IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 July 2023, accepted 10 August 2023, date of publication 14 August 2023, date of current version 18 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3305276

RESEARCH ARTICLE

Utilizing Machine Learning Models to Predict Student Performance From LMS Activity Logs

MAJID KHAN[1], SABA NAZ[1], YASHIR KHAN[2], MUNEEB ZAFAR[3], MAQBOOL KHAN[4,5], (Senior Member, IEEE), AND GIOVANNI PAU[6], (Member, IEEE)

[1]International Business Machines Corporation (IBM), Johannesburg 2196, South Africa
[2]College of Industrial Economics, School of Economics and Management, Anhui Polytechnic University, Wuhu, Anhui 241000, China
[3]10 Pearls, Islamabad 44010, Pakistan
[4]Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology, Haripur 4232, Pakistan
[5]Software Competence Center Hagenberg, 4232 Hagenberg, Austria
[6]Kore University of Enna, 94100 Enna, Italy

Corresponding author: Giovanni Pau (giovanni.pau@unikore.it)

ABSTRACT In recent years, the application of data mining techniques on educational data has grown in importance. Educational data mining can be used to find hidden patterns in students' academic conduct and predict future success by examining previous data. Because more technical tools are being used to enhance the learning environment, including learning management systems (LMS), the importance of educational data mining is growing for educational institutions. The purpose of this study is to employ data mining techniques to analyse pupil behaviour patterns and predict how well they would perform academically. According to the findings of this study, there is a considerable correlation between student performance and a number of different factors, such as resource (page) views, activity gaps, grades from the previous semester, grades from prerequisite courses, and evaluations of first-term tests. Teachers and educators can use this study to spot students who need extra assistance so they can intervene.

INDEX TERMS Activity logs, data mining, sentiment analysis, machine learning, neural networks, emotion recognition, access control, LMS.

I. INTRODUCTION

The significance of education data mining is getting progressively more relevant for educational institutions as a result of the expanding number of software solutions for improving the classroom environment, typically referred to as "e-learning," "collaborative learning," or generally the LMS [6], [31]. It is a widely held concept that by analysing educational data, patterns and reasons for a specific student's success or failure can be discovered, aiding management in the process of remedial decision-making [26].

Every student's performance history is recorded in institutions that have implemented e-learning and campus management solutions, typically in a variety of formats. According to this data analysis, one in four university students do not complete their degrees [23]. Table 1 shows a comparison between dropout rates of different countries. In a lot of developing nations, the situation is significantly worse. The same table displays the dropout rates of the students chosen from two degree programs at Pakistan's National University of Science and Technology (NUST) for this study.

TABLE 1. Insight of student dropout rates in different countries vs. selected student group.

Country	Dropout Rate
USA	25-28 %
England	10-13 %
Switzerland	10-15 %
Norway	10-15 %
Japan	<15 %
Germany	10-13 %
South Korea	<15 %
Pakistan, NUST	**Dropout Rate**
MS-IT	30 %
BS-IT	23 %

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

Many factors such as willingness, parents' backgrounds, primary school, confidence, and participation in social activities, are very important in determining a student's behaviour [1], [7], [10]. Only a small amount of research has been done to identify aspects other than the obvious that influence a student's academic achievement. By analysing the data from the past and current semesters, this study aims to forecast a student's performance at any point in their academic program [2].

Our core contribution can be summarized as follows;

- The study aims to predict student performance at any stage of an academic program by analyzing data from previous and current semesters.
- This Research analyses unique student activities such as resource views, activity gaps, previous semester grades, grades in prerequisite courses, and evaluation of first-term exams to establish the correlation of these activities with student performance.
- Educators can utilize the findings of this research to identify students who require special attention and implement appropriate measures.

The existing literature highlights significant opportunities for improvements within the field of Educational Data Mining (EDM). However, three key features have yet to be adequately addressed in the literature. These missing features represent potential areas for further research and development. Addressing these gaps could contribute to the advancement and enrichment of the EDM field. The following three features are missing in the literature.

- Previous research is mainly focused on a single course or two and doesn't span over a whole degree program in which a student is enrolled.
- A methodology that can predict a student's performance at any stage of the degree program has not been established yet.
- Input attributes or features used for modeling the classifier, mostly have an obvious relationship with the predicted value, such as using course grades to predict semester GPA. Opportunity for finding hidden patterns beyond the obvious is still present.

In the subsequent sections, the design and implementation of a series of classifiers that tackle the above three aspects of educational data mining are discussed.

The paper is organized as follows: Section II critically evaluates published literature. The strengths and weaknesses of previously presented models are noted, and comparative studies are discussed. The fundamental structure of Knowledge Discovery in Databases, data collected, data preparation, cleansing, integration, and attribute retrieval/selection are discussed in Section III. This section also covers the methods and technologies that are employed in the data pre-processing and representation processes. Section IV discusses Model pattern analysis, how and which data attributes were chosen for modelling. To choose the model that best fits our data, many modelling strategies were employed. The top-performing model Patterns and rules are also provided at the end of this section. Section V provides the findings and results. Section VI brings the process to a close and concludes the outcome of this research study along with the future direction.

II. BACKGROUND AND LITERATURE REVIEW

Finding novel patterns in a huge mass of information is known as data mining [42]. More recently, the application of data mining techniques on educational data has gained notable importance [6]. A systematic literature review is conducted to explore predictive analysis tools in higher education, with a specific focus on highlighting the most pertinent instances of predictors and early warning systems employed in practical applications [22]. Intriguing trends regarding the learning process and its results may be found using the data supplied by E-learning systems [5], known as Educational Data Mining (EDM). Data mining and machine learning techniques are used to identify and monitor student performance, teacher effectiveness, and other educational outcomes. By using data to guide decision-making and pinpoint areas for improvement, education data mining aims to increase the efficacy and efficiency of education. Predicting student performance, spotting at-risk kids, and enhancing course design are some specific applications of educational data mining. The recognized patterns aid in decision-making and serve as a foundation for forecasting future trends [31], [45]. More specifically, EDM can help in four areas, 1) building models, and defining student characteristics; 2) discovering the effectiveness of the support provided by the e-learning software; 3) improving models for the knowledge structure of the domain; and 4) scientific discovery regarding learners and learning [32].

Decision tree induction is the most common method used in EDM. In a study conducted in India on 50 university students, features such as the CGPA of the last semester, grades of quizzes and assignments of the current semester, class attendance frequency, grades in lab work, general proficiency, and final exam marks to establish student behavior and predict student performance [30]. It indicates that grades or marks earned in course activities have a direct impact on student's overall performance in any course. Some unconventional activities like the participation of students in discussions, such as posting questions and answering corresponding messages, may have a significant impact on student performance. An analysis conducted on an online business course with 17,934 server logs of 98 undergraduate students concluded that low levels of participation lead to a higher risk of poor performance [20].

Association rule mining can be used to identify patterns in student performance data that may indicate why certain students are more successful than others [32]. For instance, an association rule might reveal that students who spend a certain amount of time studying each week tend to get higher grades [36]. Association rule mining can be used to identify factors that are associated with student retention and dropout

rates [38]. An association rule might reveal that students who live on campus are more likely to remain enrolled in school than those who commute. Association rule mining can be used to identify patterns in student behavior that can inform the design of personalized learning experiences [29]. Such as an association rule might reveal that students who prefer hands on learning experiences tend to perform better in certain subjects. Association rule mining can be used to identify patterns in student performance data that can inform the design of courses and course materials. Thus, an association rule might reveal that students who engage with certain types of educational materials tend to perform better on exams. Association rule mining has also been used to discover patterns in the LMS logs [13]. A dataset of 29 students was used to predict final exam marks based on assignments and quizzes, as well as page views on the discussion forum.

University of Windsor, Canada study identify that student performance had a direct relationship with assignments. An association rule mining algorithm was used on a dataset from CLEW (Collaboration and Learning Environment Windsor) to show the association between assignments and the final marks of students. The results indicate that the weight of assignments had a positive impact on final marks, and assignments should be given the right priority [15]. The relationship between online presence and student performance in a blended course by analyzing student log data. The study revealed that both the frequency and duration of online presence had a statistically significant impact on student's final grades [35]. A similar conducted study revealed a strong correlation between student attendance and academic performance. Students with more than 60% attendance tended to achieve Good (37.7%), Very Good (32.1%), and Excellent (18.9%) academic grades compared to other categories of academic achievement. Additionally, the study indicated that assignments and exercises had a significant impact on undergraduates' final grades, as determined by the logistic regression model [16].

One of the most comprehensive studies used data from seven courses and 438 students [41]. Various data mining and statistical techniques were applied using Weka and Keel data mining tools. The outcome of this effort was to integrate a user-friendly data mining capability with Moodle. Weka has been used in other EDM efforts as well [39].

Student behavior modeling using machine learning is the process of using machine learning algorithms to analyze and predict student behavior based on data about their past actions and characteristics [44]. This can involve using machine learning techniques to identify patterns and trends in student behavior data, and to make predictions about how a student is likely to behave in the future. Machine learning algorithms can be used to analyze student behavior data in order to identify students who may be at risk of academic learning difficulties or dropping out. This can allow educators to intervene early and provide targeted support to help these students succeed [3]. Machine learning algorithms can be used to analyze student behavior data in order to identify patterns and trends that can inform the design of personalized learning experiences. This can ensure that students receive learning experiences that are tailored to their needs and preferences [43].

Machine learning algorithms can be used to analyze student behavior data in order to identify patterns and trends that can inform the design of courses and course materials. This can help ensure that courses are designed in a way that is most likely to engage and motivate students [17].

Radial Basis Function (RBF), a neural network technique was used to predict student performance [9]. The dataset was taken from a Chinese University, that included the information on marks obtained during 2010-11 and 2011-12 sessions and also previous yearś marks to predict the current semester's subject marks. On the basis of prediction the students were divided into different categories with respect to their performance. Machine learning algorithms can be used to analyze student behavior data in order to make predictions about how a student is likely to behave in the future. This can help educators anticipate and respond to potential issues before they arise [33].

The dataset of 300 undergraduate students from 2003 to 2012 is taken from the University of Illinois, USA, for this study. Students' attributes of age, sex, race, citizenship status, and student grades were used to build the Bayesian Network model to predict students' academic performance. Model the student's grades in the three major courses of the second semester and work as an alarm system for students at risk [34].

A predictor based on the Naïve Bayes algorithm modeled on a dataset of 300 records was obtained from the Bachelor of Computer Application for the 2009-2010 session [11]. The study established that student grades depend upon attributes like previous academic performance, living location, and medium of instruction. Other contributing attributes include gender, family size, annual income status, food habits, college type, parents' qualifications, and occupation.

The K-Means clustering algorithm was used to group students into high, medium, and low achievers using attributes like previous grades, GPA, number of students, and percentage [37]. Result summaries show that 8.33 percent of students required special attention. Study [27] also conducted an analysis of the effects of the two variables on students' academic performance by employing K-Means clustering techniques.

The University of Tuzla collected 257 records from the faculty (Economics) to compare the performance of Naive Bayes, decision tree and Multilayer Perceptron algorithm over attributes such as gender, family size, distance from school, GPA, scholarships, entrance exam marks, materials (books, notes), time (study hours), internet usage and earnings [28]. The performance of each algorithm was assessed based on three criteria i.e. prediction accuracy, error rate and learning time. Naive Bayes predicted more instances correctly and also performed better in prediction accuracy as compared to others. Decision Tree and Naïve Bayes performed equally well w.r.t learning time.

Introducing a framework for e-learning and employed various machine learning algorithms to predict the most beneficial e-learning sessions for students [18]. The study identified Deep Learning and Random Forest algorithms as suitable for predicting useful sessions in e-learning. From the prediction results, the research also derived factors that influence session effectiveness, such as the study environment, family commitment, and teaching style. Fuzzy probabilistic neural networks was utilized to study students performance and behavior using attributes like merit, study behavior, class behavior, interest, belief, and family background [8]. By raising the teaching standard, the students might not understand anything taught in the class, while keeping it too low may cause non-seriousness and lack of interest. The cross-validation method improved the prediction results. A deep learning approach to explore student academic performance and employs regression analysis to accurately forecast their outcomes. The dataset consists of 10140 records with 9 attributes from students who have previously completed their academic activities, sampled from three distinct colleges. With k = 3, the deep learning model achieves a mean absolute score (mean absolute error) of 1.61 and a loss of 4.7 [19].

The reviewed literature reveals ample room for enhancements in the field of Educational Data Mining (EDM). Further advancements and developments are needed to maximize its potential. Future research should focus on addressing these opportunities for improvement in EDM.

III. DATA PREPARATION AND REPRESENTATION

This section focuses on the structure, cleaning, integration and attribute retrieval/selection process of LMS data. How and what kind of tools and techniques are used to achieve the data pre-processing and representation process, is also discussed.

A. KNOWLEDGE DISCOVERY IN DATABASE (KDD) CYCLE

Knowledge Discovery in Database (KDD) is a cycle or set of iterative steps that need to be followed to complete the process of Knowledge Discovery [4]. The main steps of Knowledge Discovery process are data pre-processing, data mining or model learning, and pattern analysis as shown in Figure 1.

All the steps mentioned in the figure 1 will be discussed in the subsequent sections.

B. DATA PRE-PROCESSING AND FEATURES EXTRACTION

1) DATA SET/PARTICIPANTS

A dataset of activity logs from LMS was collected for undergraduate students. All courses of the Bachelor of Information Technology (BIT) program are included for evaluation. Each course is treated as a separate dataset and is used to build a model per course.

2) DATA CLEANING

It involves techniques that involve converting raw data into an understandable format. Real-world data is often partial, inconsistent, or deficient in certain behaviors, and is prone to many errors. Data cleansing is an interactive approach, as diverse sets of data have different sets of laws determining the validity of data [25]. Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.

TABLE 2. The original data from the LMS log.

Field	White Space	Null Value	% Complete	Valid Records
Course	0	0	100	2678
Time	0	0	100	2678
IP Address	0	0	100	2678
User Full Name	0	5	99.8	2673
Action	0	0	100	2678
Information	95	7	96.1	2576

TABLE 3. The attributes list after pre-processing.

Field	White Space	Null Value	% Complete	Valid Records
Section	0	0	100	65
Name	0	0	100	65
Active Days	0	0	100	65
Total Activities	0	0	100	65
View Gap	0	0	100	65
First Sessional Active Days	0	0	100	65
Second Sessional Active Days	0	0	100	65
Third Sessional Active Days	0	0	100	65
First Sessional Activities	0	0	100	65
Second Sessional Activities	0	0	100	65
Third Sessional Activities	0	0	100	65

A sample of the data file extracted from LMS log tables is shown in Table 2. There is not a single obvious attribute/dimension that can be directly used to train a model.

There are 94 raw log files of 47 courses, as the students were in two sections, so one file for each section. On average, each file has ten thousand records. Manually cleaning this much amount of data file was almost impossible. An automated data cleaning process was defined using C-Sharp (C#). It automatically picks the files from a directory, apply the cleaning process, derived new attributes from the existing one and make one file for every course. The attributes list after running the automatic cleaning process are shown in the Table 3.

3) DATA INTEGRATION

For each student, the data is placed into two different files. One file has the students' activity logs (after pre-processing) and the other file contains student grades. Data with different sources and representations are put together and conflicts within the data are resolved. After the cleaning process, we have 43 refined datasets one for each course. Now to integrate these files and assign a course grade for each student, an ETL process is defined using SPSS Statistics (http://www-01.ibm.com/software/ch/de/analytics/spss/). This process combines all the course data files into one file and also introduces a new key "Semester_Couse" to uniquely identify each course dataset. Some grade labels have fewer records than others, to balance the distribution of the record boasting

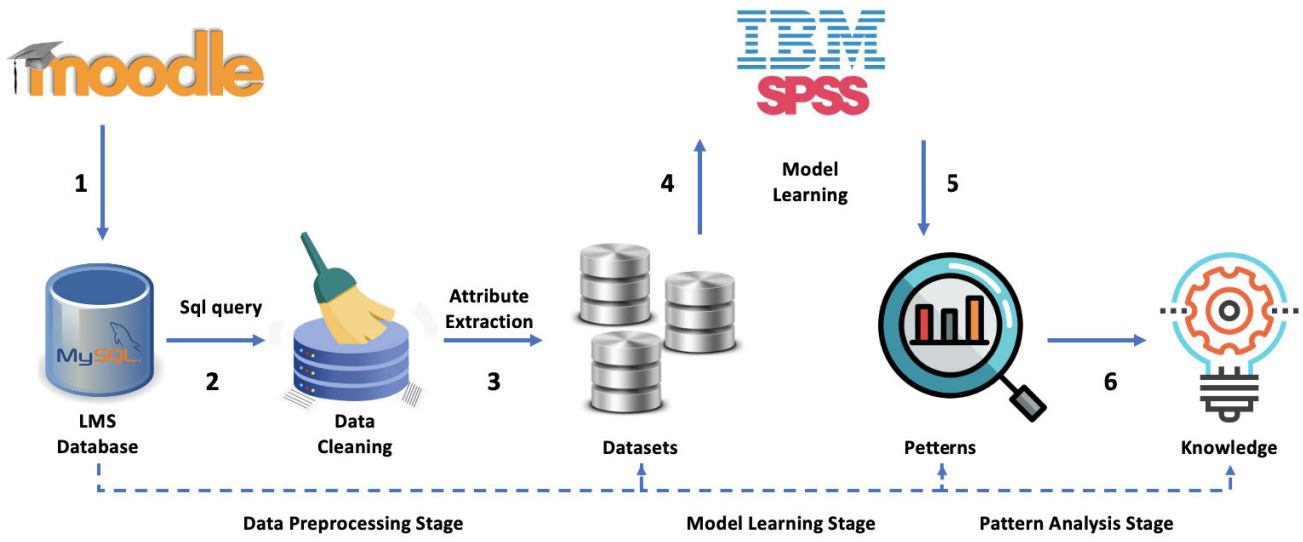


FIGURE 1. Knowledge discovery in databases (KDD) process steps.

TABLE 4. The list of attributes after data integration using the ETL process.

Field	White Space	Null Value	% Complete	Valid Records
Registration No	0	0	100	4665
View Gap	0	0	100	4665
Total Activities	0	0	100	4665
First Sessional Active Days	0	0	100	4665
Second Sessional Active Days	0	0	100	4665
Third Sessional Active Days	0	0	100	4665
First Sessional Activities	0	0	100	4665
Second Sessional Activities	0	0	100	4665
Third Sessional Activities	0	0	100	4665
Semester_Course	0	0	100	4665
Semester GPA	0	0	100	4665
Course Grade	0	0	100	4665
Previous CGPA	0	0	100	4665
Active Days	0	0	100	4665

process was also introduced. The list of attributes after completing the data integration process is shown in Table 4.

4) FEATURES SELECTION

The process of selecting a subset of relevant features used was carried out to construct a model. Not all the attributes are useful for model training. There are some attributes can create over-fitting or contribute negatively, so only those should be used that are relevant to target/dependent attribute. The final attribute set and their roles in the model are shown in Table 5.

The ''Course Grade'' is used as the target that has to be predicted, the registration number is used as record identifier and all other attributes are used as input. The ''Semester_Course'' attribute is used as a splitter to select separate datasets for each course.

TABLE 5. The final attribute set and their roles in model.

Field	Measurement	Role
Registration No	Nominal	Indentifier
Semester_Course	Nominal	Split
View Gap	Continuous	Input
Total Activities	Continuous	Input
First Sessional Active Days	Continuous	Input
Second Sessional Active Days	Continuous	Input
Third Sessional Active Days	Continuous	Input
First Sessional Activities	Continuous	Input
Second Sessional Activities	Continuous	Input
Third Sessional Activities	Continuous	Input
Previous Semester GPA	Continuous	Input
Previous CGPA	Continuous	Input
Active Days	Continuous	Input
Course Grade	Nominal	Target

IV. MODEL LEARNING AND PATTERN ANALYSIS

This section will explain how and what attributes of data after the preparation and pre-processing process were selected for modelling. What kind of modelling techniques were used; the dataset size; the model selected; the important predictors; kinds of patterns or rules were generated.

A. MODEL SELECTION AND TRAINING

1) DATASET SIZE

The most important factor to train a model accurately is to select accurate dataset. As our study focus is to predict student performance at any stage of the program, therefore to train the model accordingly, the dataset needs to be split course-wise. We have a dataset of 43 courses and a separate model was trained on each course data.

TABLE 6. The overall accuracy of models and no. of fields used.

Model	Overall Accuracy (%)	No. of Fields Used
C5	91.5	5
C&T Tree	85.2	11
Quest	85.2	10
CHAID	85.2	4
Logistic Regression	81	12
Neural Net	80	12
Discrimenant	75.7	12
Bayesian Network	15.7	12

2) MODEL SELECTION

"Every model is wrong but some are useful" Using this famous quotation, different classification techniques were applied on the same dataset to check the comparative accuracy. Eight different classification models were applied to the same dataset, to select the most appropriate one. Models, their accuracies, and a number of fields used to train a particular model are listed in the Table 6.

C5 has the highest accuracy. But the model with the highest training accuracy does not mean that it is the best model. Testing and validation of the model need to be performed in order to select the most suitable model. After testing and validation of the models, C&R(Classification and Regression) Tree [12] gave more accurate results. Even scoring on the previously unknown dataset, C&R Tree predicting accuracy is far better than any of the other models listed in the Table 6.

3) PREDICTOR IMPORTANCE

A list of attributes along with their roles, which took participation in model training according to the role assigned. Each feature/attribute has its impact which can be negative or positive in building a model. In below Figure 2, the bars show the contribution or importance of each attribute in building a more accurate model.

The figure shows that the first five attributes "Previous Semester GPA", "Second Sessional Activities", "First Sessional Active Days", "Previous CGPA" and "Total Activities" were the most important predictors to predict the target (Course Grade), with the first one having a much stronger impact.

4) COURSE WISE MODELS

In order to predict course results (Course Grade) for each course, a model was built per course. Each course data is separated on the basis of key parameter "Semester_Couse", which uniquely identifies each course dataset. Every model was trained, saved and the resultant model was a set of 43 models. Every course has a unique model and can predict course grade for that particular course. The model for each course along with the number of records used to train that model, number of field used and accuracy of every model is shown in the Table 7.

B. PATTERN ANALYSIS

Every decision tree model generates patterns or rules which provide a rational answer for all possible inputs. Patterns or rules are basically a set of if-else statements which provide "most likely" matching of the inputs. In classification algorithms, the result or target is a set of all possible outputs or class labels, so the end result of every rule is a class label from the set of all possible labels. One class label can have more than one rule.

There are 43 courses, so we cannot list down rules or patterns for every model. As an example, below are the decision tree rules of for course "Network Technologies (TCP-IP-Suite)".

Rules for A - Contains 3 rule(s)

Rule 1 for A (34: 1.0)
IF previous_semester_GPA > 3
AND First_Sessional_Activities > 8500
THEN A

Rule 2 for A (10: 1.0)
IF previous_semester_GPA > 3
AND First_Sessional_Activities < = 8500
AND Active_Days < = 41
THEN A

Rule 3 for A (3: 1.0)
IF previous_semester_GPA < = 3
AND First_Sessional_Active_Days < = 3

TABLE 7. The final attribute set and their roles in model.

Model	No. of Records in Split	No. of Fields Used	Overall Accuracy (%)
Semester1-Applied Physics	93	9	52.6
Semester1-Calculus-I	215	9	54.8
Semester1-Communication & Interpersonal Skills	94	9	86.1
Semester1-Discrete Mathematics	152	9	42.7
Semester1-Fundamental of Computer Programming	132	9	91.6
Semester1-Fundamental of ICT	167	9	64.6
Semester2-Calculus-II	171	11	71.9
Semester2-Electronics for IT	79	11	94.9
Semester2-Introduction to Management	96	11	93.7
Semester2-Linear Algebra	224	11	60.2
Semester2-Object Oriented Programming using C++	167	12	54.4
Semester3-Data Structures	143	11	66.4
Semester3-Database Design & Implementation	120	11	66.6
Semester3-Digital Logic Design	162	11	69.7
Semester3-Introduction to Java Programming	206	12	42.7
Semester3-Principles of Accounting	141	11	69.5
Semester3-Probability & Statistics	193	11	55.4
Semester4-Computer Architecture	94	11	92.5
Semester4-Computing Algorithms	201	11	57.7
Semester4-Data Communications	176	11	74.4
Semester4-Operating System	92	11	90.2
Semester4-RDBMS using Oracle	121	12	67.7
Semester4-Software Engineering	140	11	68.5
Semester5-Computer Networks	72	12	98.6
Semester5-Distributed Computing	66	11	98.4
Semester5-Object Oriented Software Engineering	119	12	91.5
Semester5-Principles of Marketing	83	11	91.5
Semester5-Technical Business Writing	104	11	94.2
Semester5-Web Technology-I	73	11	91.7
Semester6-Advance Database Systems	163	12	52.1
Semester6-Computer Graphics	71	11	95.7
Semester6-Enterprise Network Technologies Win2000 Linux	114	11	94.7
Semester6-Financial Management for IT Professionals	84	11	96.4
Semester6-Network Technologies(TCP/IP Suite)	94	11	97.8
Semester6-Numerical Analysis	86	11	97.6
Semester6-Web Technologies-II	92	11	100
Semester7-Advance Java with Emphasis on Internet Applications	106	12	93.4
Semester7-Computer Network Security	107	12	93.4
Semester7-Human Computer Interaction	97	11	94.8
Semester7-Professional Ethics	134	11	93.2
Semester7-Software Project Management	75	11	89.3
Semester8-Entrepreneurship	114	10	88.5
Semester8-Project Self Study	115	9	67.5

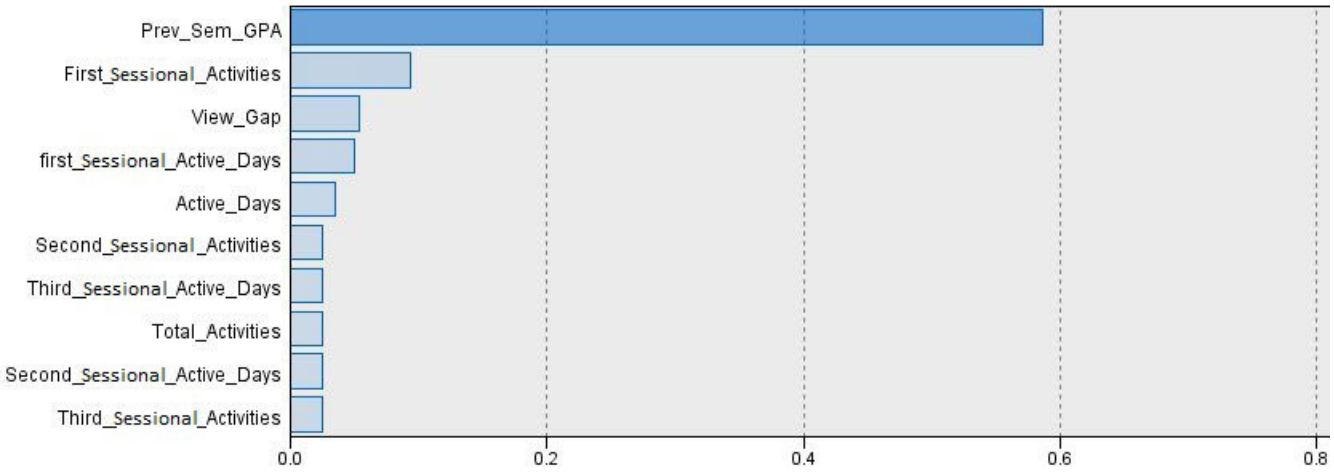


FIGURE 2. Predictor importance list.

AND View_Gap > 0.001
AND Active_Days > 30
THEN A

Rules for B - Contains 3 rule(s)

Rule 1 for B (19: 0.895)
IF previous_semester_GPA < = 3
AND First_Sessional_Active_Days > 3
THEN B

Rule 2 for B (1: 1.0)
IF previous_semester_GPA > 3
AND First_Sessional_Activities < = 8500
AND Active_Days > = 41
THEN B

Rule 3 for B (1: 1.0)
IF previous_semester_GPA < = 3
AND First_Sessional_Active_Days < = 3
AND View_Gap > 0.001
AND Active_Days < = 30
THEN B

Rules for C - Contains 1 rule(s)

Rule 1 for C (4: 1.0)
IF previous_semester_GPA < = 3
AND First_Sessional_Active_Days < = 3
AND View_Gap < = 0.001
THEN C

There are Four class labels "A","B", "C" and "F", each having one or more of the one rules to predict the label occurrence. In front of each rule, the number of records that fall under that rule, along with the confidence level. For example "Rule 1 for B (19: 0.895)", this rule predicts on 19 records with 89% confidence. Some rules might be incorrect for a particular dataset, depending on the testing and validation accuracy of that particular model.

V. MODEL TESTING AND VALIDATION RESULTS

Model and decision rules are explained in the previous section, this section will show the testing results. Those results are being validated by applying the same model rules on the unseen data. Confusion matrix, true positive rate, false positive rate, recall and precision for all the class levels is shown in this section.

A. MODEL TESTING RESULTS

In order to train the model, the dataset is split into training and testing datasets. On the training dataset, the model builds its decision rules or patterns and then test those patterns on the testing dataset to see the output. In our case, the C&R Tree decision tree algorithm is selected, and its testing results are given below in Table 9.

Total of 110 number of records that are analyze, of which 102 (92.73 %) records are accurately predict. The wrongly predict records are just 8 (7.27 %) which shows that the model accuracy is very high.

The confusion matrix is used for in-depth analysis and records distribution in different classes. It shows records that are assigned to the wrong class label. Table 10 showed the confusion matrix for the testing results.

Table 10 shows that the model accurately classified all the students of grade "A" into their respective class labels "A". Most of the students of grade "B" are accurately classified but some students (4) of Grade "B" are wrongly classified into "A". For class label "C" 4 out of 9 students having Grade "C" are wrongly classified into "B". Class label "F" is not defined because there is no record for this label so no rule is defined for "F". The True Positive (TF) rate, False Positive (FP) rate, and Recall for each class label is shown below.

B. MODEL VALIDATION RESULTS

The model testing accuracy is very high, but its validation accuracy is very important for the best model. A model cannot

TABLE 8. Model testing accuracy results.

Correct	102	92.7 %
Wrong	8	7.27 %
Total	110	

TABLE 9. Model testing confusion matrix.

Grade		A	B	C	F
A	Count	71	0	0	0
	Row %	100.00	0.00	0.00	0.00
	Total %	64.54	0.00	0.00	0.00
B	Count	4	26	0.00	0.00
	Row %	13.33	86.66	0.00	0.00
	Total %	3.63	23.63	0.00	0.00
C	Count	0	4	5	0
	Row %	0.00	44.44	55.55	0.00
	Total %	0.00	3.63	4.54	0.00
F	Count	0	0	0	0
	Row %	0.00	0.00	0.00	0.00
	Total %	0.00	0.00	0.00	0.00

TABLE 10. Model testing: True Positive (TF) rate, False Positive (FP) rate, and recall for each class.

Class	True Positive (TP)Rate	False Positive (FP)Rate	Precision	Recall
A	1	0.102	0.946	1
B	0.866	0.0006	0.866	0.866
C	0.555	0.000	1	0.555
F	0	0	0	0

TABLE 11. Model validation accuracy results.

Correct	18	64.2 %
Wrong	10	35.7 %
Total	28	

be deployed for future prediction until its validation accuracy comes up to an acceptable mark. To validate the model accuracy, the ''Network Technologies (TCP-IP-Suite)'' course dataset of the new students is selected. The model accuracy on the new dataset is given in the Table 11.

The above results show that the model predicts 64.29 percent of records correctly. The testing accuracy for most of the decision tree and neural network algorithms is very high, but in validation the accuracy of most of the algorithms is less than 40 percent. C&R Tree (Classification and Regression Tree) validation accuracy is the highest in all.

The confusion matrix for validation result is given in below Table 12.

The confusion matrix shows that most of the records of class ''A'' are accurately predicted, but in class ''B'' and ''C'' there are records that are wrongly predicted. Most of the records of class ''C'' are classified into class ''B''. There are no records classified for class ''F'', because there is no rule defined for class label ''F''. True positive (TP), False Positive (FP) and Recall for validation results are shown in Table 13.

TABLE 12. Model validation confusion matrix.

Grade		A	B	C	F
A	Count	71	0	0	0
A	Count	7	1	0	0
	Row %	87.50	12.50	0.00	0.00
	Total %	25.00	3.57	0.00	0.00
B	Count	2	10	2	0
	Row %	14.28	71.42	14.28	0.00
	Total %	7.14	35.71	7.14	0.00
C	Count	0	3	3	0
	Row %	0.00	50.00	50.00	0.00
	Total %	0.00	10.71	10.71	0.00
F	Count	0	0	0	0
	Row %	0.00	0.00	0.00	0.00
	Total %	0.00	0.00	0.00	0.00

TABLE 13. Model validation: True Positive (TF) rate, False Positive (FP) rate and recall for each class.

Class	True Positive (TP) Rate	False Positive (FP) Rate	Precision	Recall
A	0.875	0.1	0.777	0.875
B	0.714	0.307	0.714	0.307
C	0.5	0.090	0.6	0.5
F	0	0	0	0

VI. CONCLUSION AND FUTURE WORK

This study was conducted to analyse students activities and behaviors [14], to identify patterns that can help in predicting students' future performance [40]. Different data mining techniques were used to get more accurate results and make predictions on their outcomes. Activity logs from Learning Management System (LMS) were collected for undergraduate students and investigated through machine learning, data mining techniques and statistical models in an attempt to investigate how student activities, semester term activities, resource views, activities gap, previous semester grades, prerequisite course grades etc impact on the student performance. This research concludes that previous semester grades as well as first-term activities have the highest impact on student grades. The results might be employed to help students to get aware that in which course they need to focus on to improve their performance, institutes design their courseware and making process by providing information based on empirical evidence, assisting the instructor to identify the students needing special attention [24] and take desirable measures.

Though a number of valuable studies have been conducted to predict students' performance, the following three main features have only been addressed by this study.

- Instead of focusing on a course or two, this study spans the full courses of a degree program in which a student is enrolled.
- Establish a methodology that can predict a student's performance at any stage of the degree program, have it separately trained, and save a model for each course.
- Modeling was done using unsupervised predictor features or dimensions, which have no evident connection to the target or dependant value.

A. FUTURE WORK

Currently, in this research, most of the predictor attributes or dimensions that are used for modeling have an unsupervised relationship with the target or dependent value. For the future work point of view, a combined set of supervised (directly related to or affecting the target value) and unsupervised (no direct relationship with the target value) attributes will be used to predict the student's performance. Social attributes like family income, the mother's education, schooling, etc., can also play an important role in student performance. This will allow us to better understand and predict the students' performance.

Moreover, in the next subsequent work, there will be a focus on early alarm systems [21]. This means that the prediction cycle/period should be decreased and students' performance should be predicted on a sessional basis. This will help identify the students who need special attention at the start of a course. By diagnosing a problem early, it can be treated more effectively and timely.

REFERENCES

[1] J. Abbas, J. Aman, M. Nurunnabi, and S. Bano, "The impact of social media on learning behavior for sustainable education: Evidence of students from selected universities in Pakistan," *Sustainability*, vol. 11, no. 6, p. 1683, Mar. 2019.

[2] A. Q. Noori and N. Noori, "Online learning experiences amid the COVID-19 pandemic: Students' perspectives," *Academia Lett.*, vol. 2, no. 1, pp. 45–51, 2020.

[3] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, M. Bashir, and S. U. Khan, "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," *IEEE Access*, vol. 9, pp. 7519–7539, 2021.

[4] A. B. E. D. Ahmed and I. S. Elaraby, "Data mining: A prediction for student's performance using classification method," *World J. Comput. Appl. Technol.*, vol. 2, no. 2, pp. 43–47, Feb. 2014.

[5] A. Alam, "Cloud-based e-learning: Scaffolding the environment for adaptive e-learning ecosystem based on cloud computing infrastructure," in *Computer Communication, Networking and IoT*. Berlin, Germany: Springer, 2023, pp. 1–9.

[6] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics Informat.*, vol. 37, pp. 13–49, Apr. 2019.

[7] A. Alhadabi and A. C. Karpinski, "Grit, self-efficacy, achievement orientation goals, and academic performance in university students," *Int. J. Adolescence Youth*, vol. 25, no. 1, pp. 519–535, Dec. 2020.

[8] N. Arora and J. R. Saini, "A fuzzy probabilistic neural network for student's academic performance prediction," *Int. J. Innov. Res. Sci., Eng. Technol.*, vol. 2, no. 9, pp. 4425–4432, 2013.

[9] Y. Arora, A. Singhal, and A. Bansal, "PREDICTION & WARNING: A method to improve student's performance," *ACM SIGSOFT Softw. Eng. Notes*, vol. 39, no. 1, pp. 1–5, Feb. 2014.

[10] M. A. Ashraf, M. N. Khan, S. R. Chohan, M. Khan, W. Rafique, M. F. Farid, and A. U. Khan, "Social media improves students' academic performance: Exploring the role of social media adoption in the open learning environment among international medical students in China," *Healthcare*, vol. 9, no. 10, p. 1272, Sep. 2021.

[11] B. K. Bhardwaj and S. Pal, "Data mining: A prediction for performance improvement using classification," 2012, *arXiv:1201.3418*.

[12] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.

[13] M. Cantabella, R. Martínez-España, B. Ayuso, J. A. Yáñez, and A. Muñoz, "Analysis of student behavior in learning management systems through a big data framework," *Future Gener. Comput. Syst.*, vol. 90, pp. 262–272, Jan. 2019.

[14] Y.-C. Chang, W.-Y. Kao, C.-P. Chu, and C.-H. Chiu, "A learning style classification mechanism for e-learning," *Comput. Educ.*, vol. 53, no. 2, pp. 273–285, Sep. 2009.

[15] R. Chaturvedi and C. Ezeife, "Mining the impact of course assignments on student performance," in *Educational Data Mining 2013*. Citeseer, 2013.

[16] S. Gaftandzhieva, A. Talukder, N. Gohain, S. Hussain, P. Theodorou, Y. K. Salal, and R. Doneva, "Exploring online activities to predict the final grade of student," *Mathematics*, vol. 10, no. 20, p. 3758, Oct. 2022.

[17] K. F. Hew, X. Hu, C. Qiao, and Y. Tang, "What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach," *Comput. Educ.*, vol. 145, Feb. 2020, Art. no. 103724.

[18] M. Hussain, W. Zhu, W. Zhang, J. Ni, Z. U. Khan, and S. Hussain, "Identifying beneficial sessions in an e-learning system using machine learning techniques," in *Proc. IEEE Conf. Big Data Analytics (ICBDA)*, Nov. 2018, pp. 123–128.

[19] S. Hussain, S. Gaftandzhieva, M. Maniruzzaman, R. Doneva, and Z. F. Muhsin, "Regression analysis of student academic performance using deep learning," *Educ. Inf. Technol.*, vol. 26, no. 1, pp. 783–798, Jan. 2021.

[20] G. Kennedy, "What is student engagement in online learning. and how do I know when it is there," Melbourne CSHE Discuss. Papers, Tech. Rep., 2020, pp. 1–6.

[21] Z. J. Kovacic, "Early prediction of student success: Mining students enrolment data," in *Proc. InSITE Conf.*, 2010, pp. 19–24.

[22] M. Liz-Domínguez, M. Caeiro-Rodríguez, M. Llamas-Nistal, and F. A. Mikic-Fonte, "Systematic literature review of predictive analysis tools in higher education," *Appl. Sci.*, vol. 9, no. 24, p. 5569, Dec. 2019.

[23] D. Ljubobratovic and M. Matetic, "Using LMS activity logs to predict student failure with random forest algorithm," *Future Inf. Sci.*, p. 113, Nov. 2019.

[24] G. Lust, J. Elen, and G. Clarebout, "Students' tool-use within a web enhanced course: Explanatory mechanisms of students' tool-use pattern," *Comput. Hum. Behav.*, vol. 29, no. 5, pp. 2013–2021, Sep. 2013.

[25] O. Z. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, vol. 1. Berlin, Germany: Springer, 2005.

[26] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462–55470, 2020.

[27] R. Nand, A. Chand, and M. Naseem, "Analyzing students' online presence in undergraduate courses using clustering," in *Proc. IEEE Asia–Pacific Conf. Comput. Sci. Data Eng. (CSDE)*, Dec. 2020, pp. 1–6.

[28] E. Osmanbegović and M. Suljić, "Data mining approach for predicting student performance," *Econ. Rev.*, vol. 10, no. 1, pp. 3–12, 2012.

[29] N. S. Raj and V. G. Renumol, "A systematic literature review on adaptive content recommenders in personalized learning environments from 2015 to 2020," *J. Comput. Educ.*, vol. 9, no. 1, pp. 113–148, Mar. 2022.

[30] S. Ranjeeth, T. P. Latchoumi, M. Sivaram, A. Jayanthiladevi, and T. S. Kumar, "Predicting student performance with ANNQ3H: A case study in secondary education," in *Proc. Int. Conf. Comput. Intell. Knowl. Economy (ICCIKE)*, Dec. 2019, pp. 603–607.

[31] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining Knowl. Discovery*, vol. 10, no. 3, p. e1355, May 2020.

[32] S. A. Salloum, M. Alshurideh, A. Elnagar, and K. Shaalan, "Mining in educational data: Review and future directions," in *Proc. Int. Conf. Artif. Intell. Comput. Vis.* Berlin, Germany: Springer, 2020, pp. 92–102.

[33] B. Sekeroglu, K. Dimililer, and K. Tuncal, "Student performance prediction and classification using machine learning algorithms," in *Proc. 8th Int. Conf. Educ. Inf. Technol.*, Mar. 2019, pp. 7–11.

[34] A. Sharabiani, F. Karim, A. Sharabiani, M. Atanasov, and H. Darabi, "An enhanced Bayesian network model for prediction of students' academic performance in engineering programs," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, Apr. 2014, pp. 832–837.

[35] B. Sharma, R. Nand, M. Naseem, and E. V. Reddy, "Effectiveness of online presence in a blended higher learning environment in the Pacific," *Stud. Higher Educ.*, vol. 45, no. 8, pp. 1547–1565, Aug. 2020.

[36] D. Shin and J. Shim, "A systematic review on data mining for mathematics and science education," *Int. J. Sci. Math. Educ.*, vol. 19, no. 4, pp. 639–659, Apr. 2021.

[37] M. H. I. Shovon and M. Haque, "An approach of improving students academic performance by using K means clustering algorithm and decision tree," 2012, *arXiv:1211.6340*.

[38] B. K. Simon and A. P. Nair, "Association rule mining to identify the student dropout in MOOCs," *Int. Res. J. Eng. Technol. (IRJET)*, vol. 6, no. 1, 2019.

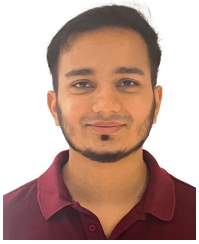
[39] A. Triayudi, W. O. Widyarto, and V. Rosalina, "Analysis of educational data mining using WEKA for the performance students achievements," in *Proc. 2nd Int. Conf. Electron., Biomed. Eng., Health Informat.* Cham, Switzerland: Springer, 2022 pp. 1–10.
[40] S. Valsamidis, S. Kontogiannis, I. Kazanidis, T. Theodosiou, and A. Karakos, "A clustering methodology of web log data for learning management systems," *J. Educ. Technol. Soc.*, vol. 15, no. 2, pp. 154–167, 2012.
[41] 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," *Comput. Hum. Behav.*, vol. 104, Mar. 2020, Art. no. 106189.
[42] W. Xiao, P. Ji, and J. Hu, "A survey on educational data mining methods used for predicting students' performance," *Eng. Rep.*, vol. 4, no. 5, May 2022, Art. no. e12482.
[43] H. Xie, H.-C. Chu, G.-J. Hwang, and C.-C. Wang, "Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017," *Comput. Educ.*, vol. 140, Oct. 2019, Art. no. 103599.
[44] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," *Comput. Hum. Behav.*, vol. 98, pp. 166–173, Sep. 2019.
[45] X. Zhang, X. Shi, Y. Khan, M. Khan, S. Naz, T. Hassan, C. Wu, and T. Rahman, "The impact of energy intensity, energy productivity and natural resource rents on carbon emissions in Morocco," *Sustainability*, vol. 15, no. 8, p. 6720, Apr. 2023.

MAJID KHAN received the B.S. degree in computer science from the University of Peshawar, in 2010, and the M.S. degree in information technology from the National University of Science and Technology, Islamabad, in 2014.

He is a highly skilled professional with extensive experience in customer value management (CVM), AI and machine learning, and telecommunication industry. With more than 13 years with IBM, he has successfully led and delivered multiple projects across Europe, Africa, and Middle-East. He is a recognized Subject Matter Expert (SME) of CVM and data analytics. He has authored numerous articles on topics, including AI and machine learning, digital marketing, and sustainable applications. His contributions have earned him several prestigious national and international awards, showcasing his dedication and expertise.

SABA NAZ received the B.Sc. degree in telecommunication engineering from the National University of Computer and Emerging Sciences, in 2012. She is a dedicated professional with a passion for the telecommunication industry. She embarked on her academic journey with the prestigious National University of Computer and Emerging Sciences(NUCES). Over the past decade, she has accumulated extensive experience and knowledge in the field, primarily focusing on enhancing IT processes within the telecommunication industry. Her dedication and commitment to her work have allowed her in identifying inefficiencies in existing IT systems and developing innovative solutions to address them. She possesses a strong analytical mindset, enabling her to analyse complex problems and devise effective strategies to optimize processes.

YASHIR KHAN received the Ph.D. degree from Southeast University, Nanjing, China. He is currently an Assistant Professor with Anhui Polytechnic University, Wuhu, China. He has published several peer-reviewed academic articles in top international journals. He is also working on various local and international projects.



MUNEEB ZAFAR received the B.Sc. degree in electrical engineering and the master's degree in computer science from the National University of Science and Technology (NUST), in 2016 and 2023, respectively. He started his educational journey with the prestigious NUST. He is a well accomplished professional with a passion for technology and computer sciences.

MAQBOOL KHAN (Senior Member, IEEE) received the M.S. degree from the Huazhong University of Science and Technology (HUST), Wuhan, in 2011, and the Ph.D. degree from Nanjing University, in 2013, China. He was with multinational companies, such as Siemens and Atos. He is currently an Assistant Professor with the Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology (PAF-IAST), Haripur, Pakistan, and an Adjunct Researcher with the Software Competence Center Hagenberg (SCCH), Austria. Recently, he completed his Post-Doctorate from SCCH, Austria. He has multidisciplinary expertise and working experience on diverse topics of big data analytics, cloud computing, predictive maintenance, explainable AI, knowledge graphs, data science, and machine learning. He has more than ten years of professional experience while working in both industry and academia. He is a certified Google Cloud Professional Architect. He is an Active Researcher working on various projects with multiple collaborators and he is currently working on European Union Project titled "Human-AI Teaming Platform for Maintaining and Evolving AI Systems in Manufacturing." He won a project during Pakistan Scientific Foundation (PSF) CRP4 call, as the Principal Investigator.

GIOVANNI PAU (Member, IEEE) received the bachelor's degree in telematic engineering from the University of Catania, Italy, and the master's degree (cum laude) in telematic engineering and the Ph.D. degree from the Kore University of Enna, Italy. He is currently an Associate Professor with the Faculty of Engineering and Architecture, Kore University of Enna. He is the author/coauthor of more than 80 refereed articles published in journals and conferences proceedings. His research interests include wireless sensor networks, fuzzy logic controllers, intelligent transportation systems, the Internet of Things, and network security. He is a member of the IEEE (Italy Section). He has involved in the organization of several international conferences, as the Session Co-Chair and a Technical Program Committee Member. He serves/served as a leading guest editor in the special issues of several international journals and the Editorial Board Member and an Associate Editor of IEEE Access, *Wireless Networks* (Springer), *EURASIP Journal on Wireless Communications and Networking* (Springer), *Wireless Communications and Mobile Computing* (Hindawi), and *Future Internet* (MDPI).

Open Access funding provided by 'Università degli Studi di Enna "KORE"' within the CRUI CARE Agreement