International Journal of Evaluation and Research in Education (IJERE)
Vol. 15, No. 1, February 2026, pp. 436~447
ISSN: 2252-8822, DOI: 10.11591/ijere.v15i1.33590

Enhancing academic performance prediction in online learning through hybrid machine learning models

Jamal Eddine Rafiq[1], Zakrani Abdelali[1], Mohammed Amraouy[2], Said Nouh[3]
[1]Laboratory of Artificial Intelligence and Complex Systems Engineering, École Nationale Supérieure d'Arts et Métiers, Hassan II University, Casablanca, Morocco
[2]National Institute of Posts and Telecommunications (INPT), Mohammed V University, Rabat, Morocco
[3]Information Technology and Modeling, Faculty of Sciences Ben M'sik, Hassan II University, Casablanca, Morocco

Article Info

Article history:

Received Nov 9, 2024
Revised Aug 5, 2025
Accepted Nov 11, 2025

Keywords:

Academic performance prediction
Feature selection
Hybrid machine learning models
Learner traces
Online learning environments

ABSTRACT

Faced with the rise of online learning platforms, predicting learners' academic performance has become a major concern to personalize and enhance educational journeys. However, traditional predictive models struggle to effectively integrate emotional and social factors. This article introduces a hybrid predictive model that combines random forests (RF) for selecting the most relevant features and multiple regression (MR) to forecast academic performance. The data is sourced from three online learning platforms and encompasses both implicit traces (learner interactions and behaviors) and explicit traces (demographic characteristics). Following a selection and merging process, the final dataset comprises 1,003,392 records and 42 features, categorized into six types of indicators: cognitive, emotional, social, normative, contextual, and demographic. The results demonstrate that this hybrid model outperforms traditional approaches and other machine learning (ML) techniques in terms of predictive accuracy, achieving an R^2 of 0.9372 and a root mean square error (RMSE) of 0.1022. The incorporation of explicit and implicit traces helps better capture the intricate interactions among the different data dimensions, significantly enhancing prediction quality. This work represents a notable advancement in the field of academic performance prediction. It also sheds light on challenges associated with the increasing complexity of models, paving the way for future research to develop more generalizable approaches.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Jamal Eddine Rafiq
Laboratory of Artificial Intelligence and Complex Systems Engineering
École Nationale Supérieure d'Arts et Métiers, Hassan II University
150 Avenue Nile Sidi Othman, Casablanca 20670, Morocco
Email: rafiq.je@gmail.com

1. INTRODUCTION

Against this background, in the era of digital education, predicting learners' academic performance much more has become huge challenging, because of the emergence of e-learning systems, and their capacity to acquire large volume of data [1]. However, the digital traces of learners are heterogeneous and difficult to process. The literature review reveals a reality that the data for online learning is heterogeneous and complex, which presents tremendous difficulties for the traditional prediction models, and thus the new methods are required to be developed considering the advanced approaches such as machine learning (ML) [2]. Traditional predictive models that model only a handful of features have limitations to capture the richness and the variety of traces learners leave behind. Waheed *et al*. [3] note that typical prediction models, which

may only rely on simplistic features, such as demographics and grades, become inadequate to characterize the diversity and complexity of learner's interactions/behaviors in virtual learning environments. These models find it difficult to integrate social interactions and emotional states, however, are crucial for precise academic performance predictions. Against this background, one key question is: how can a hybrid model better predict the learners' academic performance in the online learning by integrating both the explicit and implicit traces?

Khan and Ghosh [4] emphasize that the diversity of data generated by students in online learning environments, such as forum interactions, participation in activities, and engagement behaviors, cannot be fully leveraged by traditional predictive models. By focusing on a limited set of variables, traditional predictive models often miss the subtleties of the traces left by learners, thereby limiting their predictive effectiveness, especially in complex learning environments. In this context, it is essential to explore more robust approaches that integrate multiple dimensions of learner behavior and characteristics. Recent literature highlights the importance of considering multiple indicators: social, cognitive, emotional, normative, contextual, and demographic to better understand and predict academic performance [5]. However, several studies [6]–[8] combine these diverse dimensions within a unified model capable of handling both explicit (demographic data) and implicit (social, cognitive, emotional, normative, and contextual data) traces.

This article aims to propose and evaluate a hybrid model for predicting learners' academic performance by fully leveraging the potential of explicit and implicit data. The presented hybrid predictive model is based on advanced ML techniques and incorporates 42 features derived from six types of indicators. Our model is built from data collected on three online learning platforms, representing a wide range of learner behaviors and characteristics. Unlike previous approaches, this approach simultaneously exploits both explicit and implicit traces of learners, offering a more nuanced and systemic analysis of their academic performance.

This work stands out for the model's ability to grasp the complex interactions among various data dimensions through ML techniques, representing a significant advancement in predicting academic performance by surpassing the limitations of traditional approaches, particularly in terms of accuracy. With an R^2 of 0.9372, the model explains 93.72% of the variance in academic performance, highlighting its notable effectiveness in predicting academic outcomes through the combination of feature selection and multiple regression (MR). By fully leveraging data from interactions between learners and platforms, this work highlights the model's effectiveness in predicting learners' academic performance while opening up new perspectives for designing more tailored and personalized educational tools.

2. LITERATURE REVIEW

2.1. Predictive models of academic performance

The prediction of learners' academic performance is a widely studied research area, particularly in the context of online learning environments. Traditional predictive models primarily rely on linear regression techniques and ML methods such as random forests (RF), decision trees (DT), and support vector machines (SVM), as shown in Table 1. These approaches have proven effective in certain cases, especially when leveraging features such as previous grades, interactions with the platform, or learners' demographic profiles. However, traditional models have some limitations, especially considering the interoperability between the different dimensions of data. Furthermore, these studies are based mainly on a few features, which limits the prediction accuracies. For example, recent efforts demonstrate that models based primarily on past academic performance or demographic characteristics do not account for the learner behaviors in online learning environments [9].

Table 1. Relevant studies classified by type of indicators and predictive algorithm used

Type of indicators	Algorithms	Studies
Emotional	CNN, RF, FDN, KNN, DT	[6], [10]–[12]
Social	SVR, KNN, RF, MR, SVM, DT, ANN, CNN, NB	[5], [6], [8], [10], [13]–[15]
Cognitive	KNN, NB, CNN, LR, RF, ANN, FDN, DT, SVR, SVM	[5], [7], [16]–[21]
Normative	DT, ANN, KNN, SVM, NB, RF, FDN, DNN	[5], [6], [22], [23]
Contextual	DT, ANN, KNN, SVM, NB, DNN	[5], [22], [24]
Demographic	DT, VSM, NB, KNN, RF	[8], [7], [25], [26]

Note: CNN=convolutional neural network; RF=random forests; FDN=feedforward deep network; KNN=k-nearest neighbors; DT=decision tree; SVR=support vector regression; MR=multiple regression; SVM=support vector machine; ANN=artificial neural network; NB=Naive Bayes; LR=logistic regression; DNN=deep neural network; VSM=vector space model

2.2. Integration of implicit and explicit traces

Implicit traces, such as the perceptible interaction, emotions expressed, and engagement behavior, have great potential in predicting the academic performance. Recent work has underlined the impact of incorporating these implicit traces into explicit data, which in turn leads to improved, more faithful models. For example, Rizvi *et al.* [8] introduced an approach that forecasted occupation status with the use of social traces and demographic features, and obtained a significantly higher prediction score accuracy. Likewise, the inclusion of affective data into the prediction models, is consistently acknowledged as an essential element in the comprehension of learner's engagement [27], [28]. Research by MacCann *et al*. [9] demonstrated that adding emotional indicators into predictor models could increase the prediction of academic performance by more than 15%. By employing RF for feature selection along with MR methodologies, it was independently demonstrated that this method well models emotional and behavioral data in student and enhances academic performance prediction. For example, Said *et al*. [29] showed the superior performance of RF in choosing informative feature for academic performance. They also concluded that applying simpler techniques (e.g., MR) gives results akin to more sophisticated models and is less demanding in terms of resources. This approach does not only facilitate the interpretation of the important factors affecting learners' achievements, but also reduces computational requirements, which makes the method more applicable in resource-poor settings.

2.3. Hybrid approaches in predictive models

While deep learning techniques have gained popularity for predicting academic performance, especially due to their ability to model complex nonlinear relationships, they require significant computational resources and sometimes pose interpretation challenges. In contrast, hybrid approaches combining more traditional methods, such as RF for feature selection and MR for prediction, have recently demonstrated their effectiveness in modeling multidimensional data, including emotional and behavioral traces. For instance, Said *et al.* [29] demonstrated that the use of RF effectively selects the most relevant variables for academic performance while minimizing the risk of overfitting. Integrating these features into an MR model not only yields results comparable to deep learning models but also significantly reduces model complexity and resource requirements. These approaches thus offer a more transparent interpretation of the factors influencing learners' performance while ensuring precise and accessible predictions in resource-constrained environments.

3. CONCEPTUAL FRAMEWORK AND RESEARCH OPPORTUNITIES

Although traditional approaches, such as RF and MR, have proven effective in feature selection and predicting academic performance, several challenges remain. Current models sometimes struggle to generalize their results when data is sourced from various learning platforms or when complex interactions between different types of data are involved [30]. Furthermore, few studies have truly explored the combined integration of multiple types of indicators such as social, cognitive, emotional, normative, contextual, and demographic indicators within a single predictive model using simple and accessible methods. Most research either focuses on a limited set of features or relies on more complex models, like neural networks, which require significant computational resources [31]–[33].

This article aims to fill this gap by proposing a hybrid model based on RF for feature selection and MR for predicting academic performance. This approach, while relatively lightweight in terms of computation, effectively models multidimensional data and offers a clear interpretation of the factors influencing student outcomes in online learning environments. By emphasizing simplicity and efficiency, this approach meets the needs of resource-limited systems while ensuring accuracy comparable to more complex methods.

In fact, our approach is based on the theoretical framework of multimodal learning, which posits that academic performance is influenced by a complex interaction of cognitive, social, emotional, normative, contextual, and demographic factors. According to this perspective, each dimension of learning traces (implicit and explicit) uniquely contributes to the construction of knowledge and learner engagement [34]. Drawing on Bandura's social learning theories [35] and Pintrich's cognitive engagement models [36], our work explores how these diverse dimensions interact to impact academic success in online learning environments.

4. METHOD

This research work followed a hybrid model-based strategy using RF feature selection and MR to predict academic performance. Combining three online learning environments, this approach captures and models multi-dimensional data while taking into account diverse traces (social, cognitive, emotional, normative, contextual, and demographic). In this section, we describe the main stages of our work.

4.1. Data description

We used a dataset constructed from three distinct online learning platforms, each offering a variety of courses and learning activities. The collected data includes learner traces from diverse educational environments, allowing for a wide range of behaviors and characteristics to be captured. In total, the dataset comprises 42 features, grouped into six types of indicators: social, cognitive, emotional, normative, contextual, and demographic. The demographic data includes information such as age, gender, ethnicity, participation in sports, music listening habits, and parental education level. Implicit traces include measures of social engagement (number of messages posted by the learner in forums, number of messages viewed by others, number of collaborative activities, number of activities related to discussion forums, individualized support, and parental support); cognitive engagement (number of approved assessments in course units, average grade achieved, number of resources explored, number of chapters explored, number of quiz-type activities, number of additional activities, number of volunteer works, and number of off-course quiz-type activities); emotional state (assessed from interactions on the platform) [34]; normative data covering aspects such as the number of active days, weekly study time, and absences; and finally, contextual indicators integrating data on types of activities followed and their frequency (number of repeated activities, number of content pages viewed, consultation of the platform's homepage, number of videos explored, participation in questionnaires or surveys, total number of events, sum of clicks made by the learner, interactions with online educational content, contributions to an educational wiki, consultation of a shared sub-page, participation in online video conferencing activities, activity related to data manipulation, number of interactive activities based on HTML content, number of accesses to an external resource via a URL, consultation of a specific sub-page, participation in a glossary, interaction with a content folder, and consultation of an activity involving double discussions).

4.2. Data pre-processing and analysis

Before training the predictive model, rigorous data pre-processing was essential. This process involved several key steps. First, data cleaning was performed, which included removing missing or erroneous values, handling duplicates, and managing outliers. Next, normalization was applied to numerical variables to ensure that each feature had a balanced influence on the model; values were scaled to a common range, typically between 0 and 1. Finally, categorical features such as ethnicity, gender, and parental education level were converted into numerical format using techniques such as one-hot encoding, allowing them to be effectively integrated into the ML model.

A descriptive analysis of the data was conducted to understand the distribution of variables and identify initial correlations between them. The data presented in Figure 1 highlights significant correlations between certain variables and academic performance, measured here by the grade point average (GPA), which represents the weighted average of grades obtained by a student. Specifically, strong correlations were observed with variables such as cumulative units evaluated approved (CUEvA), cumulative units grade (CUGrade), the number of chapters explored (nchapters), the number of videos watched (nplay_video), as well as other measured learning activities. However, it is worth noting that demographic indicators, such as gender, parental support, and parental education level, show significant correlations with GPA, thus highlighting the importance of considering these variables in improving predictions.

4.3. Proposed hybrid model

The proposed predictive model, as illustrated in Figure 2, is a hybrid model combining feature selection based on RF and MR. Our model combines the advantages of RF and MR to predict students' academic performance. RF are used to select the most important features based on their degree of importance, with a threshold value of 0.5 or higher [37]. This threshold represents the midpoint of the feature importance scale, which ranges from 0 to 1, thus forming a natural cutoff between features with a significant impact and those with lesser influence. A threshold of 0.5 corresponds to features with at least a moderate or higher effect on academic performance, filtering out less impactful variables while retaining the most relevant ones. These selected features are then used as inputs for an MR model, which aims to capture the complex relationships between these features and academic performance. The MR model uses a cost function based on mean squared error to produce accurate predictions. By reducing the dataset during data testing, the model ensures its robustness and precision. The main objective of this hybrid model is to improve the accuracy of predictions of academic performance by leveraging the most relevant features identified by RF.

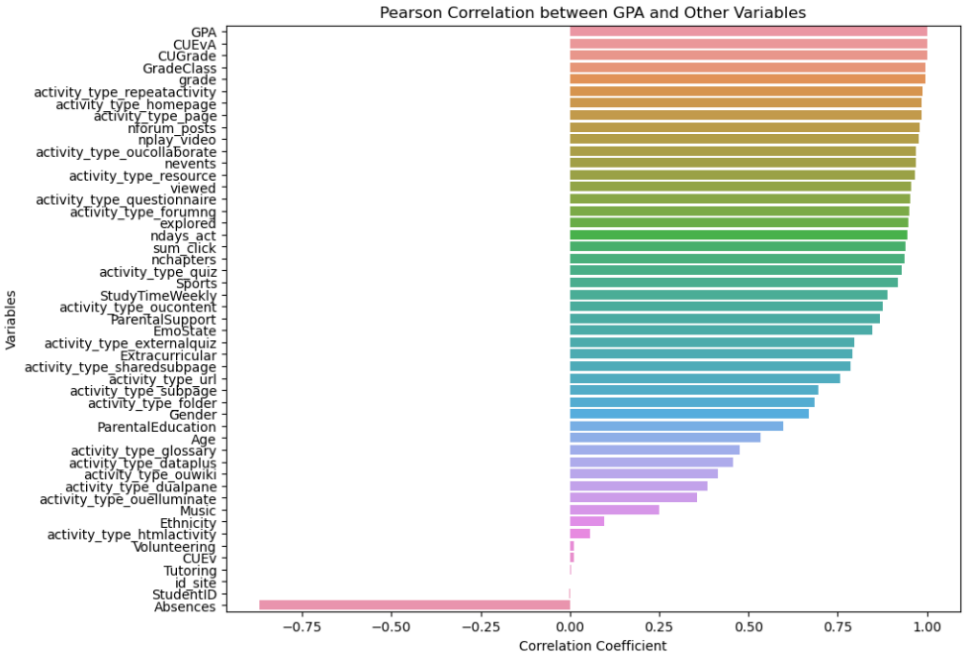

Figure 1. Pearson correlation coefficients with the GPA variable

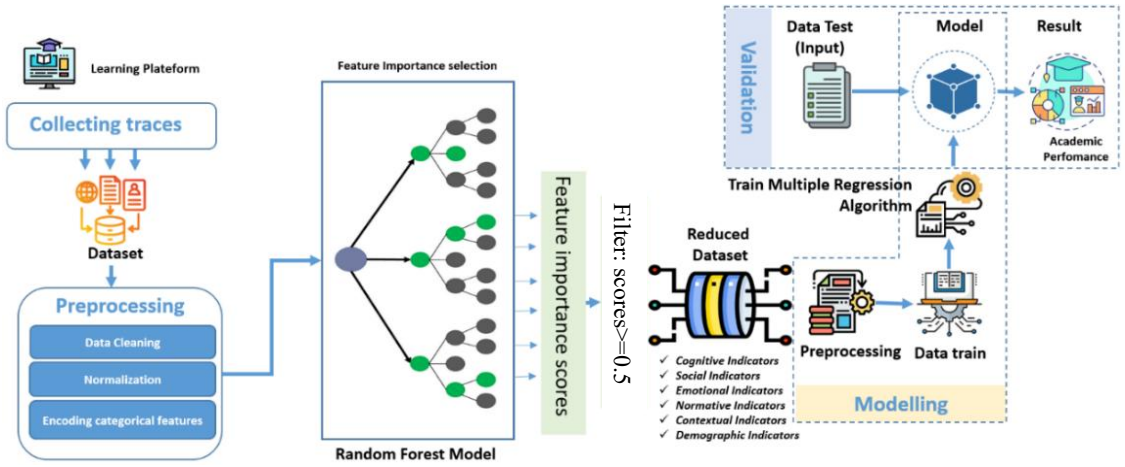

Figure 2. Flow of predictive model implementation

4.4. Model evaluation

To evaluate the performance of our hybrid model, we used several regression-appropriate metrics. The mean squared error or MSE was used to measure the discrepancy between predicted and actual values, while k-fold cross-validation ensured a robust evaluation by dividing the data into k subsets. The model utilizes an RF to select the most important features, followed by MR for the final prediction, thereby optimizing result accuracy while reducing overfitting.

To assess the performance of our hybrid model, we used several regression metrics. The coefficient of determination (R^2) measured the proportion of variance in the dependent variable that could be explained by the model (1), indicating its explanatory power. The mean squared error (MSE) and root mean squared error (RMSE) were used to quantify the accuracy of the predictions by evaluating the differences between the actual and predicted values, as in (2) and (3). Additionally, the mean absolute error (MAE) offered a more straightforward interpretation of the average deviation between predicted and actual outcomes, as in (4). To enhance the model's robustness and minimize the risk of overfitting, we applied k-fold cross-validation, as illustrated in Figure 3, which involved testing the model on multiple subsets of the data. The hybrid model integrates an RF for selecting the most relevant features and an MR approach to refine predictions, thereby increasing accuracy while avoiding overfitting.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \quad (2)$$

$$RMSE = \sqrt{MSE} \quad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \quad (4)$$

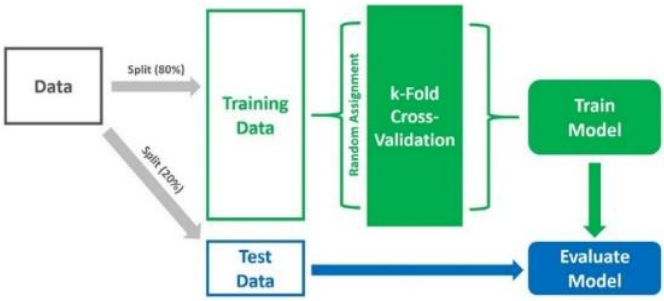

Figure 3. K-folds cross validation

5. RESULTS AND DISCUSSION

To evaluate the most influential features for predicting academic performance, our study applied an RF model to calculate importance scores. Features with a score equal to or greater than 0.5 were filtered, effectively reducing the dataset's dimensionality while retaining the most relevant variables. Figure 4 visually presents these input features, while Table 2 details the characteristics identified as the most significant. This approach has shed light on the key factors influencing academic performance and will serve as the foundation for our subsequent analyses.

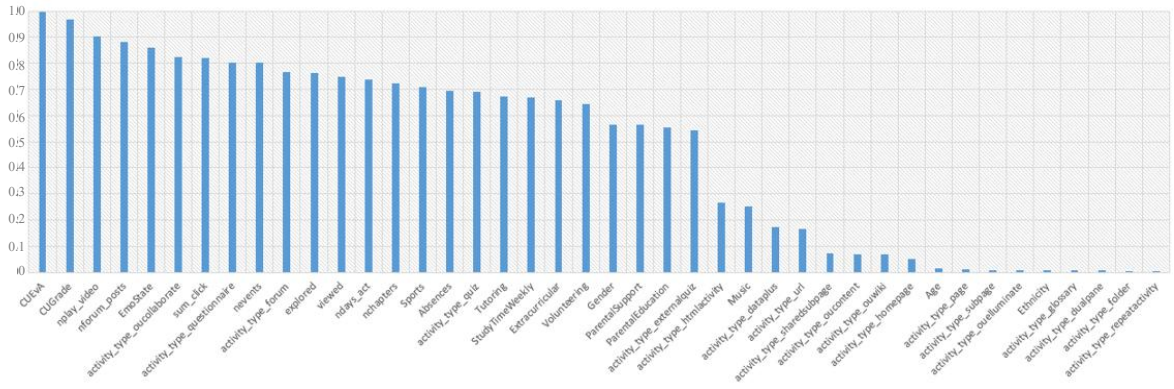

Figure 4. Graphical representation of feature importance scores

The MR phase, applied to the selected features, demonstrated an exceptional ability to capture the complex relationships between these variables and academic performance (GPA). The results indicate remarkable accuracy of the model, as evidenced by a coefficient of determination R^2 of 0.9372, suggesting that the model explains 93.72% of the variance in academic performance. A study by Baashar *et al*. [38] achieved an R^2 of 0.89 using an artificial neural network (ANN) model, while our hybrid model reached 0.9372, indicating a substantial improvement in predictive power. Furthermore, the prediction errors are extremely low, with a MSE of 0.0104, a RMSE of 0.1022, and a MAE of 0.0817. These metrics show a significant improvement over baseline models, as indicated in Table 3, highlighting the outstanding performance of the hybrid model in predicting academic performance. Moreover, this study stands out for integrating six types of indicators: cognitive, social, emotional, contextual, normative, and demographic, in contrast to previous works [5], [6], [8], which included only two, three, or four types of dimensions. These results highlight the remarkable performance of the hybrid model in predicting academic performance,

demonstrating the effectiveness of the approach that combines feature selection with MR. This synergy enhances prediction accuracy while reducing overall errors. As shown in Figure 5, the comparison of predicted and actual GPA values reflects this accuracy, with data points closely aligned around the ideal regression line, confirming the precision of the predictions.

Table 2. Most significant features (importance score ≥0.5)

Index	Feature	Importance
1	CUEvA	0.99866980695132
2	CUGrade	0.96722020595401
3	nplay_video	0.90359167229051
4	nforum_posts	0.88130615331932
5	EmoState	0.86025487655442
6	activity_type_oucollaborate	0.82612264023988
7	sum_click	0.82091760343514
8	activity_type_questionnaire	0.80424324553426
9	nevents	0.80148122191554
10	activity_type_forum	0.76780796050704
11	explored	0.76296359685862
12	viewed	0.74971104991562
13	ndays_act	0.73695945514967
14	nchapters	0.72486778905941
15	Sports	0.71094928704572
16	Absences	0.69394259333360
17	activity_type_quiz	0.69175882932148
18	Tutoring	0.67313543360429
19	StudyTimeWeekly	0.67033847650382
20	Extracurricular	0.66063171306639
21	Volunteering	0.64558441972441
22	Gender	0.56484258363454
23	ParentalSupport	0.56409859717801
24	ParentalEducation	0.55364603940077
25	activity_type_externalquiz	0.54466617937774

Table 3. Performance of ML algorithms

Algorithm	R^2	MSE	RMSE	MAE
SVM	0.9094	0.0152	0.1234	0.0970
MR	0.9097	0.0113	0.1063	0.0874
RF	0.9096	0.0121	0.1100	0.0881
Hybrid model	0.9372	0.0104	0.1022	0.0817

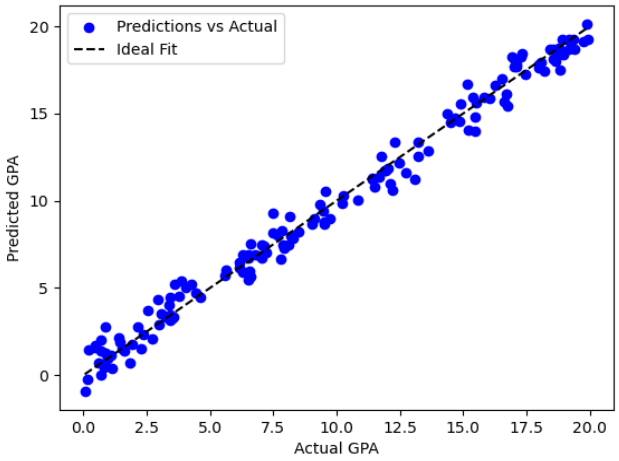

Figure 5. Regression multiple model: predicted vs actual GPA

The RF, when used for feature selection, proved to be an essential step while developing the model, as identified from the analysis of the results. As seen in Figure 6, some variables such as CUEvA, CUGrade, video viewing (nplay_video), social interactions, and forum activity (nforum_posts) seem to have an intriguing association with GPA. The 3D graphs indicate that these variables have a positive correlation with the academic performance of the learners: Figure 6(a) supports the idea that CUEvA and CUGrade are

positively related to performance and Figure 6(b) exhibits the relation between CUEvA, nplay video, and GPA. Figure 6(c) is the portrayal of the state where a participant's EmoState is positive and simultaneously active on the forum which is expected to enhance GPA greatly. These outcomes are consistent with the findings of our prior research [28], which ascertained the elevated importance of forum participation and emotional state during GPA prediction. As shown in Figure 6(d), interaction was found to have a positive impact on collaborative activity, including the number of clicks, and achieved GPA, which confirms the previously stated hypothesis.

According to Al-Zawqari *et al*. [39], feature selection using ML techniques leads to better model interpretability and performance, corroborating our results. Furthermore, engagement in extracurricular activities and parental support were also identified as key determinants of academic performance, thus supporting our conclusions on the importance of these characteristics [14], [40]. These findings collectively reinforce the importance of these factors in the context of online learning and underscores the need to consider them when evaluating learners' academic performance.

The performance of the regression model was assessed using k-fold cross-validation to measure its robustness and ability to generalize. Table 4 is for the results of the 5-fold cross-validation and Table 5 is for the results of the 10-fold cross-validation. In 5-fold cross-validation, the mean MSE across all folds was 0.01326, with a very low standard deviation of 0.0047. This indicates that the model's performance is consistent and stable across different subsets of the data. In this sense, Owusu-Boadu *et al*. [26] emphasize the importance of cross-validation in evaluating the robustness of models, which corroborates our results of consistent performance across the different folds. For 10-fold cross-validation, the mean MSE was also 0.01367, with a slightly higher standard deviation of 0.0041. The consistency of the mean MSE in both 5-fold and 10-fold validations, as well as the low standard deviations, underscore the model's reliability. These results suggest that the model maintains good accuracy even when evaluated on different subsets of the data, thereby reinforcing its robustness and reliability.

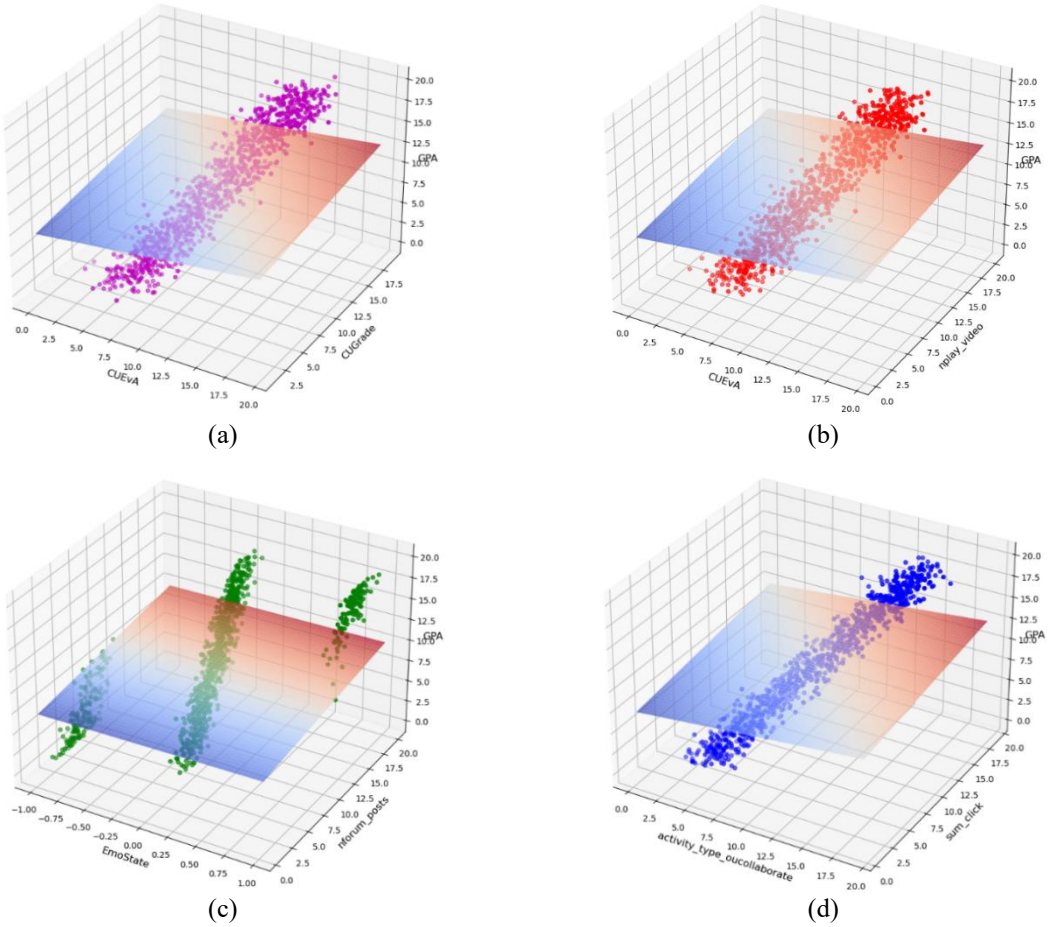

Figure 6. Four perspectives of a 3D visualization of the multiple linear regression model: (a) CUEvA vs CUGrade with respect to GPA, (b) CUEvA vs nplay_video with respect to GPA, (c) EmoState vs nforum_posts with respect to GPA, and (d) Collaborative activity vs sum_click with respect to GPA

Table 4. Evaluation outcomes using 5-fold cross-validation

Fold	1	2	3	4	5
MSE	0.0111	0.0112	0.0205	0.0121	0.0114

Mean MSE (k=5): 0.01326
Standard deviation of MSE (k=5): 0.0047

Table 5. Evaluation outcomes using 10-fold cross-validation

Fold	1	2	3	4	5	6	7	8	9	10
MSE	0.0112	0.0124	0.0203	0.0157	0.0165	0.0232	0.0151	0.0108	0.0146	0.0119

Mean MSE (k=10): 0.01367; Standard deviation of MSE (k=10): 0.0041

One of the main contributions of this research lies in the integration of both explicit and implicit traces into a single predictive model. Unlike previous studies [5], [41], which were limited to the use of four types of indicators, as well as the study by Al-Tameemi [6], which was limited to three types of indicators, our model includes six categories of data, encompassing cognitive, emotional, social, normative, contextual, and demographic indicators. This diversity in predictive indicators allows for a better understanding of the complexity of learner behaviors and enhances the accuracy of predictions. This advancement highlights the importance of enriching models with multiple and varied factors to better understand students' academic performance.

However, although the hybrid model provides excellent results, it relies on simplifying assumptions about the relationships between variables. The complex and nonlinear interactions between certain features may make the interpretation of the model more challenging. While the proposed approach is robust, future research could explore alternative models, such as deep neural networks, to capture more subtle relationships between variables [39]. A potential limitation of our model lies in the absence of a temporal component, which could capture the evolution of learner behaviors over time. Furthermore, although feature selection helped reduce the dimensionality of the dataset, some non-selected variables may also impact the results and deserve further exploration to improve the model's accuracy.

One of the strengths of our methodology is the integration of advanced ML techniques, such as RF, for the selection of predictive variables. While many studies [6], [16], [25] omit details on their selection method, our approach relies on rigorous automatic feature selection techniques. This approach has allowed us to identify the most relevant variables while reducing the dimensionality of the dataset. Indeed, RF are widely recognized as one of the most popular methods in data science for feature selection, providing a solid foundation for our methodology [42].

The results obtained from this study have significant implications for online education. They can guide educators in implementing more personalized tracking systems, allowing them to identify at-risk learners based on their interactions. The integration of explicit and implicit factors into predictive models can help better understand the variables influencing academic success and thus personalize educational pathways. In the future, it would be beneficial to validate this model on more diverse datasets, including cultural interactions and varied learning methodologies. Moreover, the inclusion of temporal data would provide a finer analysis of the behavioral evolution of learners. The proposed hybrid approach, combining feature selection by RF and MR, could also enable real-time identification of at-risk students and provide them with targeted support, thus fostering a proactive and effective teaching approach.

6. CONCLUSION

The advent of artificial intelligence is disrupting traditional teaching methods, paving the way for a new era of online learning. This study proposes a hybrid model combining feature selection by RF and MR to predict academic performance online, optimizing variable relevance while reducing complexity. The distinctive specificity of this model lies in the joint integration of indicators belonging to six complementary dimensions (cognitive, emotional, social, normative, contextual, and demographic), extracted from both explicit and implicit traces stemming from learner interactions. This multidimensional approach aims to be more comprehensive than those limited to three or four dimensions, although it could still be enriched by adding other factors, whether or not they pertain to additional dimensions. Such an extension would undoubtedly allow for an even more faithful and nuanced modeling of the factors influencing learners' academic performance. Additionally, the inclusion of a selective module based on ML techniques enhances the model's performance by identifying the most influential features while effectively reducing data complexity.

The proposed predictive model, integrating six types of indicators, significantly improves academic performance prediction with an R^2 of 0.9372 (93.72%), thereby outperforming the 89% achieved by

compared ANN models, as well as the performances of SVM, MR, and RF, while substantially reducing prediction errors (MSE=0.0104, RMSE=0.1022, MAE=0.0817). The generalization capability of our regression model was confirmed through k-fold cross-validation, with similar average MSEs (0.01326 for 5-fold with a standard deviation of 0.0047, and 0.01367 for 10-fold with a standard deviation of 0.0041), thus demonstrating stable and reliable performance across different data subsets. These results open promising perspectives for a better understanding of the determinants of academic success, while providing teachers and instructional designers with more effective predictive tools to anticipate learners' needs and support them in a personalized manner.

The proposed approach surpasses the limitations of traditional methods by offering a solution better suited to the specificities of complex and heterogeneous datasets encountered in online learning environments. For future research, it would be relevant to explore the addition of new dimensions or supplementary indicators likely to enrich the model and further improve prediction accuracy. The application of more advanced learning techniques, including hybrid models or those derived from generative artificial intelligence, as well as testing in diverse educational contexts, would also allow a more refined evaluation of the robustness and generalizability of our approach.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Jamal Eddine Rafiq	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Zakrani Abdelali				✓	✓	✓	✓			✓		✓		
Mohammed Amraouy				✓	✓	✓	✓			✓				
Said Nouh				✓	✓	✓	✓			✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data presented in this study are available on request from the corresponding author, [JER]. The data is not publicly available due to the risk of identification of study participants.

REFERENCES

[1] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H. Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: a survey study," *Education and Information Technologies*, vol. 28, no. 1, pp. 905–971, 2023, doi: 10.1007/s10639-022-11152-y.

[2] T. T. Khoei, H. Ould Slimane, and N. Kaabouch, "Deep learning: systematic review, models, challenges, and research directions," *Neural Computing and Applications*, vol. 35, no. 31, pp. 23103–23124, 2023, doi: 10.1007/s00521-023-08957-4.

[3] 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, Mar. 2020, doi: 10.1016/j.chb.2019.106189.

[4] A. Khan and S. K. Ghosh, "Data mining based analysis to explore the effect of teaching on student performance," *Education and Information Technologies*, vol. 23, no. 4, pp. 1677–1697, 2018, doi: 10.1007/s10639-017-9685-z.

[5] J. Malini and Y. Kalpana, "Investigation of factors affecting student performance evaluation using education materials data mining technique," *Materials Today: Proceedings*, vol. 47, pp. 6105–6110, 2021, doi: 10.1016/j.matpr.2021.05.026.

[6] G. Al-Tameemi, J. Xue, I. H. Ali, and S. Ajit, "A hybrid machine learning approach for predicting student performance using multi-class educational datasets," *Procedia Computer Science*, vol. 238, pp. 888–895, 2024, doi: 10.1016/j.procs.2024.06.108.

[7] K. Vimarsha, S. P. S. Prakash, K. Krinkin, and Y. A. Shichkina, "Student performance prediction: a co-evolutionary hybrid

intelligence model," *Procedia Computer Science*, vol. 235, pp. 436–446, 2024, doi: 10.1016/j.procs.2024.04.043.
[8] S. Rizvi, B. Rienties, and S. A. Khoja, "The role of demographics in online learning; a decision tree based approach," *Computers & Education*, vol. 137, pp. 32–47, 2019, doi: 10.1016/j.compedu.2019.04.001.
[9] C. MacCann, Y. Jiang, L. E. R. Brown, K. S. Double, M. Bucich, and A. Minbashian, "Emotional intelligence predicts academic performance: a meta-analysis," *Psychological Bulletin*, vol. 146, no. 2, pp. 150–186, Feb. 2020, doi: 10.1037/bul0000219.
[10] G. Wang and T. Ren, "Design of sports achievement prediction system based on U-net convolutional neural network in the context of machine learning," *Heliyon*, vol. 10, no. 10, p. e30055, 2024, doi: 10.1016/j.heliyon.2024.e30055.
[11] B. Cheng, Y. Liu, and Y. Jia, "Evaluation of students' performance during the academic period using the XG-Boost classifier-enhanced AEO hybrid model," *Expert Systems with Applications*, vol. 238, p. 122136, 2024, doi: 10.1016/j.eswa.2023.122136.
[12] M. Amraouy, M. Bellafkih, A. Bennane, and J. Talaghzi, "Sentiment analysis for competence-based e-assessment using machine learning and lexicon approach," in *The 3rd International Conference on Artificial Intelligence and Computer Vision (AICV2023)*, 2023, pp. 327–336, doi: 10.1007/978-3-031-27762-7_31.
[13] H. Waheed, S. U. Hassan, R. Nawaz, N. R. Aljohani, G. Chen, and D. Gasevic, "Early prediction of learners at risk in self-paced education: a neural network approach," *Expert Systems with Applications*, vol. 213, p. 118868, 2023, doi: 10.1016/j.eswa.2022.118868.
[14] K. S. Manic, A. S. Al-Bemani, A. A. Nizamudin, G. Balaji, and A. A. Amal, "Optimizing academic journey for high schoolers in Oman: a machine learning-enabled AI model," *Procedia Computer Science*, vol. 235, pp. 2716–2729, 2024, doi: 10.1016/j.procs.2024.04.256.
[15] Á. Hernández-García, C. Cuenca-Enrique, L. del-Río-Carazo, and S. Iglesias-Pradas, "Exploring the relationship between LMS interactions and academic performance: a learning cycle approach," *Computers in Human Behavior*, vol. 155, p. 108183, 2024, doi: 10.1016/j.chb.2024.108183.
[16] P. X. Lam, P. Q. H. Mai, Q. H. Nguyen, T. Pham, T. H. H. Nguyen, and T. H. Nguyen, "Enhancing educational evaluation through predictive student assessment modeling," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100244, 2024, doi: 10.1016/j.caeai.2024.100244.
[17] X. Wang, Y. Zhao, C. Li, and P. Ren, "ProbSAP: a comprehensive and high-performance system for student academic performance prediction," *Pattern Recognition*, vol. 137, p. 109309, 2023, doi: 10.1016/j.patcog.2023.109309.
[18] C. F. Rodríguez-Hernández, M. Musso, E. Kyndt, and E. Cascallar, "Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100018, 2021, doi: 10.1016/j.caeai.2021.100018.
[19] E. Muratov, M. Lewis, D. Fourches, A. Tropsha, and W. C. Cox, "Computer-assisted decision support for student admissions based on their predicted academic performance," *American Journal of Pharmaceutical Education*, vol. 81, no. 3, p. 46, Apr. 2017, doi: 10.5688/ajpe81346.
[20] P. R., K. P., and S. A. A., "Predicting academic performance of learners with the three domains of learning data using neuro-fuzzy model and machine learning algorithms," *Journal of Engineering Research (Kuwait)*, vol. 12, no. 3, pp. 397–411, 2024, doi: 10.1016/j.jer.2023.09.006.
[21] M. Chen and Z. Liu, "Predicting performance of students by optimizing tree components of random forest using genetic algorithm," *Heliyon*, vol. 10, no. 12, p. e32570, 2024, doi: 10.1016/j.heliyon.2024.e32570.
[22] J. Deng, X. Huang, and X. Ren, "A multidimensional analysis of self-esteem and individualism: a deep learning-based model for predicting elementary school students' academic performance," *Measurement: Sensors*, vol. 33, p. 101147, 2024, doi: 10.1016/j.measen.2024.101147.
[23] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, "Student academic performance prediction using supervised learning techniques," *International Journal of Emerging Technologies in Learning*, vol. 14, no. 14, pp. 92–104, 2019, doi: 10.3991/ijet.v14i14.10310.
[24] R. N. Marishane, "Contextual intelligence for student learning," in *Contextual Intelligence in School Leadership*, R. N. Marishane, Ed., Leiden: Brill, 2020, pp. 147–177, doi: 10.1163/9789004431263_006.
[25] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student performance prediction model based on supervised machine learning algorithms," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 928, p. 032019, doi: 10.1088/1757-899X/928/3/032019.
[26] B. Owusu-Boadu, I. K. Nti, O. Nyarko-Boateng, J. Aning, and V. Boafo, "Academic performance modelling with machine learning based on cognitive and non-cognitive features," *Applied Computer Systems*, vol. 26, no. 2, pp. 122–131, 2021, doi: 10.2478/acss-2021-0015.
[27] U. Ononye, M. Ogbeta, F. Ndudi, D. Bereprebofa, and I. Maduemezia, "Academic resilience, emotional intelligence, and academic performance among undergraduate students," *Knowledge and Performance Management*, vol. 23, no. 3, pp. 1–10, 2022, doi: 10.7605/gdlxb.2021.03.037.
[28] J. E. Rafiq, M. Amraouy, A. Namir, A. Bennane, and A. Zakrani, "Optimizing learning performance through AI-enhanced discussion forums," in *2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA)*, 2023, pp. 1–6, doi: 10.1109/SITA60746.2023.10373689.
[29] M. B. Said, Y. H. Kacem, A. Algarni, and A. Masmoudi, "Early prediction of student academic performance based on machine learning algorithms: a case study of bachelor's degree students in KSA," *Education and Information Technologies*, vol. 29, no. 11, pp. 13247–13270, 2024, doi: 10.1007/s10639-023-12370-8.
[30] Y. Manzali, Y. Akhiat, K. A. Barry, E. Akachar, and M. E. Far, "Prediction of student performance using random forest combined with Naïve Bayes," *Computer Journal*, vol. 67, no. 8, pp. 2677–2689, 2024, doi: 10.1093/comjnl/bxae036.
[31] Z. Sun, Y. Yuan, X. Xiong, S. Meng, Y. Shi, and A. Chen, "Predicting academic achievement from the collaborative influences of executive function, physical fitness, and demographic factors among primary school students in China: ensemble learning methods," *BMC Public Health*, vol. 24, no. 1, p. 274, 2024, doi: 10.1186/s12889-024-17769-7.
[32] G. Zhu, P. Raman, W. Xing, and J. Slotta, "Curriculum design for social, cognitive and emotional engagement in knowledge building," *International Journal of Educational Technology in Higher Education*, vol. 18, no. 1, pp. 1–19, 2021, doi: 10.1186/s41239-021-00276-9.
[33] O. Noroozi, H. J. Pijeira-Díaz, M. Sobocinski, M. Dindar, S. Järvelä, and P. A. Kirschner, "Multimodal data indicators for capturing cognitive, motivational, and emotional learning processes: a systematic literature review," *Education and Information Technologies*, vol. 25, no. 6, pp. 5499–5547, 2020, doi: 10.1007/s10639-020-10229-w.
[34] M. Amraouy, M. Bellafkih, A. Bennane, and M. M. Himmi, "Online Competency-based assessment (OCBA): from conceptual model to operational authoring system," *International Journal of Interactive Mobile Technologies*, vol. 16, no. 4, pp. 46–57, 2022, doi: 10.3991/ijim.v16i04.28373.
[35] V. Koutroubas and M. Galanakis, "Bandura's Social learning theory and its importance in the organizational psychology context," *Journal of Psychology Research*, vol. 12, no. 6, pp. 315–322, 2022, doi: 10.17265/2159-5542/2022.06.001.

[36] P. R. Pintrich and B. Schrauben, "Students' motivational beliefs and their cognitive engagement in classroom academic tasks," in *Student Perceptions in the Classroom*, D. H. Schunk and J. L. Meece, Eds., Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1992, pp. 149–183, doi: 10.4324/9780203052532-13.
[37] R. Genuer, J. M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010, doi: 10.1016/j.patrec.2010.03.014.
[38] Y. Baashar *et al.*, "Evaluation of postgraduate academic performance using artificial intelligence models," *Alexandria Engineering Journal*, vol. 61, no. 12, pp. 9867–9878, 2022, doi: 10.1016/j.aej.2022.03.021.
[39] A. Al-Zawqari, D. Peumans, and G. Vandersteen, "A flexible feature selection approach for predicting students' academic performance in online courses," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100103, 2022, doi: 10.1016/j.caeai.2022.100103.
[40] I. Antonio-Agirre, A. Rodríguez-Fernández, and L. Revuelta, "Social support, emotional intelligence and academic performance in secondary education," *European Journal of Investigation in Health, Psychology and Education*, vol. 9, no. 2, pp. 109–118, 2019, doi: 10.30552/ejihpe.v9i2.324.
[41] J. E. Rafiq, A. Zakrani, M. Amraouy, S. Nouh, and A. Bennane, "Improving online learning using deep learning and student's intelligences," *Turkish Online Journal of Distance Education*, vol. 26, no. 2, pp. 39–52, Apr. 2025, doi: 10.17718/tojde.1477677.
[42] R. Iranzad and X. Liu, "A review of random forest-based feature selection methods for data science education and applications," *International Journal of Data Science and Analytics*, vol. 20, no. 2, pp. 197–211, 2025, doi: 10.1007/s41060-024-00509-w.

BIOGRAPHIES OF AUTHORS

Jamal Eddine Rafiq is a pedagogical inspector in computer science in Marrakech, Morocco, and a current Ph.D. student and affiliated with Laboratory of Artificial Intelligence and Complex Systems Engineering, Hassan II University. His research interests focus on human computer interaction, artificial intelligence, and digital learning traces. He can be contacted at email: Rafiq.je@gmail.com; jamal.rafiq-etu@etu.univh2c.ma.

Zakrani Abdelali holds a Ph.D. in computer sciences at Mohammed V University, Rabat, Morocco, in 2012. He is currently professor (Habilitation à Diriger des Recherche (HDR)) at ENSAM, Hassan II University, Casablanca, Morocco. His current research interests include artificial neural network, data mining, and software engineering. He can be contacted at email: abdelali.zakrani@univh2c.ma.

Mohammed Amraouy holds a Ph.D. in computer sciences at National Institute of Posts and Telecommunications, Rabat, Morocco in 2023. He is currently a pedagogical inspector in computer science and part-time trainer at Regional Center for Education and Training Profession Oujda, Morocco. His research interests focus on human computer interaction, artificial intelligence, and online learning assessment. He can be contacted at email: amraouy.mohamed1@gmail.com.

Said Nouh holds a Ph.D. in computer sciences at National School of Computer Science and Systems Analysis (ENSIAS), Rabat, Morocco in 2014. He is currently professor (Higher degree research (HDR)) at Faculty of Sciences Ben M'Sick, Hassan II University, Casablanca, Morocco. His current research interests are artificial intelligence, machine learning, deep learning, telecommunications, information, and coding theory. He can be contacted at email: said.nouh@univh2m.ma.