logins, content and forum interactions, and quiz reviews. Additionally, it included effort data (F6-F7) representing time spent and lateness in assignments. Students were categorised as Good (G), Fair (F), or Weak (W).

The Open University Learning Analytics Dataset (OULAD) was also referenced [12]. This dataset contained demographic information, VLE behavior logs, and final results. OULAD included seven files detailing various aspects of student and course information. The AAA module for 2013J involved eight e-learning activities (F1-F8) and an

assessment score feature (F9), with student results categorised as Distinction, Pass, or Fail (F10). These features are described in Table 2.

Table 1. List of all NAU features

Features	Feature code	Description	Type	Range value
F1	N_{LG}	This indicates how many times students accessed the course site on the LMS	Interaction	0-647
F2	N_{CR}	This indicates how many times students accessed course material	Interaction	0-1007
F3	N_{PR}	This indicates how many times students read posts on the discussion forum	Interaction	0-58
F4	N_{FR}	This indicates how many times students posted on the discussion forum	Interaction	0-6
F5	N_{QR}	This indicates how many times students reviewed their quiz solution before final submission	Interaction	0-12
F6	$As1_{LI}$	This indicates how many times that students had assignment late submission	Effort	0-3
F7	Avg_D	Average amount of time between posting and submitting Assignment (in hours)	Effort	0-496
F8	T_G	Total final mark	Numeric	0-100
F9	Cl	Student category	Nominal	G,F,W

Table 2. List of all OULAD features

Features	Feature code	Description	Type	Range value
F1	*oucontent*	Total clicks on assignment's contents	Interaction	0-216
F2	*url*	Total clicks on video's links	Interaction	0-167
F3	*homepage*	Number of clicks on the homepage	Interaction	11-217
F4	*resource*	Number of clicks on the pdf's resources	Interaction	0-41
F5	*forumng*	Number of clicks on the discussion forum	Interaction	0-842
F6	*oucollaborate*	Number of clicks on the video discussions	Interaction	0-5
F7	*glossary*	Number of clicks on the basic glossary related to contents of course	Interaction	0-14
F8	*dataplus*	Total number of clicks on the additional information (videos, audios, sites)	Interaction	0-10
F9	*score*	Total assessments' scores for this module	Effort	64-453
F10	*final_result*	Student final result	Nominal	Distinction, Pass, Fail

3.2. Data Pre-processing

In this section, we describe the steps that were performed prior to training the machine learning models (clustering and classification). The pre-processing steps comprised removing missing features, scaling and extraction to improve the machine learning performance. We removed the assignment lateness indicator's feature (F6) in clustered machine learning models for some clusters because no students had late submissions in various clusters for the NAU dataset. Similarly, features F6, F7, and F8 were also removed from some clusters in OULAD because no students were logged in to those features.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

In this paper, Python was used for data scaling pre-processing, detailed in equation 1, where x is the activity type input, μ the average, and σ the standard deviation. Feature sets, explained in section 4, include output features F9 and F10 for NAU and OULAD datasets, representing final student results. "One-hot encoding" converted categorical data to numerical format for the machine learning model, mapping categories to 1 (presence) or 0 (absence).

3.3. Unsupervised Machine Learning

K-Means cluster analysis is used to determine the differences between students' levels of learning activities. The number of clusters was between two and four. Due to the high-dimensional features often containing a significant amount of redundant information, it was essential to eliminate some of the redundant data from the initial features to prevent dimension disasters and improve state recognition performance.

A highly effective technique for feature selection in unsupervised learning is the Principal Component Analysis (PCA). PCA is a valuable data analysis and processing techniques in learning analytics. The main target of using PCA is to summarise the dataset's features, indicate the directions of the largest variance, and calculate the principal components. In this work, PCA is used to reduce the dimensionality of the data and enhance its visualisation.

After that, we used the K-means algorithm with K values ranging from two to four clusters [5]. The sufficient number of clusters was assured based on results findings and the use Within Cluster Sum of Squares (WCSS), which will be described in section 4.3. With K-Means clustering, n objects are divided into K clusters, with each cluster having the greatest distinction possible. This was aimed at minimising the squared error function or the total variance within a cluster. For clarity, let's consider the following: K denotes the number of clusters, n denotes the number of students, $z_i^{(j)}$ is the case of student activity, and c_j is the number of random centroid points.

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} \left\| z_i^{(j)} - c_j \right\|^2 \tag{2}$$

Data is first divided into K clusters, which are initially defined. Randomly selected c points serve as cluster points. Next, Euclidean distance is applied to assign the objects to their nearest points. Then, the average of all objects within each cluster is calculated. Each cluster is then assigned the same points successively until all the parameters were the same in each round. The final step is to identify an optimal number of clusters by using WCSS.

3.4. Supervised Machine Learning Classifiers

This work used several types of machine learning classifiers to predict students' academic performance. These classifiers are outlined below.

3.4.1. Classical Machine Learning

In the classical machine learning section, various algorithms are discussed for their roles in predicting student academic performance and other classification tasks. The K-Nearest Neighbors (KNN) algorithm, a supervised machine learning method, is used for classification and regression tasks. It assigns values to new data points based on feature similarity to training data [13]. Support Vector Machines (SVM) operate by finding hyperplanes in an N-dimensional space to distinctly classify data points, effectively separating non-linear data in higher-dimensional spaces [14].

The Naive Bayes (NB) classifier uses the Bayesian theorem for prediction, standing out for its fast training speed compared to other classifiers. It relies on conditional probability for classifying input data [15]. Decision Trees (DT) use non-parametric criteria for classification and prediction, inferring simple decision rules from data features. This method is particularly noted for its usefulness in predicting student academic performance [16]. Random Forest (RF) is an advanced technique based on multiple decision trees. It combines predictor trees where each is based on a random vector, proving effective in classification tasks including detecting at-risk students and recommending courses [17].

3.4.2. Feedforward Dense Networks (FDN)

Artificial neural networks are employed to classify students' academic performance. Deep learning methods are also known as representational learning techniques since they comprise various layers of nonlinear modules. After which the system then can be trained to recognise complicated tasks allowing it to grasp complex and tiny details. In contrast to statistical approaches, a deep neural network aids generalisation by implying hidden patterns from data and formulating data-driven predictions [9].

Increasing the training split in a network enhances learning and accuracy. Hidden layers, nonlinear and between input and output layers, adjust weights via stochastic gradients for error computation in classification. A deep neural network, comprising input, hidden, and output layers, uses Feedforward Dense Networks (FDN) for predicting student performance. FDN architecture varies with dataset dimensionality. For OULAD, FDN inputs eight e-learning interactions (F1-F8) and assessment score (F9), processing through three Dense layers with Dropout and BatchNormalization (Batch_N). In contrast, the NAU dataset's FDN takes seven inputs (F1-F7) into three similar Dense layers (Fig. 1). The network employs batch normalization for regularization and dropout to prevent overfitting.

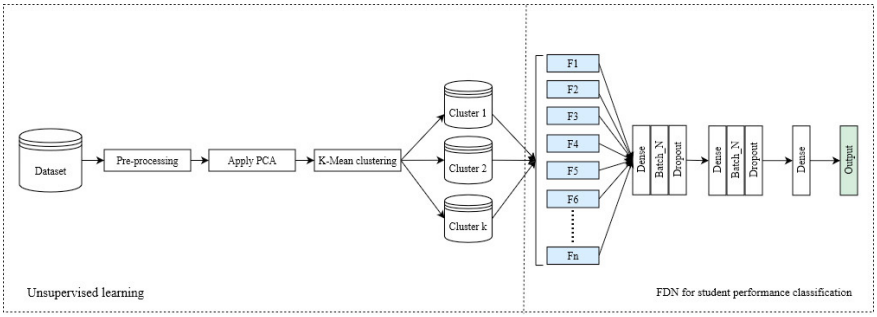


Fig. 1. Clustered FDN Architecture

3.5. Performance Metrics

We employed accuracy score (equation 3) and kappa analysis (equation 4) to evaluate our predictive models. Accuracy measures the model's overall correctness and kappa analysis compares the classifier's accuracy against random classification. WCSS is used for optimal cluster selection in clustering models. Further details are provided in section 4.

$$accuracy = \frac{\sum_{i=1}^{C} \frac{TP_i+TN_i}{TP_i+FP_i+TN_i+FN_i}}{C} \tag{3}$$

$$kappa = \frac{n \times s - \sum_{i=1}^{C} p_i \times t_i}{s^2 - \sum_{i=1}^{C} p_i \times t_i} \tag{4}$$

Here, C is the total number of classes; TP_i, TN_i, FP_i, and FN_i denote the true positives, true negatives, false positives, and false negatives for each class, respectively. In the kappa equation, n is the total number of elements correctly predicted, s is the total number of elements, p_i is the total number predicted for class i, and t_i is the number of times class i actually occurs.

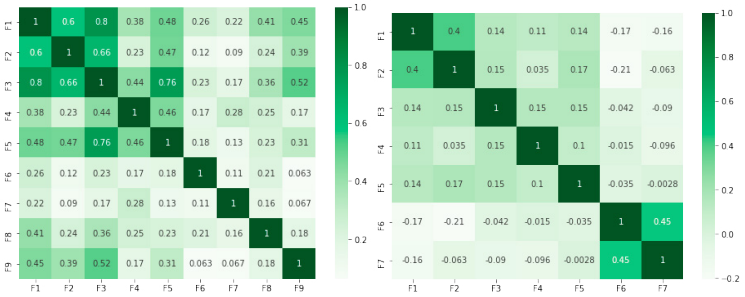
4. Experimental Results

4.1. Pairwise Correlation

The pairwise correlation coefficient, shown in Fig. 2a and Fig. 2b, reveals common variations in student learning activities. A positive correlation suggests similar magnitudes and signs between variables, while a negative one indicates similar magnitudes but opposite signs. For NAU, LMS interaction activities (F1 to F5) are positively correlated, with values ranging from .03 to .40, and are associated with students' final results (F9). Contrastingly, student effort activities (F6 and F7) negatively relate to interaction features (F1 to F5), with F6 showing a significant positive relation to the average submission time (F7). In OULAD, all e-learning interactions have positive correlations, ranging from 0.06 to 0.8. The strongest correlations are between F1 and F3 (0.8), and F3 and F5 (0.76), leading to the removal of F3 from the predictor list.

4.2. Models Setup

This section presents results from hybrid classifiers. Initially, the Pearson correlation coefficient was utilised to assess the connection between student learning activities and their final outcomes, aiding in input variable analysis and dataset dimensionality reduction to boost the machine learning model's predictive efficiency. The model, referenced in section 3.4, was trained and evaluated using FDN and other classifiers like KNN, SVM, NB, RF, and DT. We used



(a) OULAD correlation (b) NAU correlation

Fig. 2. Correlation coefficient analysis for the two datasets

categorical_cross_entropy as the loss function, ReLU for Dense layer activation, and softmax for output layers, with Dense layer units tailored to the dataset's size. FDN parameters are detailed in Table 3. Over-fitting is crucial in deep

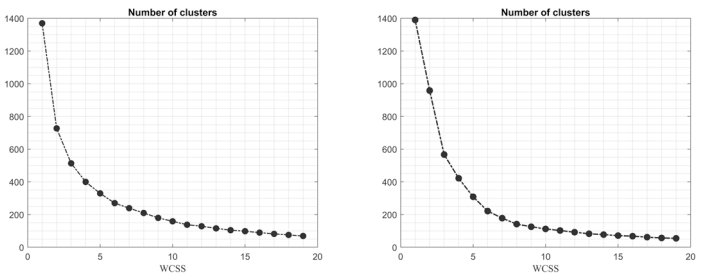
Table 3. FDN parameters

Parameter	Value
Input dimensions	7 to 9
Output dimensions	2, 3 (based on the subset classes)
Neuron at hidden layers 1, 2	200, 100
Input activation function	ReLU
Output activation function	Softmax, sigmoid
Batch-size	8
Train split	0.8
Test split	0.2
Optimizer	Adam
Loss function	categorical_crossentropy, binary_crossentropy

learning model training, arising when input layers outnumber output layers or when neuron count in hidden layers deviates from a set threshold. Prior setting of these parameters is essential before training. To counter over-fitting, we used a Dropout layer. Furthermore, sufficient training data is needed, with 80% of the dataset being allocated for training and 20% for testing.

4.3. Hybrid Machine Learning Results

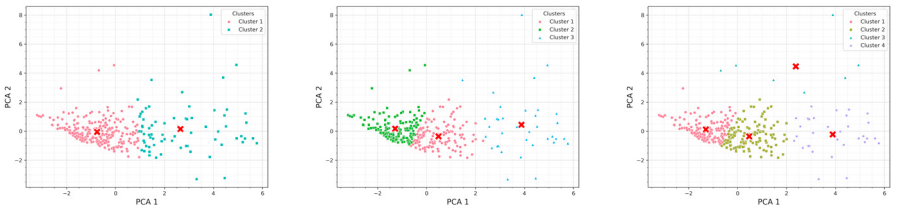
To begin with, the PCA was used to reduce dimensionality and improve the visibility of the data. Our model used a number of clusters between 2 and 4 (Fig. 3).



(a) OULAD (b) NAU

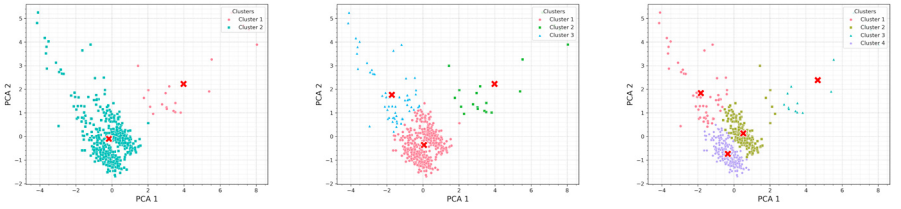
Fig. 3. WCSS results

As shown in Fig. 4 and Fig. 5, PCA analysis was used to project computed clusters into k dimensions. These figures show the results of K-Means clustering for student activities level for the NAU and OULAD datasets. Each cluster of the unsupervised model was trained individually by various machine learning classifiers to enhance the prediction results. A binary classification was turned in some clusters as the Distinction, Fail, and W classes were missing in some clusters of the OULAD and NAU datasets. Fig. 6a presents the values of evaluation metrics for the clustered machine



(a) 2K (b) 3K (c) 4K

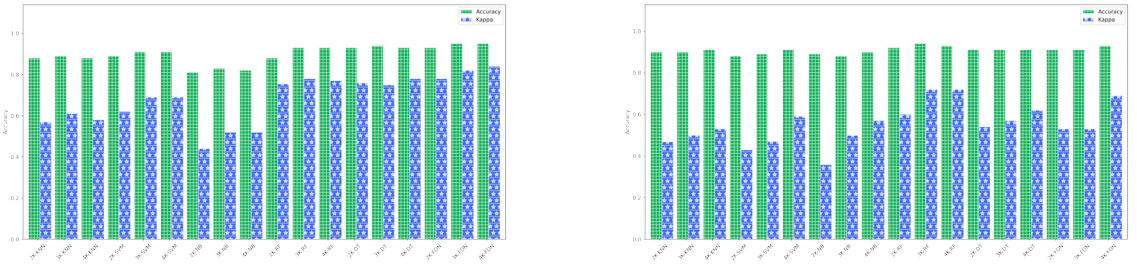
Fig. 4. K-Means clustering for OULAD



(a) 2K (b) 3K (c) 4K

Fig. 5. K-Means clustering for NAU

learning approach when trained using OULAD. It is noticeable that the 3Ks hybrid machine learning model were provided with the highest *accuracy* results compared with the other clustered machine learning classifiers. However, the FDN indicated that the four clusters *accuracy* was 95.3%. The SVM classifier has only a slight increase of about 0.3% in its *accuracy* for the three clusters hybrid model. In contrast, a slight drop was noticed in the KNN classifier from 89% to 88.7% accuracies. The RF, DT, NB classifier has a good *accuracy* increase of between 3% to 10%. The



(a) OULAD (b) NAU

Fig. 6. Accuracy and Kappa scores for OULAD & NAU

accuracies of RF, DT, NB were 93.4%, 93.7% and 83.4%. The overall accuracies for the clustered machine learning classifiers were 88.7%, 90.9%, 83.4%, 93.4%, 93.7% and 95.3% for KNN SVM, NB, RF, DT, and FDN, respectively.

In addition, the P_r, R_e, and $F1$ scores of the minor classes (Distinction and Fail) increased in the majority of the classifiers despite the KNN. The obtained *kappa* score for the RF, DT, and FDN were good as well. Fig. 6b presents the accuracies and kappa scores for the clustered machine learning approach when trained using the NAU dataset. In 2K classifiers, FDN, RF, and DT with 90.7%, 91.5%, and 90.53% accuracies were shown superior performance, respectively. Concerning 3K classifiers, RF, DT, and FDN likewise were presented the best performance with 93.8%, 90.7%, and 90.9% accuracies, respectively. In general, these 3K classifiers were provided consistent performance in all three classes compared to the non-clustered classifiers. The minimum *accuracy* result was the 3K-NB classifier with only 88.07%. The combination of interaction and effort activities provide a more comprehensive and consistent result.

5. Conclusion and Future Work

This paper aimed to predict students' academic performance using a hybrid machine learning approach, combining unsupervised and supervised methods. PCA reduced dataset dimensionality, and K-Means clustering grouped students. Clusters were trained with six classifiers (SVM, KNN, NB, RF, DT, FDN) on two educational datasets (OULAD, NAU). Results showed the clustered approach improved prediction, especially clustered FDN, which best identified at-risk students. Future work will explore deep learning and advanced clustering for early detection of at-risk students and test the methodology on additional datasets.

References

[1] D. Gašević, Y.-S. Tsai, and H. Drachsler, "Learning analytics in higher education – stakeholders, strategy and scale," *The Internet and Higher Education*, vol. 52, p. 100833, 2022.

[2] H. Khosravi, S. Shabaninejad, A. Bakharia, S. Sadiq, M. Indulska, and D. Gašević, "Intelligent learning analytics dashboards: Automated drill-down recommendations to support teacher data exploration," *Journal of Learning Analytics*, pp. 1–22, 2021.

[3] A. Al-Azawei, "What drives successful social media in education and e-learning? a comparative study on facebook and moodle.," *Journal of Information Technology Education*, vol. 18, 2019.

[4] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student engagement predictions in an e-learning system and their impact on student course assessment scores," *Computational intelligence and neuroscience*, 2018.

[5] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," *Journal of medical systems*, vol. 43, no. 6, pp. 1–15, 2019.

[6] G. Al-Tameemi, J. Xue, S. Ajit, T. Kanakis, I. Hadi, T. Baker, M. Al-Khafajiy, and R. Al-Jumeily, "A deep neural network-based prediction model for students' academic performance," in *2021 14th International Conference on Developments in eSystems Engineering (DeSE)*, pp. 364–369, IEEE, 2021.

[7] B. Sekeroglu, K. Dimililer, and K. Tuncal, "Student performance prediction and classification using machine learning algorithms," in *Proceedings of the 2019 8th International Conference on Educational and Information Technology*, pp. 7–11, 2019.

[8] E. Lau, L. Sun, and Q. Yang, "Modelling, prediction and classification of student academic performance using artificial neural networks," *SN Applied Sciences*, vol. 1, pp. 1–10, 2019.

[9] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from vle big data using deep learning models," *Computers in Human behavior*, vol. 104, p. 106189, 2020.

[10] A. Zohair and L. Mahmoud, "Prediction of student's performance by modelling small dataset size," *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, pp. 1–18, 2019.

[11] A. Moubayed, M. Injadat, A. Shami, A. B. Nassif, and H. Lutfiyya, "Student performance and engagement prediction in elearning datasets," 2020.

[12] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Scientific data*, vol. 4, no. 1, pp. 1–8, 2017.

[13] Y. Tang, J. Liang, R. Hare, and F.-Y. Wang, "A personalized learning system for parallel intelligent education," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 352–361, 2020.

[14] K. T. Chui, D. C. L. Fung, M. D. Lytras, and T. M. Lam, "Predicting at-risk university students in a virtual learning environment via a machine learning algorithm," *Computers in Human Behavior*, vol. 107, p. 105584, 2020.

[15] S.-S. M. Ajibade, N. B. Ahmad, and S. M. Shamsuddin, "A data mining approach to predict academic performance of students using ensemble techniques," in *International Conference on Intelligent Systems Design and Applications*, pp. 749–760, Springer, 2018.

[16] G. Akçapınar, M. N. Hasnine, R. Majumdar, B. Flanagan, and H. Ogata, "Developing an early-warning system for spotting at-risk students by using ebook interaction logs," *Smart Learning Environments*, vol. 6, no. 1, p. 4, 2019.

[17] H. Chanlekha and J. Niramitranon, "Student performance prediction model for early-identification of at-risk students in traditional classroom settings," in *Proceedings of the 10th International Conference on Management of Digital EcoSystems*, pp. 239–245, 2018.