IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 22 April 2023, accepted 26 May 2023, date of publication 20 June 2023, date of current version 13 July 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3287972*

RESEARCH ARTICLE

# Machine Learning in Higher Education: Students' Performance Assessment Considering Online Activity Logs

**GHAZANFAR LATIF[1,2], SHERIF E. ABDELHAMID[3], KHALED S. FAWAGREH[1], GHASSEN BEN BRAHIM[1], AND RUNNA ALGHAZO[4]**

[1]Department of Computer Science, Prince Mohammad Bin Fahd University, Khobar 31952, Saudi Arabia
[2]Department of Computer Sciences and Mathematics, Université du Québec à Chicoutimi, Chicoutimi, QC G7H 2B1, Canada
[3]Department of Computer and Information Sciences, Virginia Military Institute, Lexington, VA 24450, USA
[4]Department of Education, Health, and Behavioral Studies (EHBS), University of North Dakota, Grand Forks, ND 58202, USA

Corresponding author: Ghazanfar Latif (glatif@pmu.edu.sa)

This work was supported in part by the Commonwealth Cyber Initiative, an investment in the advancement of cyber research and development, innovation, and workforce development.

**ABSTRACT** Machine learning in Education is receiving more attention from researchers as the number of students at all levels globally is increasing. To ensure students' success in K-12 educational institutions and higher education institutions work needs to be done to assist students, teachers/professors, parents, and all stakeholders to provide the support that students need. The need and motivation for such systems are very well-established and thus the aim of this work is to develop a system based on modified machine learning models to automatically predict students' performance and subsequently identify students at risk. The DEEDs dataset is used in this study. Novel features were extracted and applied to well-known classifiers some of which are ensemble classifiers. These classifiers were also combined with base learners such as bagging and boosting. The problem was divided into three scenarios; binary classification of the pass and fail, three class scenarios, and four class scenarios. It was shown that ensemble methods combined with base learners of boosting and bagging significantly increase the accuracy for binary classification, slightly increase accuracy for three class problems, and have no significance in increasing the accuracy when the problem is 4 classes. The ensemble algorithm of bagging and boosting FDT achieved an accuracy of 98.25% for binary classification and 89.47% for three classes. The standard ensemble FDT achieved an accuracy of 77.19% for four classes. The results obtained for binary classification were compared with results reported in the extant literature using the same dataset proving that the proposed modified algorithms achieved better results than similarly proposed methods. The three-class and four-class results could not be compared because according to the author's knowledge, there are no research papers published for the same dataset for multi-class classification.

**INDEX TERMS** Automation in higher education, ensemble learning, machine learning, multiclass grade prediction, online activities, student performance prediction.

## I. INTRODUCTION

Due to the recent attention of researchers to the field of machine learning and its practical use in various fields to benefit stakeholders including decision-makers, researchers are now in the process of proposing the use of machine learning to benefit humankind in various applications. For instance, in the recent Covid 19 pandemic, researchers have proposed the use of machine learning to ease the burden on medical institutions in diagnosing positive cases of Covid [1] because medical institutions globally were running low on resources and were overwhelmed. In a different domain, machine learning was also applied to support establishing smart cities [2]. Another area that is also drawing a lot of

The associate editor coordinating the review of this manuscript and approving it for publication was Massimo Cafaro.

attention is the education sector aiming at enhancing teaching outcomes, which will be the scope of this current research work.

Machine learning and data mining can give us helpful information and insight into predicting students' performance during their studies. With this information, we can help students perform better, and guide them in the right direction by making correct and sensible decisions in a timely manner. With the help of Machine Learning and AI methods, we store different data such as previous grades, to predict students' performance with great accuracy [3], [4], [5], [7]. The application of Data Mining techniques in the educational sector has increased over the years, and many universities have been collecting massive data related to students' performance over the years for the sake of analysis and hidden knowledge extraction [8]. Recently, researchers have been paying attention to the big data being collected and using them to create intelligent solutions to enhance the learning process for students. Some of the use of machine learning and AI in recent years are government-supported grade prediction and analysis programs. Some of the reasons behind such analysis is distributing scholarships among potential students who are most likely to perform well and succeed in their undergraduate studies. Since, however, it is costly for the government to offer everyone a scholarship, many governments have invested in machine learning and AI to predict students' success in order to decide on awarding the scholarship. Another use is to help guide students in the right direction, predict their performance outcome, and support them through intervention and tutoring before receiving their grades. As can be seen, predicting a student's academic standing at graduation time can be very useful in helping institutions select among candidates, or helping potentially weak students in overcoming educational challenges. Two approaches were considered in handling such problems both of which are supervised learning-based approaches. The first one uses students' past performance and the second one is entirely event-driven data based on students' activities and interactions with educational systems and resources. Both approaches rely on machine learning classification/regression models by having these models predict students' performance as a function of multiple input features. Student performance refers to the student's grades in individual courses and also indicates the student's progression in the curriculum study plan. Student assessment refers to any assessment-graded tasks given in a particular course such as homework, quizzes, exams, projects, etc. The need for systems based on machine learning in the education sector whether K-12 Education or higher education is becoming required to ease the teaching and learning process. Many publishers are integrating machine learning into their products to offer students material and assessments based on their individual needs and past performance. Machine learning application courses are offered as multidisciplinary courses across many fields to ensure that machine learning integrates within all fields. Machine learning is being integrated into class management systems, campus life, enrollment, and all others because these systems contain vast amounts of data that can be easily processed through machine learning to help the students and decision-makers as well. Within Academics, the performance and success of students in their chosen field and within their semester-long classes is important for students, faculty, administrators, parents, and other stakeholders. Intervention, when intervention is needed, would play a great role in ensuring that all students acquire the needed support to succeed. Even for students who are not deemed at risk, wouldn't it be great if a system can inform them what they need to concentrate on in order to obtain the grade they aspire to achieve. All this can be achieved if enough data is available and machine learning systems are put to work in predicting grades and further analysis to identify students' weaknesses and strengths. It is now well established that more work is required to integrate machine learning systems to analyze the vast data available for students to produce results that eventually help students and faculty members. This is the primary motivation for the current study.

The aim of this paper is to develop a Machine Learning based system that can predict students' performance and thereby identify students at risk. The system is able to do so based on a dataset consisting not only of students' grades on certain tasks but also consisting of students' behavior while performing certain tasks.

In this paper, we start by conducting an extensive literature review to identify the different machine learning-based methods that researchers have used to predict students' performance. We will also survey the various types of datasets that were used in the context of students' performance prediction, as well as the achieved accuracy. Our proposed method which is based on ensemble methods combined with base learners (weak learners) of boosting and bagging is presented next. We consider 3 different scenarios: binary classification cases where students are classified as pass or fail students, multi-classification cases where we initially consider three class scenarios, and then four class scenarios.

The novel contributions of this paper are briefly summarized as follows:

- We propose a modified machine learning model based on the ensemble method combined with base learners of boosting and bagging for student performance prediction and identify students at risk
- We propose the extraction of novel features from within the DEEDS dataset that has not been explored in previous literature according to our knowledge.
- We combine data that is grade-based and behavioral-based on the DEEDS published dataset.
- Our proposed model shows significant improvement in Binary and Multi-class classification and identification.
- The proposed model achieved higher accuracies and metric values than other similar research in the extant literature using the same dataset or similar datasets.

This paper is organized as follows. Previous related work in this area is outlined in Section II. The proposed methodology is detailed in Section III. Results and discussion of the proposed methodology are given in Section IV. Finally, the conclusion and future research directions are provided in Section V.

## II. RECENT STUDIES

Machine learning has gained research momentum over the last decade with the advancement of computing power and the development of complex graphical processing units. Machine learning has been applied in areas of healthcare [9], business [10], agriculture [11], stock market economy [12], and others. The use of machine learning in a specific area of education is also detailed in this section.

In [13], the authors proposed a system that will predict how likely a student might drop out. The authors suggested system benefits institutions in reducing student dropout, as it is a major concern in the education and policy-making communities. The authors used three machine learning models (regularized logistic regression, k-nearest neighbors, and random forest) to predict the binary dropout variable. The author's dataset consisted of 32,500 student transcript records from the largest public universities. Their proposed system achieved 66.59% accuracy for logistic regression, 62.24% for random forests, and 64.60% for k-nearest neighbors. In [14], the authors proposed developing a system that predicts students' performance. The authors focused on using students' past academic records. The authors used a neural network as the machine learning technique. They used a dataset consisting of a training set of 60 students and validation of 10 students to predict marks on some courses. Their proposed model achieved 70.48% accuracy.

In [15], the authors proposed a method to predict students' final grades. They focused on gathering students' comments about each class that reflects their level of understanding and the difficulties that the students might face. They employed Word2Vec and Artificial Neural Network (ANN) to predict students' final grades for each class. The author's dataset includes students' comments that were collected in each class which includes 15 lessons. 123 students were asked to fill out a questionnaire about their learning level. They achieved different accuracies for each grade. For grade S, the average accuracy reported was 87.3%, Grade-A 79.1% average accuracy, grade-B 85.0% average accuracy, grade-C 88.5% average accuracy, and for grade-D 89.5% average accuracy.

In [16], the authors proposed a system that predicts dropout intentions and students' grades. They focused on gathering information that represents student academic data. The authors used Logistic Regression (LR), Gaussian Naïve Bayes (GB), Support Vector Machine (SVM), Random Forest (RF), and Adaptive Boosting (AdaBoost) classifications. Their dataset consisted of 3,463 students majoring in law, 526 students majoring in mathematics, and 455 computer science students. Their system achieved 83% accuracy for the degree in computer science, 76% average accuracy for law, and 61% average accuracy for mathematics. In [17], the authors proposed a system that predicts students' performance in academic institutions. The authors used correlations, multiple regressions, and process mining to analyze and predict students' performance. The author's dataset consisted of MOOC data from Coursera, performance data, and course evaluation to analyze on-campus courses. The authors reported a 75% average accuracy.

In [18], the authors developed a system that predicts students' performance using conventional statistical analysis and neural networks. The dataset consisted of students' socioeconomic backgrounds and entrance examination results at Chinese universities. They reported an 84.8% average accuracy prediction using ANN. In [19], the authors proposed using an ANN to predict students' grades. They suggested that with such a system, educational services will improve, and students will be able to get the assistance that they need based on their predicted performance. The authors used ANN models to predict the student's performance and grades. The author's dataset consisted of thirty students that were collected randomly from the Department of computer science at Tai Solarin University. They reported a 92.7% average accuracy in predicting students' final grades using ANN.

In [20], the authors used machine learning techniques to predict students' performance and results. They suggested that such a model will be able to identify weak students who may need some assistance and guidance. The authors applied five machine learning algorithms to the dataset and reported different accuracies. The machine learning algorithms used for classification were Decision Tree (DT) with C4.5 algorithm, ANN with Multi-Layer Perceptron (MLP), Bayesian Network (BN) with Naïve Bayes (NB), Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO), Lazy Learners with incremental learning, and 1-Nearest Neighborhood (1-NN). The dataset consisted of undergraduate students majoring in Computer Science. The authors reported 79% average accuracy for training and 66% average accuracy for testing.

In [21], the authors used ANN models to predict students' performance based on one factor. The academic students' results were used to predict the performance of the students using Multilayer Perceptron (MLP) and Generalized Regression Neural Network (RNN). The dataset consisted of 100 graduated students' academic results. The authors reported 75% average accuracy using MLP and 95% average accuracy using Generalized RNN. In [22], the authors used ANN to predict students' performance in e-learning courses. The dataset consisted of 3,518 university students who are taking e-learning courses. They reported 80.47% average accuracy. In [23], the authors used ANN with a feed-forward backpropagation algorithm to predict students' performance in the College of Engineering and information technology. The goal of the study was to identify students who will have extraordinary performance. The dataset consisted of sophomore students' records majoring in engineering. They reported 84.6% average accuracy. In [24], the authors proposed a system to predict students' future grades based on students' past performance. The authors used the Prefix tree for the Sequential Pattern (PSP) algorithm. They used the Canvas network dataset and reported a 95.24% average accuracy. In [25], the authors used data mining techniques to predict students' performance. The algorithms that were used are NB, MLP, and C4.5. The author's dataset was collected using a questionnaire survey for the 2010-2011 academic year. They reported 76.65% average accuracy for NB, 71.20% average accuracy for MLP, and 73,93% average accuracy for C4.5.

In [26], the authors applied an instance-based learning classifier, decision tree, and NB machine learning algorithm to predict student performance. The author's dataset consisted of collecting the log data from the learning management systems (LMSs) and reported 97% average accuracy. In [27], the authors used SVM and KNN algorithms to predict students' final examination grades. The dataset was obtained from the University Minho in Portugal which includes 395 data samples of math subjects. They reported 96% average accuracy using SVM and 95% average accuracy using KNN. In [28], the authors used Neural Network and NB classification to predict students' final grades. The author's dataset consisted of 181 student records from North South University. They reported 75% average accuracy.

In [29], the authors have attempted on predicting student performance using three categories of features: student engagement, demographics, and performance data. Using classification models, they were able to predict at-risk students, and using regressions models, they were able to predict students' scores. The authors used several classification and regression models. For classification, they used an SVM, DT, ANN, NB, and KNN. For regression, they used SVM, ANN, DT, Bayesian Regression (BN), K-NN, and Linear Regression (LR). These models were tested on an open university learning analytics dataset (OULAD). Performance-wise, ANN achieved better performance than other models: for classification, the model achieved an F1-score of approximately 96%, and for regression, the model achieved a Root-Mean-Square Error (RMSE) of approximately 14.59.

The author in [30] used an e-learning simulation software called DEEDS to construct a dataset to track students' interactions during an online lab class in terms of editing text, the number of keystrokes pressed, time spent on each activity, etc., along with the exam score achieved in the session. The prediction model developed was able to, using a total of 86 novel statistical features, predict whether a student's performance is low or high. In the experiments conducted, five popular classifiers were used: RF, SVM, NB, LR, and MLP. Performance-wise, the model achieved the best classification accuracy performance of 97.4% using the RF classifier. By using historic exam grades and in-progress course exam grades, authors in [31] were able to identify "at-risk" students per area and per subject. Based on real-time feedback on students' performance, appropriate remedial strategies were employed to improve retention rates. The study was conducted on 335 students and 6,358 assessment grades considering 68 subjects categorized into 7 different knowledge areas. A Decision Tree was used in their model to classify students into either passing or failing. The best model accuracy performance reported was 96.5%. Though the proposed model achieved reasonably high classification accuracy, it suffers from the complexity of the model in terms of a number of classes being performed (2 classes – good or bad performing students).

The DEEDS dataset discussed earlier was also used in [32], using several machine learning models, to perform a comparative analysis. The models used were ANN, LR, DT, SVM, and NB. Average time, total activities, average mouse clicks, related activities in an exercise, average keystrokes, and average idle time per exercise, were the features the authors extracted and considered. Compared with other models used, SVM achieved the best accuracy of 94%.

The authors in [33] have considered a similar study while considering the same DEEDS dataset. The authors have conducted their study considering the 5 features: average idle time, average time, the total number of activities, total related activity, and the average number of keystrokes. The best-achieved performance when considering 5 different classification algorithms was 80% when considering the Alpha Investing technique on an SVM-based model. The study in [33] has focused only on a binary model (good vs. bad performing students) which is a major shortcoming of the proposed model. Also, in [33] the authors have abstracted the set of proposed features and did not differentiate, for instance between the types of activities within a single exercise.

The authors in [34] have considered applying neural network-based classifiers to predict low-grade students and high-grade students. The authors concluded that students attained performance is proportional to he level of difficulty of the lab topics being covered. Similar to [33], the proposed model did not perform in terms of achieved accuracy which ranged between 0.7 and 0.8.

As mentioned in the introduction of this section, machine learning research is changing the way we view things and the way many industries operate. It is the intention of the authors to make use of machine learning in education to assist students in being informed of their academic progress and success as well as to provide a means for professors and institutions to intervene with students at-risk to assist them to succeed. Table 1 shows the summary of the recent literature for student grade prediction using machine learning techniques along with the dataset used and the best average accuracy reported.

## III. PROPOSED METHODOLOGY

The proposed methodology attempts to build a classification model to predict students' performance using an online student engagement dataset as shown in Figure 1. In this attempt, the model will also be able to provide other useful information such as identifying students at-risk. The modified method

**TABLE 1. Summary of the recent literature for student grades prediction using machine learning techniques (2016-2022).**

| Ref. | Method Name | Dataset Name | Accuracy (%) |
|---|---|---|---|
| [13] | Regularized logistic regression | 32,500 student transcripts | 66.59% |
| [16] | Logistic Regression with Adaptive Boosting | Dataset of 3,463 students majoring in law, 526 in mathematics, and 455 in computer science | 73.33% |
| [17] | Correlation features multiple regressions | MOOC performance and evaluation data from Coursera | 75% |
| [19] | ANN with feed-forward backpropagation | 30 transcripts of computer science students | 92.7% |
| [21] | Generalized Regression Neural Network | 100 graduated students' academic results | 95% |
| [22] | Artificial neural network | 3518 student e-learning courses data | 80.47% |
| [23] | ANN with feed-forward backpropagation | Sophomore engineering students' grades records | 84.6% |
| [24] | Prefix tree for Sequential Patterns algorithm | Canvas network dataset | 95.24% |
| [28] | Neural network and Naïve Bayes | 181 students record from North South University | 75% |
| [30] | Statistical features with Random Forest | DEEDS dataset | 97.4% |
| [31] | Decision Trees | 335 student's data | 96.5% |
| [32] | Artificial neural network | DEEDS dataset | 94.5% |
| [33] | SVM, ANN, Naïve Bayes classifiers, Logistic Regression, and Decision Trees | DEEDS dataset | 75% |
| [34] | J48, Random Tree, MLP, Radial Basis Neural Network | DEEDS dataset | 80% |

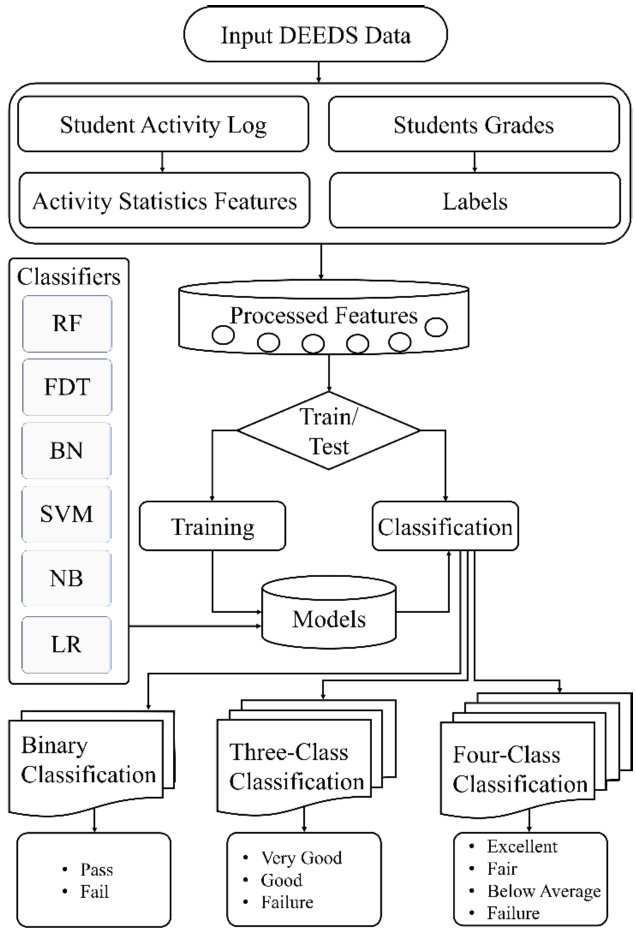



**FIGURE 1. Workflow of the proposed system.**

uses the same dataset of [30] where the author aggregates semantically similar activities resulting in a total of 9 different activities (down from 15). Our approach follows a standard machine learning-based prediction model where the dataset is initially pre-processed to exclude the non-relevant data and ensure that the dataset is properly formatted for further processing. The dataset cleaning phase is followed by a feature extraction step where a total of 254 features were extracted to be used for the performance prediction. This contrasts with the extracted features in [22] and [26], where the authors considered 84 features and 30 features, respectively. Our model considers extracting the following 7 types of features for each type of student activity and each session: (1) Total_idle_time_for_each_exercise, (2) mouse_wheel count, (3) mouse_wheel_click count, (4) mouse_click_left count, (5) mouse_click_right count, (6) mouse_movement count, and (7) keystroke count. Details of these extracted features are described in section III-B.

Contrary to the prediction models presented in [30] and [33], where the authors considered a binary classification of students in terms of either achieving a "good" performance or a "bad" one, the proposed model attempts to extend the model to consider more classes. Initially, and for the sake of performance comparison of our proposed model with the existing ones, we consider 2 classes "pass" or "fail" of students, where "pass" corresponds to students achieving at least 40% of the total grade, and "Fail" otherwise. Where fail is considered the students at-risk. In the second set of experiments, we consider a set of 3 classes where students are categorized in either of the following 3 categories: "Very Good" – where students achieve a total performance between 70% and 100%, "Good" – where students achieve a total performance between 40% and 69%, or "Failure" – where students achieve a total performance between 0% and 39%. Where Good and Failure are considered the students at risk. A further enhanced proposed model suggests increasing the number of classes into 4 classes as follows: "Excellent" – where students achieve a total grade between 80% and 100%, "Fair" – where students achieve a total grade between 60% and 79%, "Below Average" – where students achieve a total grade between 40% and 59%, and "Failure" – where students achieve a total grade between 0% and 39%. Where Below Average and Failure are considered the students at-risk. Prediction results were collected using the following set of classifiers being applied in a standard manner as well as along with ensemble learning (boosting and bagging): (1) Random Forest (RF), (2) Fast Decision Trees (FDT), (3) Bayesian Network (BN), (4) Support Vector

Machine (SVM), (5) Naïve Bayesian, and (6) Linear Regression. The performance results of our proposed model were compared to that in [30], [32], [33], and [34] where the same DEEDS dataset was used.

### A. EXPERIMENTAL DATASET

In this paper, the authors chose the Digital Electronic Education and Design Suites (DEEDS) dataset that uses a proprietary Enhanced Learning Technology (ELT) to capture in real-time, the in-class student behavior and interaction [35]. This is considered one of the few globally available datasets that provided this information. What sets DEEDS apart is the data available which consists of the learning environment, time spent on each problem, average idle time, average keystrokes applied, the performance achieved per session, and the type of online activities that each student was performing, etc. The richness of the data available provided an opportunity for developing a classification technique that can outperform similar techniques being applied to the same dataset. The detailed dataset features description is shown in Table 2.

**TABLE 2. Summary of the DEEDS dataset log statistics.**

| DEEDS data description | Statistics |
|---|---|
| Total Students | 115 |
| Non-participating Students in any Activity | 7 |
| Total Online Lab Sessions | 6 |
| Avg. number of students in each session | 87 |
| Total Exercises | 6 |
| Total Online Activities | 15 |
| Online Log Features Captured | 13 |
| Total Log Entries for all session | 230,318 |

The DEEDS dataset developed at the University of Genoa, Italy in 2015 was developed by logging the students' interactions while doing class activities and exams. Students learn a certain topic in the sessions and for each one of the 6 lab topics, students were required to perform exercises that range in number from 4 to 6. The session maximum grade was set at either 4 or 6 depending on the activity and topic. During each activity and exam, all students were logged for a complete and comprehensive view of their behaviors while performing the activities and exams. For instance, in a study session, student activities are recorded by the DEEDS platform. These activities are a collection of 1 or many of the 15 possible activities such as using a text editor, using a simulation timing diagram tool, reviewing study material, etc. After attending all lab sessions, students will be taking an exam where questions are chosen to cover each of the six lab topics. DEEDS creates a separate file for each student attending a session and adds a new entry every one-second interval. It is important to mention that the number of sessions in the DEEDS dataset is 6. Students were tested with an intermediate and final exam in all sessions except for the first one, where no exercises nor exams were taken. Session 1 data was eventually filtered from our processed dataset. The attained grades will be used to label the used data.

### B. DYNAMIC FEATURES EXTRACTION FROM ONLINE ACTIVITY

To predict students' performance, we consider an extended list of features, all features categorized in either of the following 3 categories: (1) features based on the count of each of the 9 types of Activities, (2) features based on the count of Timing statistics, and (3) features based on the count of peripheral activity. These features were collected per student and per session. we propose adopting 252 new features which include the 86 features considered in [22] to better describe each of the students' 9 activities depicted in Table 3 in terms of the 7 categories of features shown in Table 4.

**TABLE 3. Online student activity types along with their frequency.**

| Category | Order | Frequency |
|---|---|---|
| Editing | 1 | 18% |
| Aulaweb | 2 | 4% |
| Deeds Activity | 3 | 17% |
| FSM | 4 | 9% |
| Study | 5 | 10% |
| Blank | 6 | 10% |
| Other | 7 | 15% |
| Diagram | 8 | 9% |
| Properties | 9 | 8% |

**TABLE 4. New extracted features definition.**

<table>
<tr><th>Features</th><th>Description</th><th>Category</th></tr>
<tr><td>Idle_time</td><td>Idle time duration during a specific activity</td><td>Timing statistics-based features</td></tr>
<tr><td>Mouse_wheel</td><td>Mouse wheel count during a specific activity</td><td rowspan="6">Peripheral activity count-based features</td></tr>
<tr><td>Mouse_Wheel_click</td><td>Mouse click count during a specific activity</td></tr>
<tr><td>Mouse_click_left</td><td>Left mouse click count during a specific activity</td></tr>
<tr><td>Mouse_click_right</td><td>Right mouse click count during a specific activity</td></tr>
<tr><td>Mouse_movement</td><td>Distance covered during a specific activity</td></tr>
<tr><td>Keystrokes</td><td>Keystrokes count during a specific activity</td></tr>
</table>

The extracted 252 features are broken down – for each feature, into 36 per feature type (a total of 7 as shown in Table 4). For each type of feature, statistics were collected in each exercise (a total of 4) and for each type of activity (a total of 9) as shown in Figure 2. Figure 2 shows that features are extracted for each exercise using the online student activity within each exercise. From each online activity, the 7 categories listed in Table 4 are extracted as features for the particular activity for the particular exercise for each student.

For the *"Idle_time"* type of features, they were calculated by summing up the total idle time per exercise and per type of

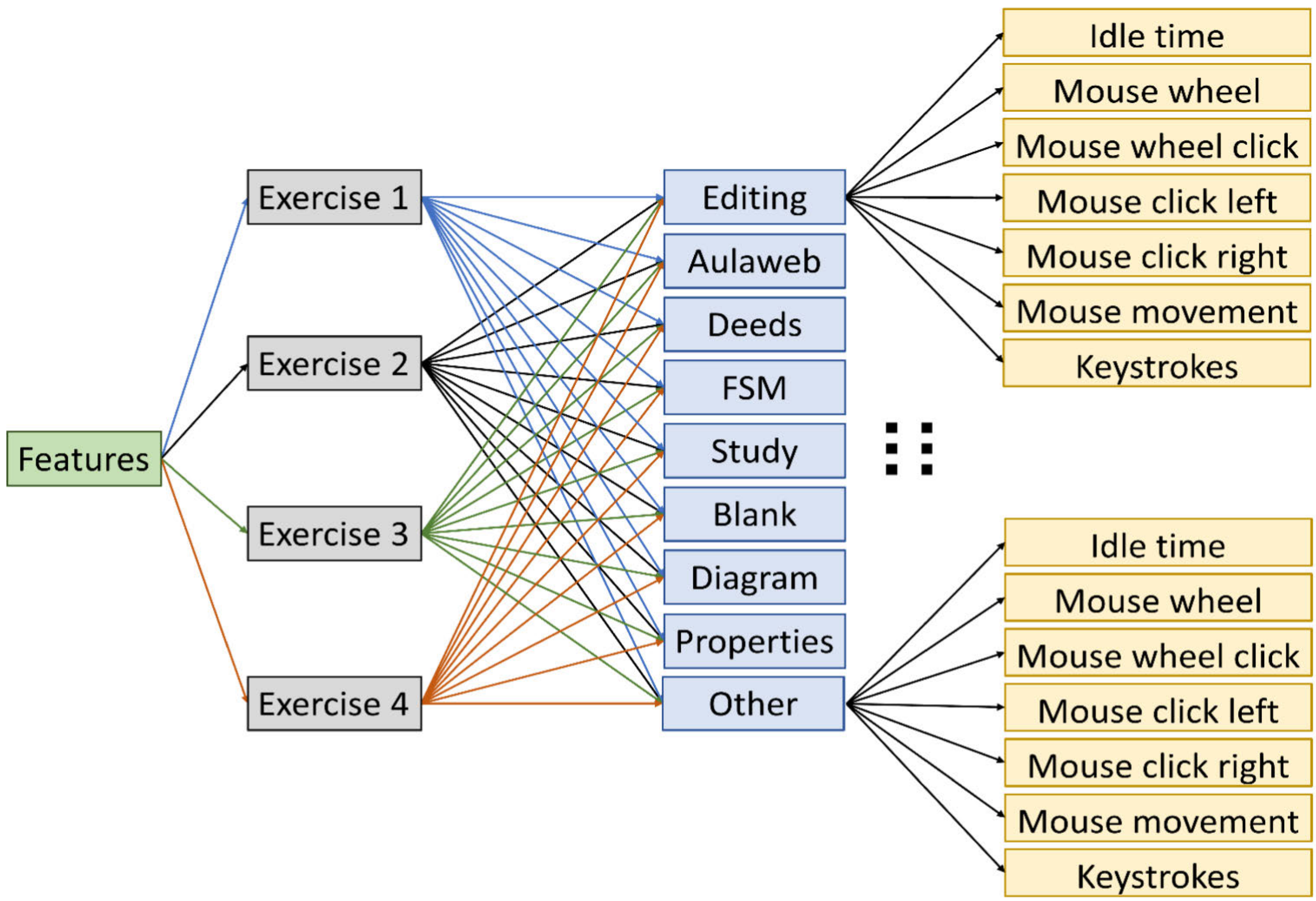



**FIGURE 2.** Characterization for the extracted 252 features based on each activity and exercise.

activity, resulting in a total of 36 features. These are captured in the matrix T as shown in Equation (1).

$$T = \begin{bmatrix} t_{1,1} & t_{1,2} & t_{1,3} & \dots & t_{1,9} \\ t_{2,1} & t_{2,2} & t_{2,3} & \dots & t_{2,9} \\ t_{3,1} & t_{3,2} & t_{3,3} & \dots & t_{3,9} \\ t_{4,1} & t_{4,2} & t_{4,3} & \dots & t_{4,9} \end{bmatrix} \quad (1)$$

Matrix T is a 4 by 9 matrix, element "$t_{i,j}$" represents the total sum of Idle_time in exercise "$i$" for activity "$j$." For instance, "$t_{1,3}$" represents the total Idle_time spent by a specific student in exercise 1 for activity 3 ("*Study*") calculated as shown in Equation (2).

$$t_{i,j} = \sum_{1 \to n} Idle_{Time} \quad (2)$$

where $n$ is the total count of occurrence of idle times in exercise $i$ and activity $j$.

Following the definition of matrix T, the 36 *"Idle_time"* related features are depicted in Equation (3).

$$F_{k=\{1 \to 36\}} = t_{\lceil k/9 \rceil, k-9\lfloor k/9 \rfloor} \quad (3)$$

In Equation 2, $F_1$ for instance $= t_{1,1}$ represents the total idle time spent in exercise_1 for the "*Editing*" type of activity.

The next category of peripheral-based features – features 2 through 7 as depicted in Table 4) were extracted per exercise and per activity type in the same way as the *"Idle_time"* related features were extracted. For instance, the *"Mouse_Wheel"* type of features was calculated by summing up the total *"Mouse_Wheel"* count per exercise and per type of activity, resulting in a total of 36 features. These are captured in the matrix MW as shown in Equation (4).

$$MW = \begin{bmatrix} mw_{1,1} & mw_{1,2} & mw_{1,3} & \dots & mw_{1,9} \\ mw_{2,1} & mw_{2,2} & mw_{2,3} & \dots & mw_{2,9} \\ mw_{3,1} & mw_{3,2} & mw_{3,3} & \dots & mw_{3,9} \\ mw_{4,1} & mw_{4,2} & mw_{4,3} & \dots & mw_{4,9} \end{bmatrix} \quad (4)$$

Matrix MW is a 4 by 9 matrix, element "$mw_{i,j}$" represents the total sum of the *"Mouse_Wheel"* count in exercise "$i$" for activity "$j$." For instance, "$mw_{1,3}$" represents the total *"Mouse_Wheel"* spent by a specific student in exercise 1 for activity 3 ("Study") calculated as shown in Equation (5).

$$mw_{i,j} = \sum_{1 \to n} Mouse\_Wheel \quad (5)$$

where $n$ is the count of occurrence of Mouse Wheel in exercise $i$ and activity $j$. Following the definition of matrix MW, the 36 Idle_time related features are depicted in Equation (6).

$$F_{k=\{37\to72\}} = mw_{\lceil k/9\rceil, k-9\lfloor k/9\rfloor} \quad (6)$$

In Equation 6, $F_{37}$ for instance $= mw_{1,1}$ represents the count of the *"Mouse_Wheel"* in exercise 1 for the "*Editing*" type of activity.

The same process is being considered to extract the remaining 180 features ($F_{k=\{73\to252\}}$) for the remaining 5 peripheral-related activity features namely: *"Mouse_Wheel_click," "Mouse_click_left," "Mouse_click_right," "Mouse_movement,"* and *"Keystrokes"* resulting in 36 new features per peripheral activity.

### C. CLASSIFICATION

In the context of Machine Learning, classification is defined as the process of grouping or categorizing data into a pre-defined number of classes. It is considered a pattern recognition technique where various classification algorithms may be applied to historic training data to find hidden patterns with the purpose of making future predictions on data with the same set of characteristics (aka. Features) and unknown classes. Based on the application domain, the data distribution, and the number of targeted classes, classification tasks may fall under any of these 4 categories: (1) Binary classification – where targeted classes can be any of 2 values, (2) Multi-class classification – where targeted classes can be more than two values, (3) Multi-label classification – where one instance of the dataset can have more than one label, and (4) Imbalanced classification – where the instances of the dataset under study have a skewed or biased distribution of labels. In the literature, many existing algorithms can be applied to classify instances. These include Naive Bayes, Decision Tree, K-Nearest Neighbor, Support Vector Machine, Logistic Regression, etc. In this research work, we consider modeling the student prediction problem as a binary classification problem in the first step, then extend the model to a multi-class classification problem. Our model applies the following six base learner models: Random Forest (RF), Fast Decision Trees (FDT), Bayesian Network (BN), Support Vector Machine (SVM), Naïve Bayesian (NB), and Linear Regression (LR). We propose the use of the ensemble method to include boosting and bagging on the 6 standard classifiers to increase the accuracy.

In the next subsections, a brief description of the top performing classifiers as well as the boosting and bagging techniques are provided.

#### 1) RANDOM FOREST (RF)

Random Forest (RF) is a supervised ensemble learning method used primarily for classification and regression [36], [37]. The method forms an ensemble by constructing a collection of decision trees with controlled variation by combining Breiman's bagging sampling technique [38], and the random selection of features, initiated independently. Compared with other ensemble learning methods, RF is known for its high accuracy and robustness to noise and overfitting. Consequently, it caught the attention of many researchers who produced several extensions of it to further improve its performance.

Although increasing the number of trees in an RF has the potential of increasing the accuracy, RF often involves higher time and space to train the model as the number of trees increases. Furthermore, empirical and theoretical studies conducted by [39] clearly demonstrated that adding more trees to an RF beyond a certain limit (i.e. 500 trees) does not necessarily improve accuracy.

Given a training set X $= x_1, \ldots, x_n$ with class labels Y $= y_1, \ldots, y_n$, bagging repeatedly selects random samples with replacement from the training set and constructs trees from these samples to form an RF ensemble:

For m = 1, . . . , M.

- Sample with replacement n training examples from X, Y and call these $X_m$ and $Y_m$.
- Using $X_m$ and $Y_m$, train a classification or regression tree $f_b$.

By averaging the predictions from all the individual regression trees (or by taking majority voting in the case of classification trees), predictions for unseen samples $x'$ can be made based on Equation 7.

$$f^{\wedge} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \quad (7)$$

#### 2) FAST DECISION TREE (FDT)

FDT is an extension of the traditional Decision Tree (DT) that was mainly developed to improve the performance of DT. Like a DT, an FDT is a supervised learning technique that can be used for both classification and regression problems. It is a tree-based classifier, where the features of a dataset are represented as internal nodes, branches represent the decision rules, and each leaf node represents the outcome. To improve the performance of DTs, many researchers developed FDTs that are much faster to train and evaluate [40]. Due to their efficiency, FDTs are used in many real-time applications where data is created in real-time and tend to perform much better than DTs.

#### 3) BAYESIAN NETWORK (BN)

BNs are also known as belief networks (or Bayes nets for short). Bayesian networks (BNs) are graphical structures used to represent knowledge about an uncertain domain [41]. Each node in the graph corresponds to a random variable, and each edge represents the conditional probability for the corresponding random variables. Because conditional probabilities in the graph are often estimated by using known statistical and computational methods, BNs are considered multi-disciplinary combining principles from graph theory, probability theory, computer science, and statistics.

Key advantages of BNs include efficiency in learning and query answering, and their ability to handle small and incomplete datasets. Due to the lack of a universally acknowledged method for constructing BNs from data, the design of BNs is hard to make compared to other networks.

#### 4) BOOSTING ENSEMBLE LEARNING

Boosting is an incremental process of building a sequence of classifiers, where each classifier works on the incorrectly classified instances of the previous one in the sequence [42]. Boosting is known to be resilient and robust to overfitting. On the downside, it tends to be sensitive to outliers and is considered hard to implement in real time due to the increased complexity of the algorithm.

A good representative of boosting ensemble learning is AdaBoost which is based on Equation (8).

$$F_T(x) = \sum_{t=1}^{T} f_t(x) \tag{8}$$

where each $f_t$ is a weak learner that, for each sample in the training set, produces an output hypothesis that fixes a prediction for that sample. A weak learner is selected at each iteration t, and by assigning it a coefficient $\alpha_t$, the total training error $E_t$ of the resulting t-stage boosted classifier is minimized as shown in Equation (9).

$$E_t = \sum_{i} E[F_{t-1}(x_i) + \alpha_t h(x_i)] \tag{9}$$

#### 5) BAGGING ENSEMBLE LEARNING

The bagging technique, also known as bootstrap aggregation, is considered the most famous method [43]. It was introduced by [44] where each decision tree in the ensemble is created using a sample with replacement from the training data. Statistically speaking, the sample is likely to have about 64% of instances appearing at least once in the sample and are referred to as in-bag-instances. The remaining instances (about 36%), are referred to as out-of-bag (OOB) instances. Minimizing overfitting, improving the model's accuracy, and dealing with higher dimensional data efficiently are key advantages of bagging. Despite its ability to improve accuracy, bagging can be computationally expensive.

#### 6) NAÏVE BASED (NB)

Considered one of the simple and most effective classification algorithms, NB helps in building fast machine-learning models that can make quick predictions [45]. Because it predicts on the basis of the probability of an object, it is considered a probabilistic classifier. It is highly scalable and is not sensitive to irrelevant features.

One noticeable drawback of NB is that it assumes that all features in the dataset are independent or unrelated, which rarely happens in real life. Consequently, this assumption has the potential of limiting its application to real-world datasets.

#### 7) SUPPORT VECTOR MACHINE (SVM)

An SVM classifier is a supervised learning algorithm capable of performing classification and regression analysis [9]. It uses several types of kernels to solve non-linear problems by transforming linearly inseparable data into linearly separable ones. A major drawback of SVM is that it doesn't perform well for large datasets due to high training time.

#### 8) LOGISTIC REGRESSION (LR)

Used for both classification and regression, LR is considered one of the most commonly used binary classification algorithms [30]. It is easy to implement, interpret, and very efficient to train. It performs well when the dataset is linearly separable. Its major drawback, however, is that it is limited to binary classification.

## IV. RESULTS AND DISCUSSION

In this section, we report the performance results of the proposed models on the DEEDS dataset. These results were reported for each of the classification scenarios (binary, 3-classes, and 4-classes) while considering the following set of metrics: accuracy, precision, recall, and F1-Measure. For binary classification, out of a total of 575 records, Pass has 408 and Fail has 167 records. For the 3-classes, Very Good has 285, Good has 123 and Failure has 167 records. Similarly, for the 4-classes, Excellent has 107, Fair has 178, Below Average has 123 and Failure has 167 records. It is important to mention that the experimental setup consists of a random distribution-based model where 80% of the data was randomly used (resulting in 460 records $\sim$ 4 data sessions) for training and 20% being used for testing (115 entries $\sim$ 1 data session).

All the experiments are performed using the standard parameters for different classifiers. For the Boosting based experiments, AdaBoost.M1 is used which was proposed by Yoav and Robert [46], and for Bagging based experiments, used Bagging Predictor method proposed by Leo [44] . For both algorithms (Boosting and Bagging), a batch size of 100 for 10 epochs while the parameters of each classifier were kept to their default values as follows. For the SVM classifier, the learning rate, Loss, Soft margin, and Batch size were set to 0.001, 0.1, 1, and 100, respectively. The Kernel Radial Basis Function (RBF) and Optimizer Sequential Minimal Optimizer (SMO) were also used. For the RF classifier, the model was run with the batch size, Epochs, and Tree depth being set to 100, 100, and 20, respectively. For the NB classifier, the model was run with the batch size, Distribution, and Discretization set to 100, Normal, and False, respectively. For the FDT classifier, the model was run with the batch size and Learning rate set to 100 and 0.001, respectively. For the LR and BN classifiers, both models were run with the batch size set to 100. In the case of BN, the Simple Bayesian Net estimator and Hill Climbing Search Algorithm were used.

Table 5 shows the experimental results for binary performance prediction using the standard learning algorithm, and the base learners using boosting and bagging on the 6 standard classifiers Random Forest (RF), Fast Decision Trees (FDT), Bayes Network (BN), Support Vector Machines (SVM), Naïve Bayes (NB), and Logistic Regression (LR). Comparing the results in Table 5, it can be observed that the best results (in bold) were obtained using the FDT classifier enhanced with the base learners boosting and bagging which does the learning process independently and in parallel. For boosting, however, the learning is done sequentially and based on the results from the previous learner. The highest achieved results in this experiment were 98.25% accuracy, 0.983 precision, 0.982 recall, and 0.982 F1-measure. In this experiment, only binary classification of pass/fail was used as labels. The results obtained in this experiment can also be compared with the results reported in the extant literature for the same binary classification and using the same dataset as used in [28], [29], and [30].

**TABLE 5.** **Experimental results for two-class (Pass/Fail) performance prediction using the proposed features with different ensemble learning classifiers (boosting and bagging).**

| | | RF | FDT | BN | SVM | NB | LR |
|---|---|---|---|---|---|---|---|
| Standard | Accuracy | 96.49 | **96.49** | 94.74 | 92.98 | 85.96 | 84.21 |
| | Precision | 0.967 | 0.965 | 0.951 | 0.936 | 0.876 | 0.854 |
| | Recall | 0.965 | 0.965 | 0.947 | 0.93 | 0.86 | 0.842 |
| | F1 | 0.964 | 0.965 | 0.946 | 0.927 | 0.863 | 0.845 |
| Boosting | Accuracy | 96.49 | **98.25** | 98.25 | 92.98 | 85.96 | 87.72 |
| | Precision | 0.967 | 0.983 | 0.983 | 0.936 | 0.876 | 0.88 |
| | Recall | 0.965 | 0.982 | 0.982 | 0.93 | 0.86 | 0.877 |
| | F1 | 0.964 | 0.982 | 0.982 | 0.927 | 0.863 | 0.878 |
| Bagging | Accuracy | 96.49 | **98.25** | 94.74 | 92.98 | 89.47 | 85.96 |
| | Precision | 0.967 | 0.983 | 0.951 | 0.936 | 0.895 | 0.866 |
| | Recall | 0.965 | 0.982 | 0.947 | 0.93 | 0.895 | 0.86 |
| | F1 | 0.964 | 0.982 | 0.946 | 0.927 | 0.895 | 0.862 |

Figure 3 shows the confusion matrix for the standard FDT and the base learners of boosting FDT and Bagging FDT. Enhancing the learning algorithms with the base learners of boosting and bagging has increased the True Positive TP to 100% as compared to the TP using the standard methods where the TP was only 97.5%. This increase of 2.5% becomes more significant as the complexity of the classification problem increases.

The experiment was repeated under the same setup for three classes: Very Good, Good, and Failure with corresponding scores of 70%-100%, 40%-69%, and 0%-39% as shown in Table 6. Using the standard classifier applied to the DEEDS dataset, the highest achieved accuracy was obtained when using FDT with an accuracy of 89.47, precision of 0.913, recall of 0.895, and F1-measure of 0.888. When using the boosting base learners (basically the learning is done sequentially and based on the results from the previous learner), the highest results, in this case, were for the BN classifier with an accuracy of 84.21%, precision of 0.836, recall of 0.842, and F1-measure of 0.838. We notice a slight drop in the accuracy metrics (compared to binary classification) which is due to

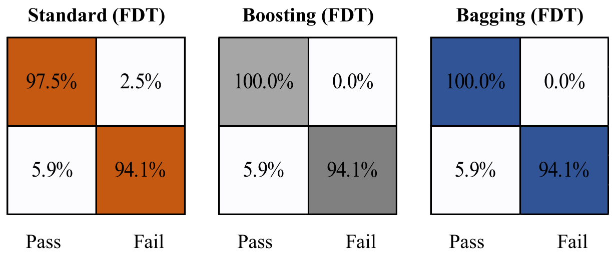



**FIGURE 3.** **Confusion matrix-based comparison for FDT for two class performance prediction.**

the added complexity of considering a third label. When using the bagging base learners, FDT again achieved the best results with an accuracy of 89.47%, precision of 0.9, recall of 0.895, and F1-measure of 0.892. Compared to the standard, the F1 measure was the only metric that showed improvement. Overall, in the case of three label performance prediction problem, the FDT along with the bagging ensemble method achieved the best overall results taking all metrics into account.

Figure 4 describes the confusion matrices for the standard FDT, boosting BN, and bagging FDT. Though we achieved almost similar results when using standard FDT and bagging FDT in Table 6, however, now zooming into the Bagging FDT confusion matrix, it becomes apparent that the bagging FDT performed better, and it justifies the relatively high F1-measure being observed in this case compared to other models.

The experiments were repeated for four classes: excellent 80%-100%, Fair 60%-79%, Below Average 40%-59%, and Failure 0%-39% as shown in Table 7. In this case, we observe that the ensemble learning algorithms (boosting and bagging) did not have a positive effect on the results. On the contrary,

**TABLE 6.** **Experimental results for three-class performance prediction using the proposed feature with different classifiers with ensemble learning (boosting and bagging).**

| | | RF | FDT | BN | SVM | NB | LR |
|---|---|---|---|---|---|---|---|
| Standard | Accuracy | 84.21 | **89.47** | 80.7 | 71.93 | 71.93 | 71.93 |
| | Precision | 0.843 | 0.913 | 0.817 | 0.721 | 0.775 | 0.724 |
| | Recall | 0.842 | 0.895 | 0.807 | 0.719 | 0.719 | 0.719 |
| | F1 | 0.834 | 0.888 | 0.808 | 0.719 | 0.726 | 0.72 |
| Boosting | Accuracy | 82.46 | **82.46** | 84.21 | 71.93 | 71.93 | 71.93 |
| | Precision | 0.826 | 0.828 | 0.836 | 0.719 | 0.775 | 0.724 |
| | Recall | 0.825 | 0.825 | 0.842 | 0.778 | 0.719 | 0.719 |
| | F1 | 0.819 | 0.824 | 0.838 | 0.765 | 0.726 | 0.72 |
| Bagging | Accuracy | 85.97 | **89.47** | 82.46 | 71.93 | 78.95 | 68.42 |
| | Precision | 0.861 | 0.9 | 0.831 | 0.778 | 0.786 | 0.632 |
| | Recall | 0.86 | 0.895 | 0.825 | 0.719 | 0.789 | 0.684 |
| | F1 | 0.855 | 0.892 | 0.823 | 0.765 | 0.787 | 0.645 |

the highest achieved results were recorded using the standard algorithms applied to the DEEDS dataset using the FDT classifier. For standard FDT, the accuracy was 77.19%, the precision was 0.809, the recall was 0.772, and the F1 measure was 0.78. The lowest metric results were obtained when using the boosting ensemble learning with the RF classifier where an accuracy of 73.68%, a precision of 0.781, a recall of 0.737, and an F1 measure of 0.745 were achieved. We still see in this case that the standard FDT is closely followed by the ensemble algorithm bagging using the BN classifier. The problem of having four classes is extremely complex for a complex dataset used in the work, in addition, to the number of features being extracted. The FDT decision tree is an ensemble method on its own and thus is extremely powerful in the classification process. Hence, it is arguably logical that as the number of classes increases with the same number of features, the traditional ensemble algorithms such as FDT will achieve the highest accuracy, and the base learners such as boosting, and bagging will have little to no effect.

The experiments were repeated for four classes: excellent 80%-100%, Fair 60%-79%, Below Average 40%-59%, and Failure 0%-39% as shown in Table 7. In this case, we observe that the ensemble learning algorithms (boosting and bagging) did not have a positive effect on the results. On the contrary, the highest achieved results were recorded using the standard algorithms applied to the DEEDS dataset using the FDT classifier. For standard FDT, the accuracy was 77.19%, the precision was 0.809, the recall was 0.772, and the F1 measure was 0.78. The lowest metric results were obtained when using the boosting ensemble learning with the RF classifier where an accuracy of 73.68%, a precision of 0.781, a recall of 0.737, and an F1-measure of 0.745 were achieved. We still see in this case that the standard FDT is closely followed by the ensemble algorithm bagging using the BN classifier. The problem of having four classes is extremely complex for a complex dataset used in the work, in addition, to the number of features being extracted. The FDT decision tree is an ensemble method on its own and thus is extremely powerful in the classification process. Hence, it is arguably logical that as the number of classes increases with the same number of features, the traditional ensemble algorithms such as FDT will achieve the highest accuracy, and the base learners such as boosting, and bagging will have little to no effect.

Figure 5 shows the confusion matrix for the standard FDT, boosting RF, and bagging BN. We can observe through the confusion matrix that standard FDT achieves the best results followed by bagging BN and finally boosting RF.

Figure 6 shows a summary of predictions using one session of testing and the rest of the four sessions for the training. As can be seen, the boosting and bagging ensemble set of algorithms in the case of binary classification, both compete in achieving the highest accuracy in each session. Looking at the sessions S2-S6, we observe that the achieved performance of the boosting and bagging depends on the session being considered, however, as was shown in Table 4 the overall average accuracy was the same for both. For the case of

**TABLE 7. Experimental results for four-class performance prediction using the proposed feature with different classifiers with ensemble learning (boosting and bagging).**

| | | RF | FDT | BN | SVM | NB | LR |
|---|---|---|---|---|---|---|---|
| Standard | Accuracy | 84.21 | **89.47** | 80.7 | 71.93 | 71.93 | 71.93 |
| | Precision | 0.843 | 0.913 | 0.817 | 0.721 | 0.775 | 0.724 |
| | Recall | 0.842 | 0.895 | 0.807 | 0.719 | 0.719 | 0.719 |
| | F1 | 0.834 | 0.888 | 0.808 | 0.719 | 0.726 | 0.72 |
| Boosting | Accuracy | 82.46 | **82.46** | 84.21 | 71.93 | 71.93 | 71.93 |
| | Precision | 0.826 | 0.828 | 0.836 | 0.719 | 0.775 | 0.724 |
| | Recall | 0.825 | 0.825 | 0.842 | 0.778 | 0.719 | 0.719 |
| | F1 | 0.819 | 0.824 | 0.838 | 0.765 | 0.726 | 0.72 |
| Bagging | Accuracy | 85.97 | **89.47** | 82.46 | 71.93 | 78.95 | 68.42 |
| | Precision | 0.861 | 0.9 | 0.831 | 0.778 | 0.786 | 0.632 |
| | Recall | 0.86 | 0.895 | 0.825 | 0.719 | 0.789 | 0.684 |
| | F1 | 0.855 | 0.892 | 0.823 | 0.765 | 0.787 | 0.645 |

three classes, the standard and bagging algorithms achieved comparable high performance in terms of accuracy in most sessions. In the case of the four classes, the standard is showing more superiority in achieving the highest accuracy. It is easily observed in this case, that when using standard ensemble methods such as FDT, the accuracy achieved when increasing classes is already high, and combining them with base learners of boosting and bagging has little to no effect on the overall model performance.

It is important to mention that, though the achieved results for the 3-class and 4-class prediction models, which were constructed by combing "ensemble-based methods" and standard classification algorithm, were close to 90% accuracy. Higher performance will be achieved with a larger dataset in terms of the number of students taking part in this exercise (115 students in the case of DEEDS) leading to a larger training dataset and eventually a more accurate prediction model.

As stated above, most of the previous literature used the same dataset for binary classification, and thus we can compare the results obtained in this work for binary classification with those reported in the extant literature. Figure 7 shows the comparison of the results obtained in this work with three works from the extant literature that used the same dataset, and we observe that for all metrics, the proposed modified process and method in this work has the highest performance compared with those reported in the extant literature.

The three-class and four-class results could not be compared because according to the author's knowledge, there are no research papers published for the same dataset for multi-class classification. However, the results obtained for multi-class classification using the proposed method were significant. The results obtained for multi-class classification

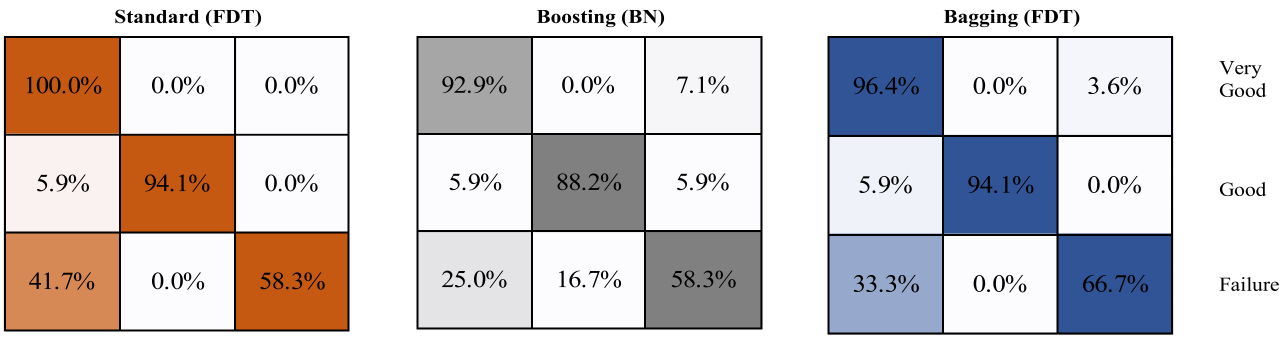



**FIGURE 4.** Confusion matrix-based comparison for top classifiers for three class performance prediction.

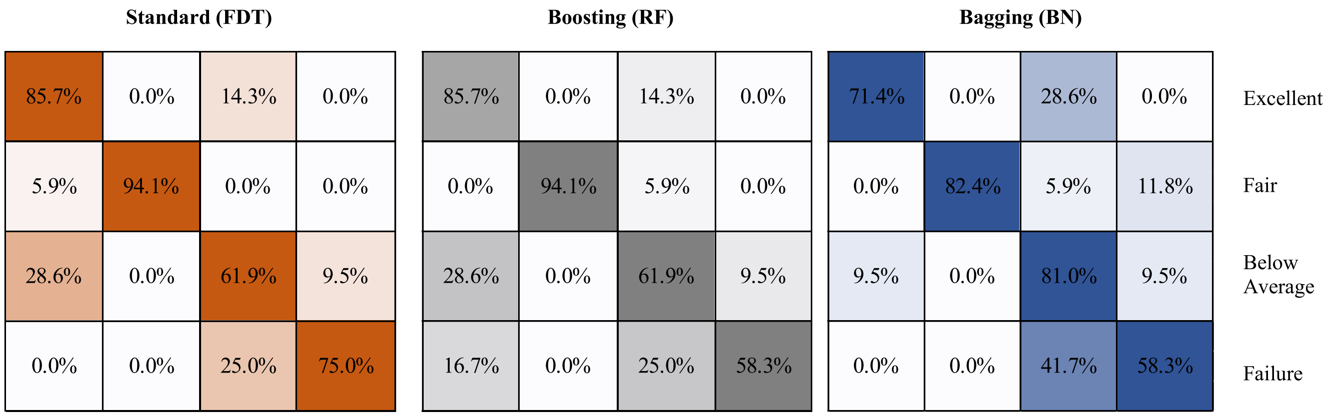



**FIGURE 5.** Comparison of best-performing classifiers (FDT, RF, BN) based on the confusion matrix for four grading classes prediction.

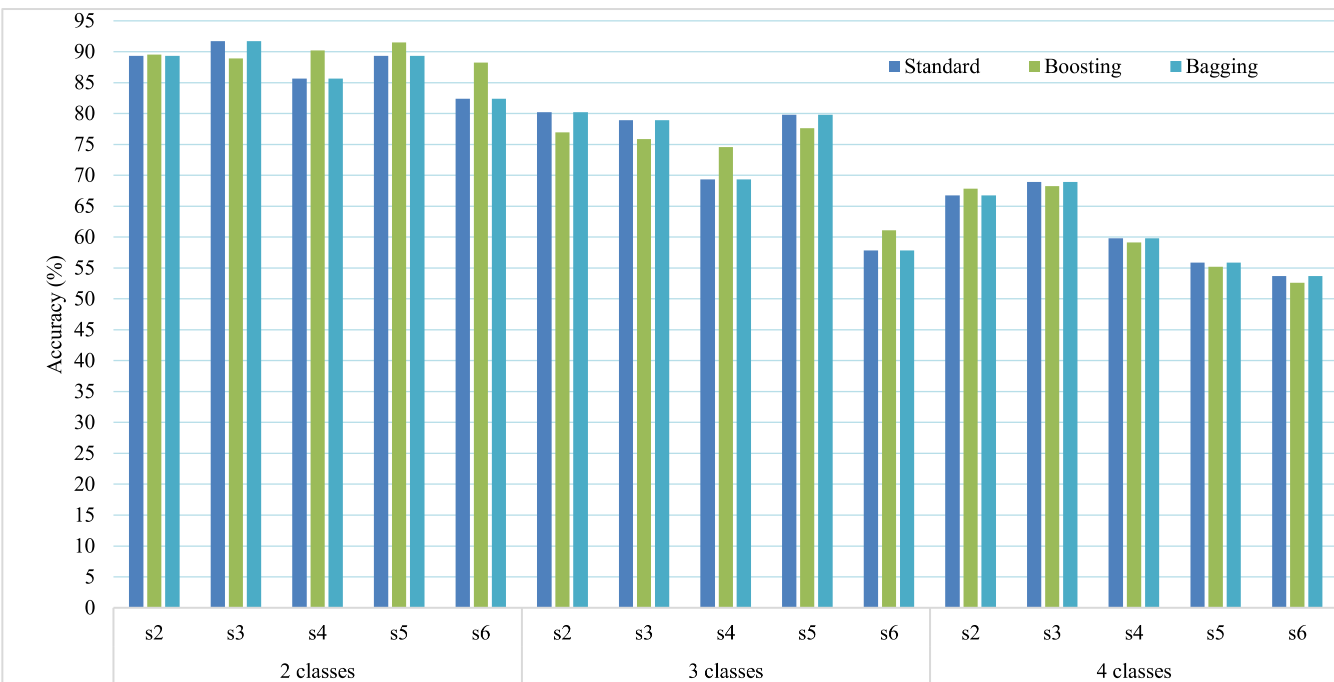



**FIGURE 6.** Comparison of accuracies per session for 2 classes, 3 classes, and 4 classes of student grades.

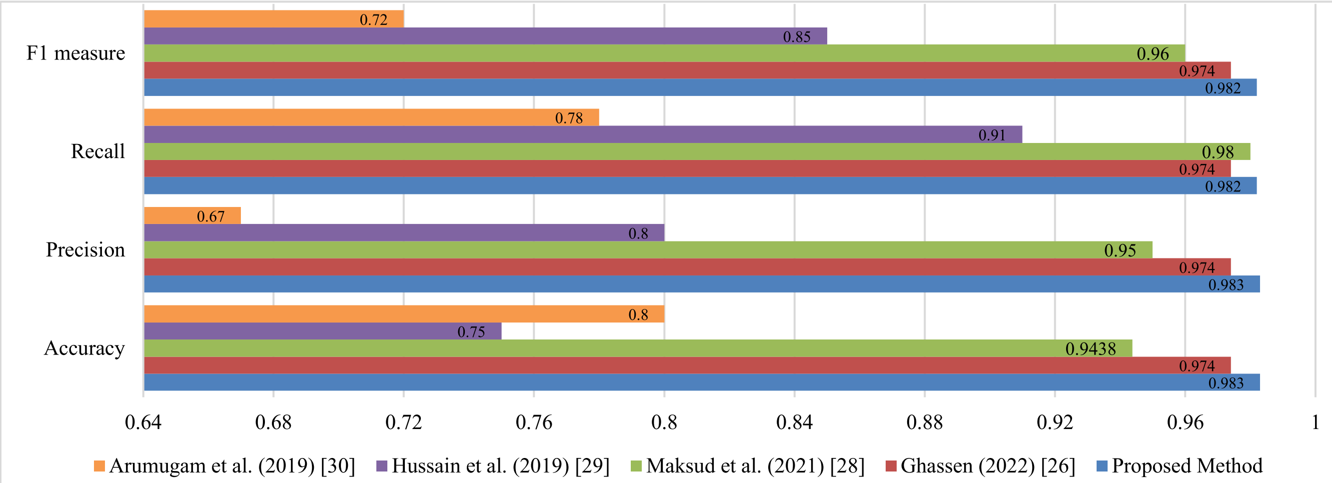



**FIGURE 7.** Proposed model performance comparison with the latest literature performance using the same dataset for binary classification.

are in some cases higher than results reported for binary classification in the extant literature. The multi-class problem is usually complex and stands for improvement in future research work.

The research presented in this work can stand as a baseline for other researchers tackling the students' performance prediction problem using dynamic data obtained from students and combined with the student's static data. Dealing with datasets of students' static data such as grades on various assessments is a much easier problem than approaching the student performance from a holistic perspective that takes students' behavior while taking the assessments into account in addition to their grades. In the future, this can lead researchers to identify students with learning disabilities and be able to diagnose the disability from behavior in assessments. It can be also expanded to identify students with emotional and psychological problems that might need intervention from counselors.

## V. CONCLUSION

In this work, we propose a modified machine learning model for student performance prediction and identification of students at risk. The dataset used in this work is one of few that is globally available that provides information on student interaction with classroom activities as well as their results. The uniqueness of the dataset is that it deals with the dynamic aspects of a student rather than just the static data related to the student like grades in homework, quizzes, and exams. This makes it unique in such a way that it sets the pace for developing a holistic dataset that can capture both dynamic and static data related to students. For example, we hope to see future datasets that can capture not only students' interaction but also the student's state of mind and the student's psychological and socioeconomic aspects as well as the student's static data. This kind of dataset will bring us closer to developing a machine learning system that is able to know the student individually to predict their success and be able to concurrently suggest remedial action if the student is not doing well. As these remedial actions are not always academic and could be related to other aspects such as counseling.

Using the available dataset, we propose a modified machine learning algorithm that predicts students' performance and identifies students at risk. The problem was divided into three scenarios; binary classification of the pass and fail, three class scenarios, and four class scenarios. It was shown that ensemble methods combined with base learners of boosting and bagging significantly increase the accuracy for binary classification, while slightly increasing the accuracy for three-class and four-class problems. The ensemble algorithm of bagging and boosting FDT achieved an accuracy of 98.25% for binary classification and 89.47% for three classes. The standard ensemble FDT achieved an accuracy of 77.19% for four classes. The results obtained for binary classification were compared with results reported in the extant literature using the same dataset proving that the proposed modified algorithms achieved better results than similarly proposed methods. The three-class and four-class results could not be compared because according to the author's knowledge, there are no research papers published for the same dataset for multi-class classification. However, the results obtained for multi-class classification using the proposed method were significant. The results obtained for multi-class classification are in some cases higher than results reported for binary classification in the extant literature. The multi-class problem is usually complex and stands for improvement in future research work.

The research presented in this work can stand as a baseline for other researchers tackling the students' performance prediction problem using dynamic data obtained from students and combined with the students' static data. Dealing with datasets of students' static data such as grades on various assessments is a much easier problem than approaching the student performance from a holistic perspective that takes students' behavior while taking the assessments into account in addition to their grades. In the future, this can lead researchers to identify students with learning disabilities and be able to diagnose the disability from behavior in assessments. It can be also expanded to identify students with emotional and psychological problems that might need intervention from counselors.

Future work might include developing a more holistic dataset that includes a more comprehensive set of dynamic and static aspects of students. The authors also plan to continue to work on the DEEDs dataset for multi-class classification to increase the classification according now that a baseline has been set in this work.

## ACKNOWLEDGMENT

For more information about Commonwealth Cyber Initiative, visit https://cyberinitiative.org

## REFERENCES

[1] A. Bashar, G. Latif, G. Ben Brahim, N. Mohammad, and J. Alghazo, ''COVID-19 pneumonia detection using optimized deep learning techniques,'' *Diagnostics*, vol. 11, no. 11, p. 1972, Oct. 2021, doi: 10.3390/diagnostics11111972.

[2] A. A. Mahmoud, I. N. A. Alawadh, G. Latif, and J. Alghazo, ''Smart nursery for smart cities: Infant sound classification based on novel features and support vector classifier,'' in *Proc. 7th Int. Conf. Electr. Electron. Eng. (ICEEE)*, Apr. 2020, pp. 47–52, doi: 10.1109/iceee49618.2020.9102574.

[3] G. Latif, R. Alghazo, M. A. E. Pilotti, and G. B. Brahim, ''Identifying 'at-risk' students: An AI-based prediction approach,'' *Int. J. Comput. Digit. Syst.*, vol. 11, no. 1, pp. 1051–1059, Mar. 2022, doi: 10.12785/ijcds/110184.

[4] S. D. A. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, H. Fujita, and N. A. Md. Ghani, ''Multiclass prediction model for student grade prediction using machine learning,'' *IEEE Access*, vol. 9, pp. 95608–95621, 2021.

[5] A. Gupta, D. Garg, and P. Kumar, ''An ensembling model for early identification of at-risk students in higher education,'' *Comput. Appl. Eng. Educ.*, vol. 30, no. 2, pp. 589–608, Mar. 2022.

[6] G. Kostopoulos, S. Kotsiantis, and P. Pintelas, ''Predicting student performance in distance higher education using semi-supervised techniques,'' in *Model and Data Engineering*. Cham, Switzerland: Springer, 2015, pp. 259–270.

[7] X. Ma and Z. Zhou, ''Student pass rates prediction using optimized support vector machine and decision tree,'' in *Proc. IEEE 8th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2018, pp. 209–215.

[8] M. C. Sáiz-Manzanares, J. J. Rodríguez-Díez, J. F. Díez-Pastor, S. Rodríguez-Arribas, R. Marticorena-Sánchez, and Y. P. Ji, ''Monitoring of student learning in learning management systems: An application of educational data mining techniques,'' *Appl. Sci.*, vol. 11, no. 6, p. 2677, Mar. 2021, doi: 10.3390/app11062677.

[9] G. Latif, G. Ben Brahim, D. N. F. A. Iskandar, A. Bashar, and J. Alghazo, ''Glioma tumors' classification using deep-neural-network-based features with SVM classifier,'' *Diagnostics*, vol. 12, no. 4, p. 1018, Apr. 2022, doi: 10.3390/diagnostics12041018.

[10] Y. Al-Anqoudi, A. Al-Hamdani, M. Al-Badawi, and R. Hedjam, ''Using machine learning in business process re-engineering,'' *Big Data Cognit. Comput.*, vol. 5, no. 4, p. 61, Nov. 2021, doi: 10.3390/bdcc5040061.

[11] G. Latif, J. Alghazo, R. Maheswar, V. Vijayakumar, and M. Butt, ''Deep learning based intelligence cognitive vision drone for automatic plant diseases identification and spraying,'' *J. Intell. Fuzzy Syst.*, vol. 39, no. 6, pp. 8103–8114, Dec. 2020, doi: 10.3233/jifs-189132.

[12] I. K. Nti, A. F. Adekoya, and B. A. Weyori, ''A systematic review of fundamental and technical analysis of stock market predictions,'' *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 3007–3057, Aug. 2019, doi: 10.1007/s10462-019-09754-z.

[13] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, ''Predicting student dropout in higher education,'' 2016, *arXiv:1606.06364*.

[14] H. Agrawal and H. Mavani, ''Student performance prediction using machine learning,'' *Int. J. Eng. Res.*, vol. 4, no. 3, pp. 111–113, Mar. 2015.

[15] J. Luo, S. E. Sorour, K. Goda, and T. Mine, ''Predicting student grade based on free-style comments using Word2Vec and ANN by considering prediction results obtained in consecutive lessons,'' *Int. Educ. Data Mining Soc.*, Jun. 2015. [Online]. Available: https://eric.ed.gov/?id=ED560772

[16] S. Rovira, E. Puertas, and L. Igual, ''Data-driven system to predict academic grades and dropout,'' *PLoS ONE*, vol. 12, no. 2, Feb. 2017, Art. no. e0171207, doi: 10.1371/journal.pone.0171207.

[17] R. Conijn, A. Van den Beemt, and P. Cuijpers, ''Predicting student performance in a blended MOOC,'' *J. Comput. Assist. Learn.*, vol. 34, no. 5, pp. 615–628, May 2018, doi: 10.1111/jcal.12270.

[18] E. T. Lau, L. Sun, and Q. Yang, ''Modelling, prediction and classification of student academic performance using artificial neural networks,'' *Social Netw. Appl. Sci.*, vol. 1, no. 9, pp. 1–10, Aug. 2019, doi: 10.1007/s42452-019-0884-7.

[19] O. L. Usman and A. O. Adenubi, ''Artificial neural network (ANN) model for predicting students' academic performance,'' *J. Sci. Inf. Technol.*, vol. 1, no. 2, pp. 23–37, 2013.

[20] A. Acharya and D. Sinha, ''Early prediction of students performance using machine learning techniques,'' *Int. J. Comput. Appl.*, vol. 107, no. 1, pp. 37–43, Dec. 2014.

[21] A. R. Iyanda, O. D. Ninan, A. O. Ajayi, and O. G. Anyabolu, ''Predicting student academic performance in computer science courses: A comparison of neural network models,'' *Int. J. Modern Comput. Sci.*, vol. 10, no. 6, pp. 1–9, Jun. 2018, doi: 10.5815/ijmecs.2018.06.01.

[22] S. Aydogdu, ''Predicting student final performance using artificial neural networks in online learning environments,'' *Educ. Inf. Technol.*, vol. 25, no. 3, pp. 1913–1927, Nov. 2019, doi: 10.1007/s10639-019-10053-x.

[23] S. S. Abu-Naser, I. S. Zaqout, M. Abu Ghosh, R. R. Atallah, and E. Alajrami. (2015). *Predicting Student Performance Using Artificial Neural Network: In the Faculty of Engineering and Information Technology*. [Online]. Available: http://dstore.alazhar.edu.ps/xmlui/handle/123456789/391

[24] T. N. Doan and S. Sahebi, ''Rank-based tensor factorization for student performance prediction,'' in *Proc. 12th Int. Conf. Educ. Data Mining (EDM)*, 2019, pp. 1–6. [Online]. Available: https://par.nsf.gov/biblio/10185066

[25] E. Osmanbegovic and M. Suljic, ''Data mining approach for predicting student performance,'' *Econ. Rev., J. Econ.*, vol. 10, no. 1, pp. 3–12, 2012. [Online]. Available: http://hdl.handle.net/10419/193806

[26] E. Er, ''Identifying at-risk students using machine learning techniques: A case study with IS 100,'' *Int. J. Mach. Learn. Comput.*, vol. 2, no. 4, p. 476, 2012.

[27] H. Al-Shehri, A. Al-Qarni, L. Al-Saati, A. Batoaq, H. Badukhen, S. Alrashed, J. Alhiyafi, and S. O. Olatunji, ''Student performance prediction using support vector machine and K-nearest neighbor,'' in *Proc. IEEE 30th Can. Conf. Electr. Comput. Eng. (CCECE)*, Apr. 2017, pp. 1–4, doi: 10.1109/ccece.2017.7946847.

[28] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, ''Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique,'' *Decis. Anal.*, vol. 2, no. 1, pp. 1–25, Mar. 2015, doi: 10.1186/s40165-014-0010-2.

[29] N. Tomasevic, N. Gvozdenovic, and S. Vranes, ''An overview and comparison of supervised data mining techniques for student exam performance prediction,'' *Comput. Educ.*, vol. 143, Jan. 2020, Art. no. 103676, doi: 10.1016/j.compedu.2019.103676.

[30] G. B. Brahim, ''Predicting student performance from online engagement activities using novel statistical features,'' *Arabian J. Sci. Eng.*, vol. 47, no. 8, pp. 10225–10243, Jan. 2022, doi: 10.1007/s13369-021-06548-w.

[31] D. Buenaño-Fernández, D. Gil, and S. Luján-Mora, ''Application of machine learning in predicting performance for computer engineering students: A case study,'' *Sustainability*, vol. 11, no. 10, p. 2833, May 2019, doi: 10.3390/su11102833.

[32] M. Ahamad and N. Ahmad, ''Machine learning approaches to digital learning performance analysis,'' *Int. J. Comput. Digit. Syst.*, vol. 10, no. 1, pp. 963–971, Nov. 2021, doi: 10.12785/ijcds/100187.

[33] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, ''Using machine learning to predict student difficulties from learning session data,'' *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 381–407, Feb. 2018, doi: 10.1007/s10462-018-9620-8.

[34] S. Arumugam, A. Kovalan, and A. E. Narayanan, "A learning performance assessment model using neural network classification methods of e-learning activity log data," in *Proc. Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Nov. 2019, pp. 441–448, doi: 10.1109/icssit46314.2019.8987958.

[35] V. Luca. (2014). *Educational Process Mining (EPM): A Learning Analytics Data Set Data Set*. Academic Torrents. Accessed: Jun. 18, 2023. [Online]. Available: https://academictorrents.com/details/e24e083cc337695bb84a2b68707695579c0ab4d8

[36] R. L. L. Sie, J. Delahunty, K. Bell, A. Percy, B. Rienties, T. Cao, and M. de Laat, "Artificial intelligence to enhance learning design in UOW online, a unified approach to fully online learning," in *Proc. IEEE Int. Conf. Teaching, Assessment, Learn. Eng. (TALE)*, Dec. 2018, pp. 761–767, doi: 10.1109/tale.2018.8615283.

[37] G. Latif, A. Anezi, D. N. F. A. Iskandar, A. Bashar, and J. Alghazo, "Recent advances in classification of brain tumor from MR images—State of the art review from 2017 to 2021," *Current Med. Imag.*, vol. 18, pp. 903–918, Aug. 2017. [Online]. Available: http://www.eurekaselect.com/article/120229

[38] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/a:1010933404324.

[39] C. Manapragada, G. I. Webb, and M. Salehi, "Extremely fast decision tree," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1953–1962, doi: 10.1145/3219819.3220005.

[40] S. Bernard, L. Heutte, and S. Adam, "On the selection of decision trees in random forests," in *Proc. Int. Joint Conf. Neural Netw.*, Jun. 2009, pp. 302–307.

[41] H. Yin, M. Xue, Y. Xiao, K. Xia, and G. Yu, "Intrusion detection classification model on an improved K-dependence Bayesian network," *IEEE Access*, vol. 7, pp. 157555–157563, 2019, doi: 10.1109/access.2019.2949890.

[42] R. Guo, D. Fu, and G. Sollazzo, "An ensemble learning model for asphalt pavement performance prediction based on gradient boosting decision tree," *Int. J. Pavement Eng.*, vol. 23, no. 10, pp. 3633–3646, Aug. 2022, doi: 10.1080/10298436.2021.1910825.

[43] L. Wen and M. Hughes, "Coastal wetland mapping using ensemble learning algorithms: A comparative study of bagging, boosting and stacking techniques," *Remote Sens.*, vol. 12, no. 10, p. 1683, May 2020, doi: 10.3390/rs12101683.

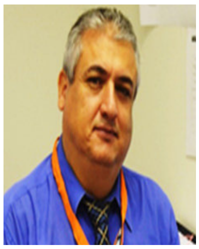
[44] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/bf00058655.

[45] G. Latif, F. Y. Al Anezi, M. Zikria, and J. Alghazo, "EEG-ECG signals classification for arrhythmia detection using decision trees," in *Proc. 4th Int. Conf. Inventive Syst. Control (ICISC)*, Jan. 2020, pp. 192–196.

[46] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *J.-Jpn. Soc. Artif. Intell.*, vol. 14, p. 1612, Sep. 1999.


**GHAZANFAR LATIF** received the B.S. degree in computer science from the FAST National University of Computer and Emerging Sciences, in 2010, by remaining on the Dean's honor list, the M.S. degree in computer science from the King Fahd University of Petroleum and Minerals, Saudi Arabia, in 2014, and the Ph.D. degree from the University of Malaysia, Sarawak, Malaysia. He was an Instructor with the CS Department, Prince Mohammad Bin Fahd University, Saudi Arabia, for three years, and has two years of industry work experience. He is currently a Research Coordinator (Deanship of Graduate Studies and Research) with Prince Mohammad Bin Fahd University. He is also continuing his postdoctoral fellowship with the University of Quebec, Canada. Throughout his educational carrier, he got a number of achievements like a full scholarship for the F.Sc., B.S.-CS, and M.S.-CS degrees, and a Gold Medal in the Ph.D. degree. His research interests include image processing, artificial intelligence, neural networks, and medical image processing.

**SHERIF E. ABDELHAMID** received the B.Sc. and M.Sc. degrees in computer science from AAST, Alexandria Campus, Egypt, and the M.Sc. and Ph.D. degrees in computer science from Virginia Tech. He was also an Infrastructure Software Engineer with the Center for Open Science, Charlottesville, VA, USA. He is currently an Assistant Professor with the Department of Computer and Information Sciences, Virginia Military Institute (VMI). Before joining VMI, he was an Assistant Professor with the College of Computing and Information Technology (AAST-Smart Village Campus), Egypt. His research work spans three main fields, such as computer science, STEM education, and public health. His research interests include high-performance services-based computing solutions, novel digital educational technologies, and tools for the social network analysis of complex systems.



**KHALED S. FAWAGREH** received the B.Sc. degree in computer science from York University, Canada, the M.Sc. degree in computer science from Dalhousie University, Canada, and the Ph.D. degree from the School of Computing Science and Digital Media, Robert Gordon University, U.K. He also has over six years of industry experience in the areas of document engineering and network management. He is currently a Lecturer with the Department of Information Technology, Prince Mohammad Bin Fahd University. Prior joining PMU, he was a Lecturer with the Department of Computer Science and Software Engineering, University of Hail, Saudi Arabia. His main research interests include data mining and machine learning.

**GHASSEN BEN BRAHIM** received the Ph.D. degree in computer science from Western Michigan University (USA). He was a Systems Analyst Engineer with the Integrated Defense Systems of Boeing and a Research Visitor with the U.S. Naval Research Laboratory. He is currently an Assistant Professor of computer science with Prince Mohammed Bin Fahd University. His research interests include machine learning, computer and network security, wireless networks, QoS routing in large-scale MANETs, routing in all-optical networks, and the design and analysis of network protocols.

**RUNNA ALGHAZO** received the M.Sc. degree in rehabilitation counseling and the Ph.D. degree in rehabilitation counseling and administration from Southern Illinois University, Carbondale, IL, USA, in 2003 and 2008, respectively. She joined the Humanities and Social Sciences Department, Prince Mohammad Bin Fahd University (PMU), as an Assistant Professor of education, in 2008. She was an Assistant Professor with the Special Education Department, King Faisal University (KFU), for four years, before rejoining PMU. She is currently an Educational Researcher and a Rehabilitation Counselor. She has various publications in conferences and journals. Her most current work focuses on the psychological and systemic variables that may contribute to students' academic success in higher education. Her research interests include special education, accommodations for students with disabilities, assistive technology, inclusion, cognitive psychology, and psychological studies on university students. She was on the Dean's List for best academic performance with SIUC throughout the master's and Ph.D. degrees and received the Guy A. Renzaglia Scholarship for best academic performance and the Dissertation Research Award, in 2007.