IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 5, 2021, accepted December 8, 2021, date of publication December 14, 2021, date of current version December 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3135309

Integrating Learning Analytics and Collaborative Learning for Improving Student's Academic Performance

ADNAN RAFIQUE[1], MUHAMMAD SALMAN KHAN[1], MUHAMMAD HASAN JAMAL[1], MAMOONA TASADDUQ[1], FURQAN RUSTAM[2], ERNESTO LEE[3], PATRICK BERNARD WASHINGTON[4], AND IMRAN ASHRAF[5]

[1]Department of Computer Science, COMSATS University Islamabad (CUI), Lahore 54000, Pakistan
[2]Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Punjab 64200, Pakistan
[3]Department of Computer Science, Broward College, Fort Lauderdale, FL 33301, USA
[4]Division of Business Administration and Economics, Morehouse College, Atlanta, GA 30314, USA
[5]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38544, South Korea

Corresponding authors: Ernesto Lee (elee@broward.edu) and Imran Ashraf (ashrafimran@live.com)

This work was supported by the Florida Center for Advanced Analytics and Data Science funded by Ernesto.Net (under the Algorithms for Good Grant).

ABSTRACT Big data analytics has shown tremendous success in several fields such as businesses, agriculture, health, and meteorology, and education is no exception. Concerning its role in education, it is used to boost students' learning process by predicting their performance in advance and adapting the relevant instructional design strategies. This study primarily intends to develop a system that can predict students' performance and help teachers to timely introduce corrective interventions to uplift the performance of low-performing students. As a secondary part of this research, it also explores the potential of collaborative learning as an intervention to act in combination with the prediction system to improve the performance of students. To support such changes, a visualization system is also developed to track and monitor the performance of students, groups, and overall class to help teachers in the regrouping of students concerning their performance. Several well-known machine learning models are applied to predict students performance. Results suggest that experimental groups performed better after treatment than before treatment. The students who took part in each class activity, prepared and submitted their tasks perform much better than other students. Overall, the study found that collaborative learning methods play a significant role to enhance the learning capability of the students.

INDEX TERMS Collaborative learning, data analytics, machine learning, learning management system, learning analytics, educational data mining.

I. INTRODUCTION

Educational institutes are striving to enhance the learning capabilities of the students and this is primarily achieved by devising novel methods and approaches for elevating the standards of the content taught to the students. Students and teachers are the key stakeholders of institutes and their success makes a huge impact on the social and economic development of a country [1]. However, in the sprint for providing the rich educational material the most important element is either ignored or completely overlooked which is analytics. Learning analytics aims at data gathering, analysis, and reporting of important indicators for students with respect to their academic performance. It provides insights about students' performance using various performance markers. However, such analysis is not possible without the temporal data of students. It is obviously very difficult for teachers to maintain that data and predict students' performance at the start of their teaching session. As a result, an efficient and user-friendly system is needed to collect the data, predict students' performance and visualize the important indicators that can help teachers take proper measures to elevate the learning ability of the students.

Learning analytics also known as data analytics in education is rapidly gaining growing attraction in education administration, besides other domains. It helps to

The associate editor coordinating the review of this manuscript and approving it for publication was Wai-Keung Fung.

integrate traditional teaching methods with ICT technology to improve teaching and learning quality in large institutions. Learning analytics provide powerful tools to help teachers improve the effectiveness of their courses and enhance students' performance through an iterative process. The key purpose of learning analytics is to provide a structure for educational institutes and administrators to standardize student data collections to analyze their academic success.

Learning analytics is the assessment, compilation, interpretation, and monitoring of students within the provided settings [2], [3]. It is considered an effective method because it helps to predict individual students' success, improve future course outcomes, boost the retention rate, improve the overall quality of teaching and provide better support in decision making [3], [4].

Learning analytics improves the performance of a student by identifying the factors that influence their performance. For this purpose, the data is gathered from educational institutions or by using a questionnaire to evaluate student progress in the early weeks of the course. After identifying the student at risk, teacher then can execute timely intervention and additional training for the weak students [5]–[7]. A solution to introduce early intervention is developed where the instructor can incorporate learning analytics for providing feedback to a student in real-time.

Learning analytics is used in different ways to predict student performance such as the speed at which students grasp the concepts from the courses taught during a semester [8]. Similarly, students' performance can be predicted from students' digital footprints i.e., demographics, behavioral logs, and face emotion monitoring while using an intelligent tutoring system [8]–[10].

Collaborative learning is a scenario where students work or learn in small communities in which they can collaborate and learn with respect to each other's expertise. Group work impacts many facets of a group project, such as working together to solve challenges, completing assignments, or discovering new ideas. The collaborative structure is established to access the students' teamwork behaviors, cooperation, individual effort, common experience, and self-adopted teamwork roles.

Collaborative learning has been regarded as important owing to its capability of incorporating powerful strategies to enhance learning. The implementation of collaborative learning depends upon teacher perceptions and opinions of collaborative learning. There are a lot of discrete ways of organizing student groups, such as teacher-selected groups and student-selected groups [11], [12]. In student-selected groups, the group structure is often homogeneous where the high performing students, low performing students, or the students from the same-sex may form a group. On the other hand, in the case of teacher-selected groups, the composition is often heterogeneous involving the students with different skill levels or sex groups. Teacher-selected groups perform much better because teachers are strongly skilled to form a group of relative students to enhance their learning capabilities.

The computer-supported learning process has attracted considerable interest in the education field over the years. In addition to offering a more flexible way for students to learn skills, it encourages students to collaborate closely with their peers. In traditional learning especially in Pakistan, teachers face problems dealing with a large number of students. One way to overcome this problem is to support collaborative learning through online education portals and learning systems [13]–[15]. Interestingly, a lot of work has been done to predict the learning outcomes of students and future performance with respect to e-learning tools [7], [16]. However, very limited work has been done in the conventional learning domain, especially for Asian students and particularly for Pakistani students [17]. This study investigates the effects of group composition, collaborative learning behavior, and academic performance in a computer-supported collaborative learning environment, especially in the context of Pakistan. The objective of this research is to fill this gap and develop an integrated framework to provide collaborative learning. In a nutshell, this study makes the following contributions

- A framework is designed that can highlight the students at-risk during the early weeks of a course and assist weak students by promoting collaborative learning. It provides a visualization system to track and monitor the performance of students, groups, and overall class to help teachers in the regrouping of students concerning their performance.
- A large dataset is gathered using a questionnaire from the COMSATS university Pakistan undergraduate students. Several machine learning classifiers are applied such as logistic regression (LR), K nearest neighbor (KNN), random forest (RF), support vector machine (SVM), naive Bayes (NB), decision trees (DT), and an ensemble-based model, to predict the future performance of the students.
- This research explores the potential of learning analytics and collaborative learning for an introductory programming course. It evaluates the potential of collaborative learning as an intervention to act in combination with the prediction system to improve the performance of students.

To achieve the above-mentioned goals, this study formulates the following three research questions:

1) Whether existing identified factors for students' performance prediction are valid in the local context of Pakistan?
2) Which are the most important features that help accurate prediction of students' performance?
3) Is it possible to improve the performance of a student by collaborative learning?

The rest of the paper is divided into four sections. Section II discusses several important works related to the current study. The proposed framework is described along with its modules in Section III. Results and discussions are provided in Section IV while Section V concludes this study.

II. RELATED WORK

The general motivation for this research is supporting higher education students and institutions to achieve better academic performance. On different occasions, prominent work has been done for learning analysis in educational sciences, such as predicting course outcomes, identifying students at risk, automating tutoring, detecting student characteristics, predicting learning behaviors, determining the level of engagement, and much more. Since this study incorporates both learning analytics and collaborative learning, the research works related to both these topics are discussed here.

A. LEARNING ANALYTICS

The authors [3], [18] provide a brief overview of learning analytics, its use in educational institutions, available learning analytics tools, and how students' performance can be monitored. In addition, the study elaborates on several challenges associated with using learning analytics concerning higher education. Studies aim at highlighting the data that instructors can use to improve the students' performance. Similarly, the authors focused on the understanding of learning analytics applications, testing approaches, methodology, and use of learning analytics in advanced studies [4]. The goal of this study is to validate four hypotheses regarding learning analytics as to whether learning analytics boosts learning outcomes, facilitates learning and teaching, and its use is comprehensive and ethical. Overall, the findings suggest that the evidence of developments in student outcomes is supported by 9%, teaching and learning assistance by 35%, comprehensive use by 6%, and ethical use by 18%. Study [5] investigates the use of various machine learning algorithms to predict the students' performance in the early weeks of the course. For this purpose, NB, DT, RF, SVM, neural network (NN), CN2 rules, and KNN are used to analyze their performance on collected data from 76 students. Results showed that the KNN algorithm shows better performance at the end of the term with 89% accuracy and 74% accuracy in the early weeks.

The authors propose an approach to identify students at risk at the start of the course in [6]. A predictive framework is developed to estimate student final grades in the third week of the course by using data gathered for teachers. The framework is applied to five courses taught by three different professors at two different universities. Various standard machine learning techniques such as RF, SVM, and AdaBoost are applied to simulate the success of two different courses and then transfer-learning techniques are applied to analyze their performance. Results indicate that RF performs well with respect to F1- scores in conventional machine learning systems but transfer learning models perform significantly better than conventional models. Similarly, the authors propose a new PredictCS tool in [19] to automatically detect and give feedback to low-performing learners in a programming course. PredictCS takes students' profiles, prior academic outcomes, programming course assignments, and contextual experiences as inputs to predict the success of a student. Results show that laboratory grades and overall performance is improved for those students who modify their projects and resubmit them based on the feedback. In a similar fashion, the authors in [20] identify students facing the risk of poor performance and suggest feedback to improve the probability of success. Additionally, students receive personalized feedback to avoid possible failure and increase the chances to pass the course. The data gathered for 53000 students from the online platform is used with a gradual at-risk (GAR) model. The case study showed that a small improvement in performance is found but results do not show conclusive evidence that such improvements are based on the early warning system (EWS). The authors investigate predicting classifiers in [21] that can be used to analyze students' data. Three machine learning are developed for prediction, such as linear regression, DT, and NB. Results suggest that NB can obtain the highest accuracy for predicting high school students' scores for a math course.

For the early estimation of the final academic success of students, [7] recommends blended learning comprising both traditional and online learning. The study also identified some critical factors that affect students' performance. In addition, the study investigates the efficacy of various datasets, i.e., online vs traditional vs blended for predicting the final performance of students. For analysis, data from 59 students with 21 variables are collected. In five datasets of various weeks, principal component regression (PCR) is used to estimate the final academic success of students. The findings reveal that the final academic success of students in a merged calculus course can be estimated by a dataset comprising data from weeks 1-6 with high accuracy. Moreover, the study also found that the blended dataset had a higher predictive performance. The study [22] makes use of several learning techniques using both tree-based and artificial neural network (ANN) models to find important factors concerning university studies. A dataset is gathered for this purpose that comprises 120 university students. The frequency of university resources utilization is found to be the most significant factor to impact students' performance. Similarly, the number of clicks on university education resources is the highest impact factor for university students' performance. It suggests that interaction with education resources and educational platforms has higher significance.

Making accurate predictions about students' final grades/success has several associated challenges and data imbalance is one of the important factors that can influence the prediction accuracy. In this regard, [23] compare various resampling techniques such as borderline synthetic minority oversampling technique (SMOTE), random over sampler, SMOTE, SMOTE-edited nearest neighbor (SMOTE-ENN), SVM-SMOTE, and SMOTE-Tomek to handle the imbalanced data problems. For performance comparison, various machine learning classifiers including

RF, KNN, ANN, XG-boost, SVM, DT, LR, and NB are used. Furthermore, the random hold out and 5-fold cross-validation methods are used to validate the used models. Additionally, the results of the Friedman test suggest that the SVM-SMOTE shows significant performance as compared to other methods. Results indicate that RF outperforms all other models when trained with the data resampled using SVM-SMOTE. Similarly, the authors in [24] use an imbalanced dataset for predicting student performance with SMOTE oversampling techniques. The study produces clusters with affinity propagation (AP) SMOTE which are later used for oversampling. The dataset comprising of 10 features and 2112 instances with an imbalance ratio of 17:85 is used and the data are resampled using SMOTE and AP SMOTE with J48 and NB classifiers to predict the student at risk. F measure, G-mean, and areas under the curve (AUC) are used to evaluate the performance where results show that AP SMOTE outperforms the original SMOTE for the NB and J48 classifiers [1].

The study [25] observes the factors that affect the student's performance using statistical analysis techniques. The study uses a dataset of 400 students collected over a period of two semesters with the proposed ensemble meta-based tree model (EMT). Experimental results show that the proposed EMT gained a high accuracy of 98.5% which is superior compared to the other techniques. Similarly, [26] obtained data from 141 students in the University of West Scotland using a student record system, learning management system (LMS), and survey. The authors compare the classification accuracy of DT, NN, SVM, and an ensemble model. PCA is applied to identify the appropriate features that should be used with the models. Seven models are created using different combinations of variables from different information sources. The ensemble technique using variables from the three sources showed the best accuracy of approximately 80%. The authors make early predictions of students at risk of low performance in [27]. The study identifies the students with a high probability to withdraw and highlights important factors to enhance the performance of students. Furthermore, the effectiveness of ANN is evaluated using handcrafted features. The features are derived from virtual learning environments and are used to predict the risk of students' poor performance. Additionally, appropriate remedies to elevate students' performance are also introduced. The results show that the proposed model achieves a classification accuracy of 84%–93% and outperforms SVM and LR by 4.3% and 8.6%, respectively.

The study [28] takes one step ahead and predicts applicants' academic performance before admitting them. A data set of 2,039 students is used to validate the proposed methodology for the enrolled students in a Saudi public university. Four prediction models ANN, DT, SVM, and NB are used to train the model. The results show that SAT test score which is the pre-admission criterion makes the most accurate prediction of students' future performance.

The author utilizes a deep LSTM model in [29] on a collection of features obtained from video clickstream data. The underlying goal is to predict learners' weekly performance and allow teachers to initiate appropriate measures for timely intervention. The dataset is collected from two independent MOOC-enabled Computer Science courses and contains the learners' interaction logs and assessment grades. The proposed LSTM model outperforms baseline ANNs, SVM, and logistic regression with an accuracy score of 93%. Similarly, study [30] used the RF to predict student dropout in the self-paced MOOC course. The data was generated from a self-paced math course offered on the MOOC platform. The designed model can predict the probability of students' dropout in the MOOC course, with an accuracy of 87.5%, the precision of 88%, recall of 87.5%, F1-score of 87.5%, and AUC of 94.5%.

Along the same lines, the author in [22] used different automated learning algorithms, including tree-based models and ANN. The study tested these methods on a group of 120 students who were pursuing a master's degree in the virtual learning environment (VLE). The algorithms used in this study include DT, RF, EGB, and MLP. The best-performing model is MLP with an average accuracy of 78% with cross-validation. The findings suggest that the number of clicks on the VLE resources, as well as, engagement with the resources and tasks with the educational platform has the greatest impact on university students' performance. The author uses KNN, LR, SVM, NB, DT, ANN, and ensemble-based techniques to increase student success in [31]. The data is obtained from 1491 students from different UAE academic institutions. Performance is measured using accuracy and kappa coefficients. Results indicate that a 75% accuracy using machine learning approaches can be obtained with a Kappa of 0.5 while the ensemble models achieve the precision of 83%.

B. COLLABORATIVE LEARNING

Collaborative learning has gained large attention due to its potential role in effective learning. Collaborative learning or group-based learning environment provides several ways of predicting student performance despite its challenges and difficulties. The authors investigate the challenges and disputes for effective implementation of collaborative learning with respect to teachers and students [11], [32]. The study involves 19 teachers and 23 students from different disciplines at a Vietnam university and interviews them using a semi-structured interview method. The participants are chosen through a snowball sampling technique. The interview aims at determining the objectives of collaborative learning, student preparedness for collaboration, the process of collaborative learning, and learning assessment. The authors apply a grounded theory approach to analyze the interviewed data with the help of Nvivo10. Results suggest that the effectiveness of collaborative learning is affected by students' inability to indulge in collaborative learning, free-riding, competence status, and friendship. Similarly, three interrelated antecedents are found including 'teachers set collaborative learning goals', 'provided instruction for collaborative skills', and 'assessed student collaboration'.

Similarly, the authors in [33] found indicators based on students' interaction and involvement for the assessment of the individual progress within the framework of teamwork. The findings show that the final grade depends on the participation of individuals and the direct relationship across passive and active associations.

The study [34] performed experiments to test the feasibility of online classes for collaborative learning. With the data divided into experimental and control groups, parallelized point-based (PPbA) for massive open online courses (MOOCs) and parallelized action-based (PAbA) for the collaborative environment are applied to measure the level of engagement. Results indicate that both the MOOCs and interactive environments have higher levels of self-regulation and interaction with experimental group students. Similarly, the authors in [35] explored an established collaborative structure to access the student's teamwork behaviors, cooperation, common experience, and self-adopted teamwork roles. To this end, they introduced an automated framework for guiding in real-time to help educators evaluate coordination fairly and effectively. The findings of the study helped to develop a system for manual interpretation of conversations on coordination.

Despite several important works in collaborative learning in general, very little work has been done within the context of Pakistan.

For instance, the authors in [12] devised a learning approach to support collective learning methods to uplift the performance of weak students. The study showed significant variations in the performance of low achievers in their final term exams compared to their mid-term where they achieve high grades. The findings suggest that the collaborative learning engagement of peer group participants is beneficial for enhancing the learning success of weak students. Similarly, the authors in [13] evaluated the impact of collaborative learning on student achievement. The students are divided into two groups to see the disparity between the two groups and a post-evaluation is carried out. Different collaborative learning techniques are used with the experimental group, including STAD, TGT, and Jigsaw II. Results show that the experimental group performs better than a control group and collaborative learning methods have a significant effect on student success.

The influence of cooperative learning on the academic performance of Asian students is evaluated in [36]. The students are divided into five groups for the analysis. The findings of the study reveal that collaborative learning has a beneficial effect on students' achievements, as well as improving participation, school and class engagement, encouragement, and freedom. Similarly, [37] analyzed the perceptions of Asian students about collaborative learning in discussion and group projects. For this purpose, one-hour face-to-face formal survey interviews of students are organized. Experimental results show that Asian students

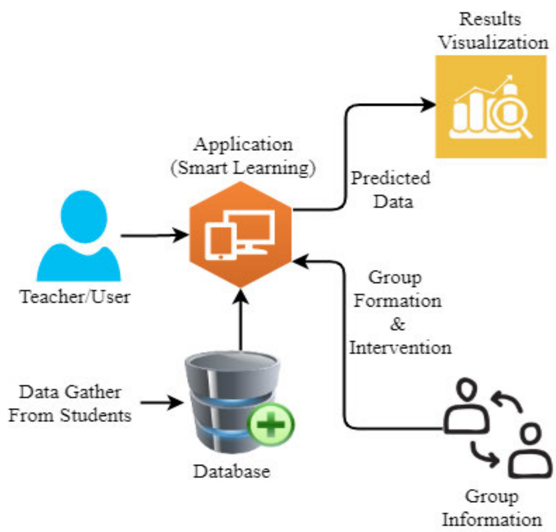


FIGURE 1. Modules of the proposed framework.

highly appreciate the importance of group work in classrooms where they can communicate with students from other cultures, improve learning skills and talents.

The above-discussed research works reveal that predominantly the work has been done concerning online learning environments for predicting students' performance and timely intervention but very limited work has been done for traditional learning environments. Therefore, this research explores various personal and academic information of students that can help in predicting the performance of students. Owing to the burden of the teachers, especially in the context of Pakistan, this study makes use of already known features of the students to predict their performance for conventional class environments. Additionally, an efficient and user-friendly tool is developed to help teachers in defining and managing student groups for collaborative learning.

III. PROPOSED FRAMEWORK

A web-based software tool named 'Smart Learning' is designed to analyze and improve the learning of students, as shown in Figure 1. Python's Django web platform and SQL Lite3 are used to build the tool. The teachers use it to determine the progress of their students and act accordingly. The proposed framework comprises three modules including a visualization module, a module for group formation and intervention, and a prediction module.

A. DASHBOARD

Learning dashboard, a class of personal informatics, is a great tool for analyzing users' personal traits to enhance self-knowledge which in turn increases intuition, self-control, and supporting proper work [38]. Dashboards involve graphical representations of the present and prospective state of a learner or a program. Using the data-driven technologies from such dashboards, educational outcomes can be predicted

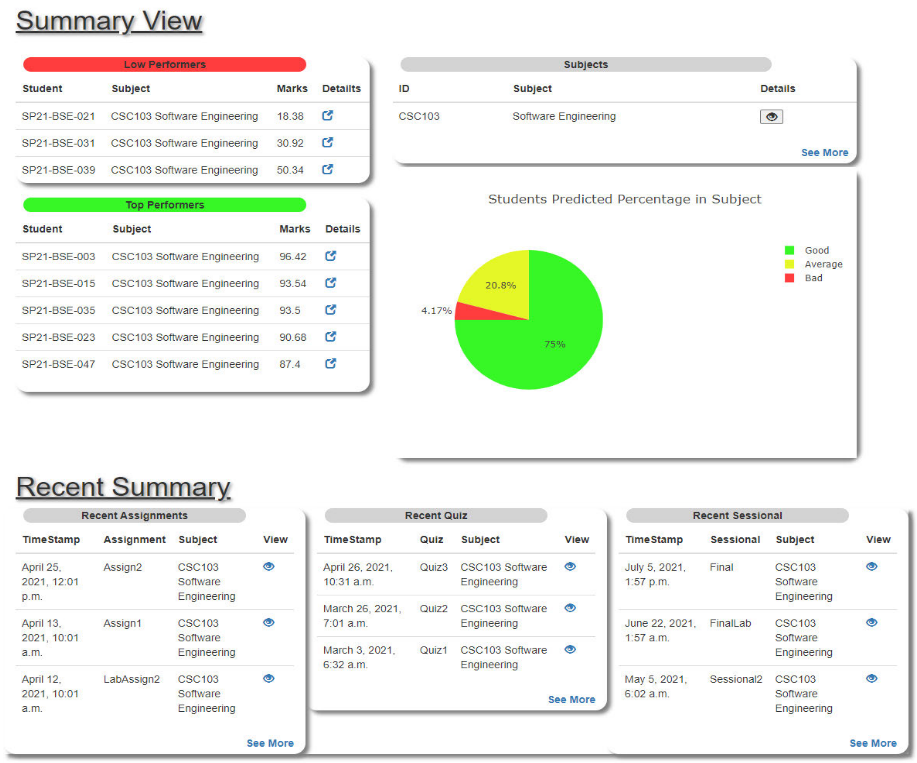


FIGURE 2. **Dashboard and prediction modules of the proposed framework.**

at an early stage. Instructors can point out and assist students at risk of academic failure [5]. In 'Smart Learning', teachers can set up a class by inserting subject, course, and student information. Teachers establish assessment tools for students such as assignments, quizzes, mid and final-term, and their evaluations. The system dashboard demonstrates the student's participation in classes, subject-wise information, top scorers, and a recent summary of assignments, quizzes, mid and final terms based on grades. A screenshot of the dashboard of the Smart Learning framework is given in Figure 2

B. PREDICTION MODULE

The proposed framework incorporates a prediction module that predicts the final outcome of a student with respect to a particular course in terms of their grades and success or failure. Several attributes related to student academic and personal information are used for the prediction module to predict students' performance. Similarly, the actual performance of a student is used such as quizzes, term grades, and assignments, etc. The prediction module contains both students' initially predicted performance and present performance in the class called actual performance. Initial performance prediction is carried out at the start of the semester where the prediction is based on previous academic and personal information such as good (71-100), average (50-70), and poor (0-49). Once the grades are predicted, teachers can make heterogeneous groups for collaborative learning. Teachers can also establish assessment tools for students such as assignments, quizzes, mid and final terms. Similarly, the performance of individual students can also be visualized as shown in Figure 3.

For predicting the performance of the students, two necessary elements are the appropriate data and classifier with higher accuracy. To obtain higher accuracy both generating the suitable data and classifier's fine-tuning is required.

1) DATASET PREPARATION

The main objective of this study is to predict student performance at the start of the semester and enhance their performance by finding factors responsible for poor performance. For this purpose, data are collected from first-year undergraduate students of the Computer Science Department of COMSATS University. The primary data are collected using a questionnaire that includes questions related to several personal, socio-economic, psychological, and academic-related variables. The questionnaire is improvised from several articles related to problem in hand [1], [25], [27], [39], [40]. After consultation and supervision from several faculty members, a questionnaire with 34 questions is finalized. A total of 164 students participated in the questionnaires with a response rate of 99% by both females and males. Table 1 presents the main features of the collected dataset.

After data gathering, data are prepared to be used for the machine learning models as it is necessary to clean and prepare the data to obtain significant results from the classifiers [6], [39]. To clean the data from errors, noise,

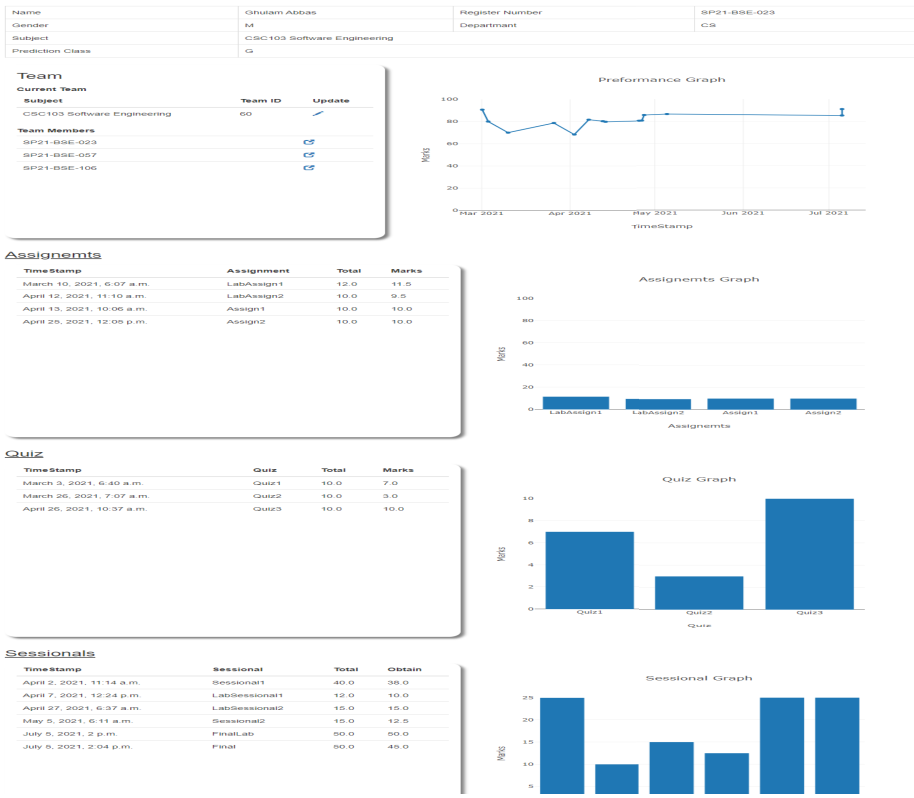


FIGURE 3. Visualization of individual student's performance.

and irrelevant information, several preprocessing steps are performed. In addition, the relevance of the attributes is studied to remove redundant, noisy, or irrelevant features. In this study, missing values are replaced with the mean value of the most frequent values for an attribute to mitigate the loss of information that might be potentially useful. Data preprocessing is applied based on attributes type. For numeric features standard scalar is used, for categorical features binary labeled features label encoder is used but for multi labeled features one hot encoder is used.

Imbalanced data problem is prevalent in many real-world datasets and it occurs where the class distributions are highly different. Studies suggest that machine learning models work best when the number of instances of each class is approximately equal [9], [13]. Analyzing the gathered dataset reveals that they are highly imbalanced with 50% samples for the 'Good' class, 40% for the 'Average' class, and 10% for the 'Poor' class. This study uses an adaptive synthetic (ADASYN) over-sampling approach to resolve the dataset imbalance problem [41]. It follows the adaptive generation of minority data samples where more synthetic data is generated for minority classes to equal the sample distribution of different classes. The ADASYN method reduces the learning bias due to imbalanced data distribution and can adaptively shift the decision boundary. Table 2 shows the distribution of the dataset after applying the ADASYN technique.

2) PREDICTION APPROACH

Figure 4 shows the flow of the approach used for predicting the students' performance. After data gathering and preparation, non-contributing features are removed such as 'Name', 'Registration number', 'Email' and 'Date of birth', etc. Also, feature importance is carried out for the most important predictable feature [12]. For training and testing, several machine learning models are used and an ensemble model is proposed as well. For this purpose, KNN, SVM, LR, RF, MLP, and DT are optimized by fine-tuning their hyperparameters.

K-NN is a non-parametric technique widely used for classification and regression problems [42]. Based on the concept of neighbors, an instance is classified using the plurality vote of neighbor's instances. For determining a neighbor different distance estimation metrics are used such as Euclidean, Minkowski, and Manhattan distance, etc.

SVM is one of the widely used models for classification and regression problems that works well for many practical applications [43]. SVM creates hyperplanes to separate the data into classes. SVM kernels are used to transform the low-dimensional input space into a high-dimensional space. By doing so, it can transform a non-separable problem into a separable problem. DT is one of the popular techniques used for prediction due to its simplicity and comprehensibility [39]. With a tree structure, DT has a root node, edge, and leaf nodes to represent the data. DT can easily

TABLE 1. Main features of student's dataset.

Name	Type	Description
Gender	Nominal	Male or Female
Family Structure	Nominal	Joint or Individual
Total number of siblings including yourself	Numeric	Total number of brother and sister student has (Should be Numeric)
Your number among siblings	Numeric	Which is the number among sibling's student has (Should be Numeric)
Parental Status	Nominal	Both or Mother Only or Father Only
Your Parents Education	Nominal	Both Educated or Mother Educated or Father Educated or Both Uneducated
Father Occupation	Nominal	Non-Services or Services or Business Man or Dead
Mother Occupation	Nominal	Housewife or Job or Business Woman or Dead
How much time you spent on Job / Business?	Numeric	Total number of hour student spent to do any job / business (Should be Numeric)
What is your Family Financial Status?	Ordinal	Strong (stable financial conditions) or Medium (sometimes stable and sometimes not stable) or Weak (not stable)
How many hours you use Mobile?	Numeric	Total number of hour student use mobile (Should be Numeric)
How many hours you have access to Internet?	Numeric	Total number of hour student has access to internet (Should be Numeric)
How many hours you play Sports?	Numeric	Total number of hour student play sports (Should be Numeric)
How much is the workload upon you, from your family?	Numeric	Family workload upon student (Should be Numeric)
Previous Degree	Nominal	Last degree student attain (F.Sc Pre-Engineering or ICS or A-Level or F.Sc Pre-Medical)
Marks Percentage in previous study	Numeric	Student's last degree marks percentage
Math marks in previous study	Numeric	Student's last degree math marks
Computer (Programming) marks in previous study	Numeric	Student's last degree computer marks
Matriculation or O-level marks percentage	Numeric	Marks obtained at secondary level
Medium of Education before University	Nominal	Mostly English or Mostly Urdu
Type of School	Nominal	Co-Education or Only Boys or Only Girls
NAT Marks	Numeric	Total NAT marks students Obtained
Your Permanent Home City	Nominal	From which city student belong
Residential Area	Nominal	Urban or Rural
Institution before Matriculation	Nominal	Government or Private
Institution during Matriculation	Nominal	Government or Private
Institution during Intermediate	Nominal	Government or Private
What is your current mode of transportation?	Nominal	By Walk or Motorcycle or Car or University Transport or Public Transport
University distance from currently living area in Km	Numeric	University distance from student currently living area
Type of Student	Nominal	Day Scholar (Living with your family) or Hostelite (Not living with your family)
Hobby	Nominal	Computer Science Related or Indoor or Outdoor or Indoor-Outdoor or None
How social are you?	Numeric	Socially confident Student (Excellent or Bad range from 1 to 5 respectively)
Why did you choose Computer Science?	Nominal	Why Student choose Computer Science as a field (describe in keyword)
Grades	Ordinal	Student's grades at end of semester in the subject are used as class label of data-set (Good, Average and Poor)

TABLE 2. Distribution of the student's dataset before and after ADASYN.

Example	Non-ADASYN case	ADASYN case
Total samples	164	233
Good	78	78
Average	76	76
Poor	10	79

be converted to classification rules. DT model has been used in many real-world applications such as financial analysis, medicine, molecular biology, manufacturing production, and astronomy. RF is a tree-based ensemble model that produces highly accurate predictions by combining many weak learners [5]. RF uses the bagging technique to train several decision trees using different bootstrap samples to enhance the performance. However, DT is challenged by the identification of appropriate split criteria that is required at the root node at each level.

In addition to the above-discussed models, this study proposes a voting classifier to enhance the prediction accuracy. The voting classifier is an ensemble model that combines different base models to make the final predictions as voting classifiers tend to show better performance than individual models [44]–[47]. This study uses SVM, RF, and K-NN as sub-estimators and final predictions are obtained using both hard and soft voting methods. The hard and soft voting models can be modeled as follows

$$KNN_p = KNN(data) \tag{1}$$
$$SVM_p = SVM(data) \tag{2}$$

and,

$$RF_p = RF(data) \tag{3}$$

where, KNN_p, SVM_p, and RF_p are the prediction by the KNN, SVM, and RF models, respectively and these predictions are used for majority voting. The class prediction with higher votes will be the final class and is decided as

$$Hard\ Voting = mode\{KNN_p,\ SVM_p,\ and\ RF_p\} \tag{4}$$

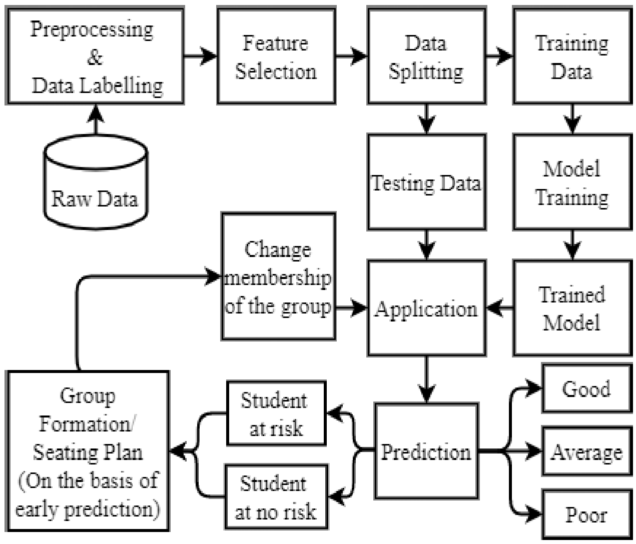


FIGURE 4. Detailed system architecture for predicting student's performance.

Contradictory to the hard voting criteria, the soft voting criterion does not consider the class with the highest prediction. Instead, it takes the probability of each class predicted from each classifier. Probabilities for soft voting criteria as found as follows

$$KNN_p(GC), KNN_p(AC), KNN_p(PC) = KNN(data) \quad (5)$$

$$SVM_p(GC), SVM_p(AC), SVM_p(PC) = SVM(data) \quad (6)$$

and

$$RF_p(GC), RF_p(AC), RF_p(PC) = RF(data) \quad (7)$$

where, GC, AC, and PC represent the good class, average class, and poor class, respectively, and $KNN_p(GC)$ is the prediction for good class by the KNN model and similarly for other models. After finding the probability by each model soft voting averages the per-class probabilities as follows:

$$GC_P = \frac{KNN_p(GC) + SVM_p(GC) + RF_p(GC)}{3} \quad (8)$$

$$AC_P = \frac{KNN_p(AC) + SVM_p(AC) + RF_p(AC)}{3} \quad (9)$$

$$PC_P = \frac{KNN_p(PC) + SVM_p(PC) + RF_p(PC)}{3} \quad (10)$$

where, $GC_P, AC_P, andPC_P$ are the probabilities for each class using each model and soft voting will make the final prediction with the highest probability as follows

$$Soft\ Voting = argmaxGC_P, AC_P, PC_P \quad (11)$$

The soft voting criterion is explained using the following example. Let the probability scores of each class given by SVC be, Good = 0.812, Average = 0.616 & Poor = 0.176 and the RF probability scores against each class be, Good = 0.828, Average = 0.594 & Poor = 0.359. Similarly, let K-NN probability scores against each class be, Good = 0.964, Average = 0.539 & Poor = 0.439 then tee soft voting can be used in the following way to predict the final class.

$$Goodclass = \frac{0.812 + 0.828 + 0.964}{3} = 0.868 \quad (12)$$

$$Averageclass = \frac{0.616 + 0.594 + 0.539}{3} = 0.583 \quad (13)$$

$$Poorclass = \frac{0.176 + 0.359 + 0.439}{3} = 0.325 \quad (14)$$

The soft voting classifies predicts that test example belongs to the 'Good' class. On the other hand, hard voting does not consider the probability of each class for individual classifiers. Instead, it takes only the predicted class from the individual classifier and predicts the final class on their voting. Let SVM, RF, and K-NN predict the output class as 'Good', 'Good', and 'Average', respectively, then as per the hard voting the final predicted class would be 'Good'.

For evaluating the performance of the prediction models, several evaluation metrics are used such as accuracy, precision, recall, and F1 score. Accuracy is the ratio between the number of correct predictions and the total number of predicted samples. For accuracy and other metrics, true positive (TP), true negative (TN), false positive (FP), and false negative (FN) samples are used.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

Precision represents the correctness of a classifier. It is the proportion of correctly classified positive samples by the total number of positive predicted samples. A high precision score means that an algorithm in the actual class is better at making observations.

$$Precision = \frac{TP}{TP + FP} \quad (16)$$

Recall shows the completeness of a classifier. It is often referred to as sensitivity, which signifies the ability of a classifier to detect a class properly. The number of true positives divided by the number of true positives plus the number of false negatives is known as recall.

$$Recall = \frac{TP}{TP + FN} \quad (17)$$

The F1 score, also called the F measure, is the harmonic mean of precision and recall. F1 score has been regarded as a better metric to determine the performance of a classifier especially when there are chances of overfitting due to data imbalance problems.

$$F_1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (18)$$

C. GROUPS FORMATION AND INTERVENTIONS

Collaborative learning is also implemented in the 'Smart Learning' framework. Collaborative learning is a scenario in which teachers divided the students into groups so that students' learning can be improved. Students can communicate with each other, share their expertise and skills

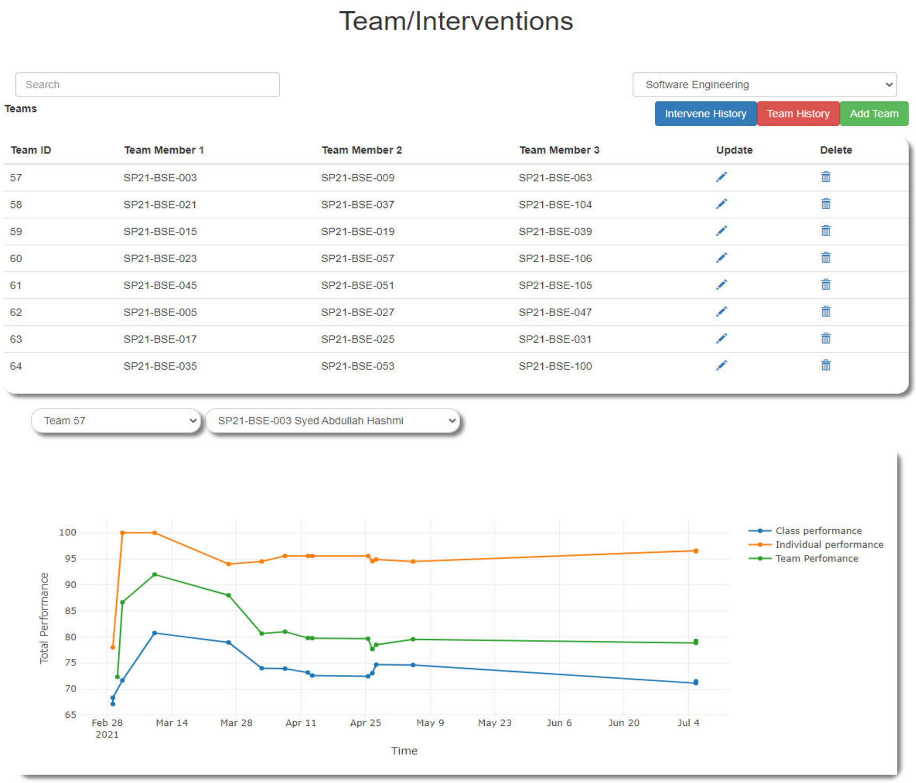


FIGURE 5. **Student's groups formation and interventions.**

and enrich the learning process. Cognitive styles and general performance indicators of the students serve as the foundation for grouping the students. On the dashboard, the instructor can create a Teams and Interventions list. The instructor prefers a seating plan that involves 3 students. The class's time to time seating arrangements are displayed and new seating plans are added. To enhance the overall team's and individual success of the students, instructors can also take part in the context of the team as per the predictive results. The instructor can also track the efficiency and individual performance of the team member. Students 'individual results are shown in both tabular form and visualization graphs that include the activities, quiz, session, and final-term marks of the student as shown in Figure 5.

IV. RESULTS AND DISCUSSIONS

Experiments are performed using a train-test split ratio of 80:20 for training and testing, respectively. All models are trained and tested on the preprocessed data and the selected evaluation metrics are used to evaluate the performance of all models. The formulated research questions are discussed here to validate the study. Results from all models are discussed in the perspective of these questions as well.

A. RQ1. WHETHER EXISTING IDENTIFIED FACTORS FOR PREDICTING STUDENTS' PERFORMANCE ARE VALID IN THE LOCAL CONTEXT (PAKISTAN)?

This question determines the validity of factors related to students' performance in the local context. This question is important as the norms, traditions, and culture of one country vary from the other and so does the education system. Therefore, it is significantly important to determine whether the existing factors in the literature are applicable and valid in the context of Pakistan. Classification algorithms have been used to predict students' performance based on demographic, pre-university, and institutional factors. Results show that the factors found in the literature are also valid for the most part in Pakistan. The data for analysis is collected from undergraduate students and contains three types of attributes to predict the performance of students. The first type is demographic that includes gender, age, residence, location, and father's qualification. The second type is institutional and has final grades, financial status, and university distance attributes. The third type is pre-qualification and includes secondary school grade, higher secondary grade, pre-college, pre-program, NAT score, and marks of intermediate and matriculation programs.

In addition to existing factors, several local factors are used with the existing factors. The input data for classifiers contains previous course marks/grade, grade point average (GPA), secondary school grade (SSG), higher secondary school grade (HSSG), gender, and parental status. Table 3 summarizes the attributes of each factor with the corresponding accuracy scores and other evaluation parameters for our study used models shown in Table 4. Results from the literature are compared with the proposed classifier, as well as, other machine learning models used in this

TABLE 3. Classification results with respect to attributes.

Reference	Attributes	Classifiers' accuracy (%)											
		LR	KNN	RF	SVM	NB	DT	MLP	Ense.	SMO	J48	REP	ANN
Current work	Student demographic, Pre-qualification, Institutional, pre-admission test score	76.60	82.01	80.85	80.85	76.60	72.34	78.72	82.98	-	-	-	-
[28] (2020)	Student demographic, Social interactions and Pre-admission test scores (HSGA, SAAT and GAT scores)	-	-	-	75.28	73.61	75.91	-	-	-	-	-	79.22
[29] (2021)	Interaction logs with MOOC enabled course and assessment grades	84	-	-	85	-	-	-	-	-	-	-	85
[30] (2021)	Skew, standard deviation, variance, kurtosis, overall trajectory, and final trajectory	-	-	87.5	-	-	-	-	-	-	-	-	-
[22] (2021)	User's interaction (clicks) with the VLE	-	-	75.5	-	-	70.5	78.2	-	-	-	-	-
[32] (2021)	Student demographic, academic information	70.8	69.2	-	73.8	72.0	69.2	-	75.9	-	-	-	71.0
[40] (2013)	Student demographic, High school background	-	-	-	-	49.5	-	72.38	-	57.25	64.88	60.13	-
[49] (2020)	Student Demographic, Social interaction, Internal assessment and Admission test scores	70.8	69.2	-	73.8	72	69.2	-	75.9	-	-	-	71

TABLE 4. Precision, recall and F1 scores for all the models.

Algorithm		Precision (%)	Recall (%)	F1 score (%)
LR		72	71	72
RF		77	77	77
DT		68	67	67
KNN		80	79	79
SVM		77	77	77
MLP		76	75	75
NB		72	72	72
Ensemble Model	Soft Voting	80	79	79
	Hard Voting	80	79	79

study. The findings show that the identified factors used for grade prediction are comparable with the local factors. Experimental results indicate that 'previous grade', 'internal marks' and HSSC grades are found to be effective for grade prediction.

B. RQ2. WHICH ARE THE MOST IMPORTANT FEATURES THAT HELP ACCURATE PREDICTION OF STUDENT'S PERFORMANCE?

The second question determines important features to make accurate predictions of students' performance. For this purpose, a deep understanding of the factors, as well as, features is required. To this end, study [41] conducted a survey of 357 papers related to students' performance with respect to 29 features. These features fall under the banners of course and pre-course performance, student engagement, student demographics, high school performance, and psycho-motor skills. Findings show that the degree dropout rate is mainly influenced by student motivation, habits, social and financial issues, lack of progress, pre-qualification, institutional, admission test score, and career transition.

This study makes deeper and explorative analysis s to examine and determine the key factors responsible for students' performance. The objective is to understand the factors that might be used in obtaining the key strengths and weaknesses for obtaining students' better performance.

Figure 6 shows the importance of all the features of the data set. It can be observed that features such as residential area, NAT marks, time spent for sports and mobile usage, previous degree grades and father occupation, etc. are more important than gender, parental and job status, etc. This solidifies our observation that most students' grades are affected by student involvement in the class. These factors are important because results show that those students perform significantly better who have good marks in admission tests and previous degrees than those who do not participate in such activities.

C. RQ3. IS IT POSSIBLE TO IMPROVE THE PERFORMANCE OF A STUDENT BY COLLABORATIVE LEARNING?

The third question of this research evaluates the efficacy of collaborative learning to boost the performance of students. This research uses the concept of the control group and experiment group for this purpose. The study uses the impact of collaborative learning on students' academic achievement over specified periods of time and the variance across experimental and control groups is studied. After exposing the experimental group to collaborative learning while providing traditional study to the controlled group the post-test is administered. The pre-test and post-test are used to analyze the academic achievement of the control group and experimental group, respectively.

For handling the potential pre-existing differences concerning students' ability in the experimental and control groups, a pre-test is conducted. Later, these scores are used to divide the students into groups. This study uses teacher-selected groups for collaborative learning. For a fair comparison, the effect of teacher quality on the groups is controlled by using the same instructor and the same course contents. The activities for the experiment and control groups are however different where the former followed learning activities in small heterogeneous groups while the latter follows the traditional teaching method. The contents are

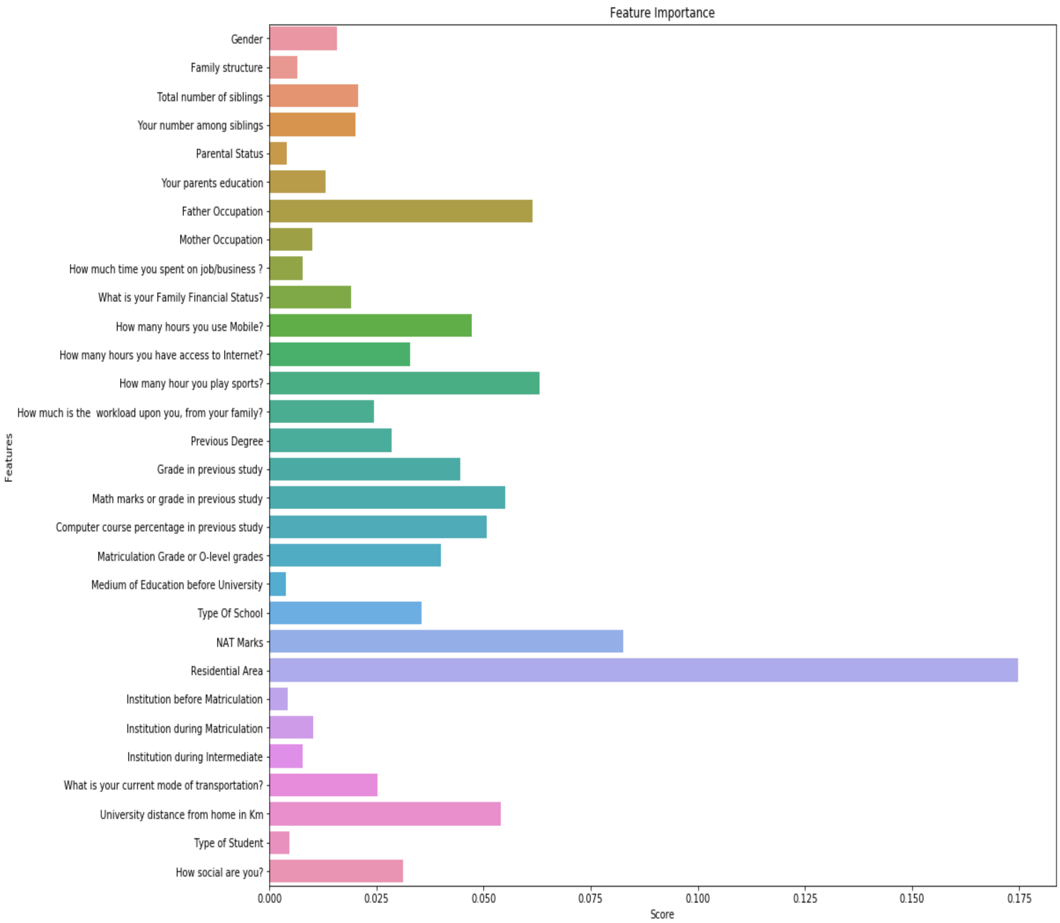


FIGURE 6. Feature importance ranked for the student learning outcomes.

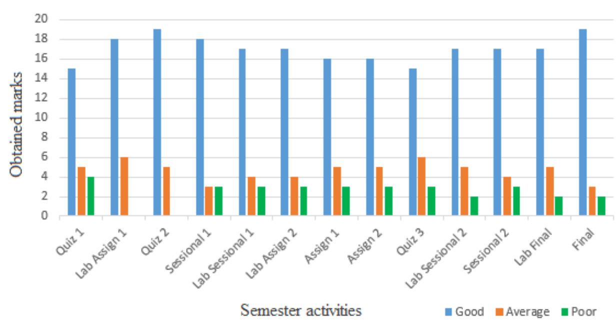


FIGURE 7. Detailed semester assessment.

taught to the experimental group through multiple activities tests, as shown in Figure 7

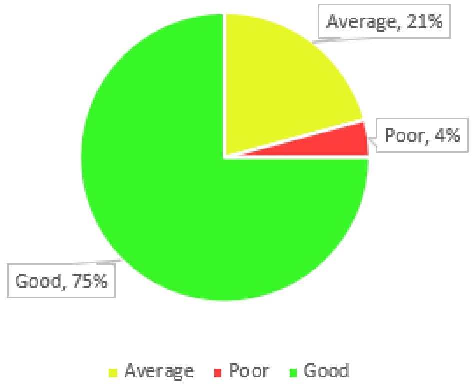
First, we analyze the performance of the experimental group before and after the treatment. For experiments, the experiment group has a total of 24 students. Table 5 shows the percentage score of performance prediction by the 'Smart Learning' system.

We have 75% students for the 'Good' class, 21% for the 'Average' class, and 4% for the 'Poor' class of students before the treatment for the experiment group. Results show that the ratio of good performance students is the highest in the experiment group even before the treatment.

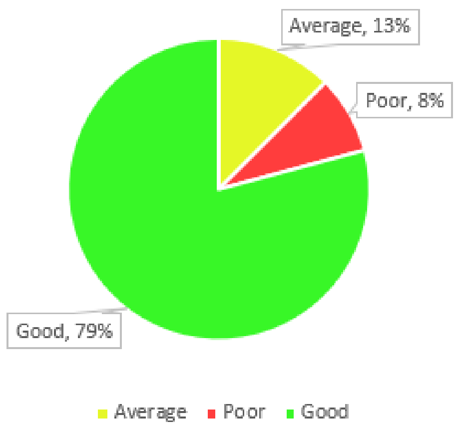
TABLE 5. Results of achievement scores of experimental groups before and after treatment.

Experimental Group	Percentage Score		
	Good	Average	Poor
Pre Assessment	75 %	21 %	4 %
Post Assessment	79 %	13 %	8 %

Now, the teacher divides the experimental group students into 8 sub-groups, each with 3 students. A large number of activities are performed such as assignments, quizzes, sessions, and projects performed by teachers in groups. Teachers observe the performance of both individuals and groups from time to time. Appropriate and necessary interventions are made by the teachers accordingly. After the term ends, the performance of the subgroups in the experiment group is analyzed again to see the change in students' performance. Results of students' performance for pre-assessment and post-assessment are given in Figure 8. Results show that the performance of the students in the experiment group has been improved after the treatment. Currently, 79 % of the total students fall under the 'Good' class which was 75% before the treatment. Similarly, the



(a)



(b)

FIGURE 8. **Achievement scores of experimental groups, (a) Pre assessment, and (b) Post assessment.**

percentage of the students in the 'Average' class has been reduced to 13% from 21%. However, the ratio of 'Poor' class students has been slightly increased but this ratio is not substantial as compared to the ratio of the students who improved their performance.

After receiving the treatment, the achievement scores of both groups have a significant difference. Results show that the experimental group after treatment is performing better than before the treatment. It is so because the students took part in each class activity, prepared and submitted their tasks and so perform much better than control groups. Overall, we can say that the collaborative learning method has a significant effect on student success, and thus it can enhance the learning ability of students.

V. CONCLUSION AND FUTURE WORK

The relevance and importance of big data analytics and collaborative learning have been proven by several research works, however, its validity and efficacy vary from one locality to another due to demographic, institutional, and personal attributes. This study is designed with twofold objectives; building a user-friendly and efficient system to predict students' performance and adopting proper corrective measures to elevate their learning, and investigating the potential and scope of collaborative learning in the context of Pakistan. An extensive literature review is performed to find the appropriate factors for predicting students' performance. A system, called Smart Learning, is built where the instructor can predict, and track the students' performance, utilize collaborative learning and analyze the results. For prediction, besides the well-known machine learning models, an ensemble model is proposed which shows better performance than state-of-the-art approaches. Three research questions are formulated to meet the objective of this study. For the first research question, results show that the existing identified factors for students' performance prediction are valid to a large extent in the local context of Pakistan as well. With respect to the most important features (2nd research question), experiments reveal that demographic, pre-university, and institutional factors (NTS marks, and previous degrees marks, etc.) are the most appropriate factors to predict students' final performance. For research question 3, experiments are performed with experiment and control groups to analyze the impact of collaborative learning on improving students' performance. With a teacher-selected group strategy, students in the experiment group are given several tasks, their performance is monitored over time, and students are regrouped if needed. Results suggest that by using collaborative learning, the student's learning capability can be elevated substantially. This study confirms that classroom-based collaborative learning techniques are found to be more effective than traditional learning methods. The study is not without limitations. The dataset collected for experiments is relatively small, and a large dataset can provide more conclusive results. We intend to predict teacher performance as well and build a course recommender system for students in the future.

REFERENCES

[1] A. Daud, N. R. Aljohani, R. A. Abbasi, M. D. Lytras, F. Abbas, and J. S. Alowibdi, "Predicting student performance using advanced learning analytics," in *Proc. 26th Int. Conf. World Wide Web Companion (WWW)*, 2017, pp. 415–421.
[2] N. Patwa, A. Seetharaman, K. Sreekumar, and S. Phani, "Learning analytics: Enhancing the quality of higher education," *Res. J. Econ.*, vol. 2, no. 2, pp. 13–29, 2018.
[3] B. Dietz-Uhler and J. E. Hurn, "Using learning analytics to predict (and improve) student success: A faculty perspective," *J. Interact. Online Learn.*, vol. 12, no. 1, pp. 17–26, 2013.
[4] O. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, "The current landscape of learning analytics in higher education," *Comput. Hum. Behav.*, vol. 89, pp. 98–110, Dec. 2018.
[5] G. Akçapınar, A. Altun, and P. Aşkar, "Using learning analytics to develop early-warning system for at-risk students," *Int. J. Educ. Technol. Higher Educ.*, vol. 16, no. 1, p. 40, Dec. 2019.
[6] S. N. Liao, D. Zingaro, K. Thai, C. Alvarado, W. G. Griswold, and L. Porter, "A robust machine learning technique to predict low-performing students," *ACM Trans. Comput. Educ.*, vol. 19, no. 3, pp. 1–19, Jun. 2019.
[7] O. H. T. Lu, A. Y. Q. Huang, J. C. H. Huang, A. J. Q. Lin, H. Ogata, and S. J. H. Yang, "Applying learning analytics for the early prediction of students' academic performance in blended learning," *J. Educ. Technol. Soc.*, vol. 21, no. 2, pp. 220–232, 2018.
[8] K. Casey and D. Azcona, "Utilizing student activity patterns to predict performance," *Int. J. Educ. Technol. Higher Educ.*, vol. 14, no. 1, p. 4, Dec. 2017.

[9] D. Azcona, I.-H. Hsiao, and A. F. Smeaton, ''Detecting students-at-risk in computer programming classes with learning analytics from students' digital footprints,'' *User Model. User-Adapted Interact.*, vol. 29, no. 4, pp. 759–788, Sep. 2019.

[10] W. Chango, M. Sánchez-Santillán, R. Cerezo, and C. Romero, ''Predicting students' performance using emotion detection from face-recording video when interacting with an its,'' in *IEDM Tech. Dig.*, 2020, pp. 1–3.

[11] A. Shahzad, M. Valcke, and R. Bahoo, ''A study to analyze the teacher's perceptions about the adoption of collaborative learning in post-graduate classes of IUB,'' *Proc.-Social Behav. Sci.*, vol. 46, pp. 3056–3059, Jan. 2012.

[12] S. Khatoon and M. Akhter, ''An innovative collaborative group learning strategy for improving learning achievement of slow learners,'' *J. Res. Reflections Educ.*, vol. 4, no. 2, pp. 1–19, 2010.

[13] F. Gull and S. Shehzad, ''Effects of cooperative learning on students' academic achievement,'' *J. Educ. Learn.*, vol. 9, no. 3, pp. 246–255, 2015.

[14] S.-L. Wang and S. S. J. Lin, ''The effects of group composition of self-efficacy and collective efficacy on computer-supported collaborative learning,'' *Comput. Hum. Behav.*, vol. 23, no. 5, pp. 2256–2268, Sep. 2007.

[15] B. Akram and L. Ghazanfar, ''Self efficacy and academic performance of the students of Gujrat University, Pakistan,'' *Academic Res. Int.*, vol. 5, no. 1, p. 283, 2014.

[16] C. Malliarakis, M. Satratzemi, and S. Xinogalos, ''Integrating learning analytics in an educational MMORPG for computer programming,'' in *Proc. IEEE 14th Int. Conf. Adv. Learn. Technol.*, Jul. 2014, pp. 233–237.

[17] N. Thota and A. Berglund, ''Learning computer science: Dimensions of variation within what Chinese students learn,'' *ACM Trans. Comput. Educ.*, vol. 16, no. 3, pp. 1–27, Jun. 2016.

[18] J. Lagus, K. Longi, A. Klami, and A. Hellas, ''Transfer-learning methods in programming course outcome prediction,'' *ACM Trans. Comput. Educ.*, vol. 18, no. 4, pp. 1–18, Nov. 2018.

[19] D. Azcona, I.-H. Hsiao, and A. F. Smeaton, ''Predictcs: Personalizing programming learning by leveraging learning analytics,'' in *Proc. Int. Conf. Learn. Anal. Knowl. (LAK)*. Sydney, NSW, Australia: Univ. Sydney, 2018.

[20] D. Baneres, M. E. Rodriguez, and M. Serra, ''An early feedback prediction system for learners at-risk within a first-year higher education course,'' *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 249–263, Apr. 2019.

[21] J. L. Harvey and S. A. Kumar, ''A practical model for educators to predict student performance in K-12 education using machine learning,'' in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2019, pp. 3004–3011.

[22] A. Rivas, A. González-Briones, G. Hernández, J. Prieto, and P. Chamoso, ''Artificial neural network analysis of the academic performance of students in virtual learning environments,'' *Neurocomputing*, vol. 423, pp. 713–720, Jan. 2021.

[23] R. Ghorbani and R. Ghousi, ''Comparing different resampling methods in predicting students' performance using machine learning techniques,'' *IEEE Access*, vol. 8, pp. 67899–67911, 2020.

[24] B. L. Lanie, ''Affinity propagation SMOTE approach for imbalanced dataset used in predicting student at risk of low performance,'' *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 5066–5070, Aug. 2020.

[25] A. Almasri, E. Celebi, and R. S. Alkhawaldeh, ''EMT: Ensemble meta-based tree model for predicting student performance,'' *Sci. Program.*, vol. 2019, pp. 1–13, Feb. 2019.

[26] O. W. Adejo and T. Connolly, ''Predicting student academic performance using multi-model heterogeneous ensemble approach,'' *J. Appl. Res. Higher Educ.*, vol. 10, no. 1, pp. 61–75, Feb. 2018.

[27] 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.

[28] 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.

[29] A. A. Mubarak, H. Cao, and S. A. M. Ahmed, ''Predictive learning analytics using deep learning model in MOOCs' courses videos,'' *Educ. Inf. Technol.*, vol. 26, no. 1, pp. 371–392, Jan. 2021.

[30] S. Dass, K. Gary, and J. Cunningham, ''Predicting student dropout in self-paced MOOC course using random forest model,'' *Information*, vol. 12, no. 11, p. 476, Nov. 2021.

[31] H. Zeineddine, U. Braendle, and A. Farah, ''Enhancing prediction of student success: Automated machine learning approach,'' *Comput. Electr. Eng.*, vol. 89, Jan. 2021, Art. no. 106903.

[32] H. Le, J. Janssen, and T. Wubbels, ''Collaborative learning practices: Teacher and student perceived obstacles to effective student collaboration,'' *Cambridge J. Educ.*, vol. 48, no. 1, pp. 103–122, Jan. 2018.

[33] Á. Fidalgo-Blanco, M. L. Sein-Echaluce, F. J. García-Peñalvo, and M. Á. Conde, ''Using learning analytics to improve teamwork assessment,'' *Comput. Hum. Behav.*, vol. 47, pp. 149–156, Jun. 2015.

[34] O. H. T. Lu, J. C. H. Huang, A. Y. Q. Huang, and S. J. H. Yang, ''Applying learning analytics for improving students engagement and learning outcomes in an MOOCs enabled collaborative programming course,'' *Interact. Learn. Environ.*, vol. 25, no. 2, pp. 220–234, Feb. 2017.

[35] R. Vivian, K. Falkner, N. Falkner, and H. Tarmazdi, ''A method to analyze computer science students' teamwork in online collaborative learning environments,'' *ACM Trans. Comput. Educ.*, vol. 16, no. 2, pp. 1–28, Mar. 2016.

[36] G. Babu, P. Suresh, and K. Pariventhan, ''Enhancement of learning through collaborative learning techniques,'' *New Man J. Multidisciplinary Stud.*, vol. 4, no. 1, pp. 29–38, 2017.

[37] M. Li and J. Campbell, ''Asian students' perceptions of group work and group assignments in a New Zealand tertiary institution,'' *Intercultural Educ.*, vol. 19, no. 3, pp. 203–216, Jun. 2008.

[38] K. Verbert, E. Duval, J. Klerkx, S. Govaerts, and J. L. Santos, ''Learning analytics dashboard applications,'' *Amer. Behav. Scientist*, vol. 57, no. 10, pp. 1500–1509, Oct. 2013.

[39] V. Ramesh, P. Parkavi, and K. Ramar, ''Predicting student performance: A statistical and data mining approach,'' *Int. J. Comput. Appl.*, vol. 63, no. 8, pp. 35–39, Feb. 2013.

[40] A. Y. Q. Huang, O. H. T. Lu, J. C. H. Huang, C. J. Yin, and S. J. H. Yang, ''Predicting students' academic performance by using educational big data and learning analytics: Evaluation of classification methods and learning logs,'' *Interact. Learn. Environ.*, vol. 28, no. 2, pp. 206–230, Feb. 2020.

[41] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S. N. Liao, ''Predicting academic performance: A systematic literature review,'' in *Proc. Companion 23rd Annu. ACM Conf. Innov. Technol. Comput. Sci. Educ.*, 2018, pp. 175–199.

[42] B. K. Yousafzai, M. Hayat, and S. Afzal, ''Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student,'' *Educ. Inf. Technol.*, vol. 25, no. 6, pp. 4677–4697, Nov. 2020.

[43] A. L. Dyckhoff, D. Zielke, M. Bültmann, M. A. Chatti, and U. Schroeder, ''Design and implementation of a learning analytics toolkit for teachers,'' *J. Educ. Technol. Soc.*, vol. 15, no. 3, pp. 58–76, 2012.

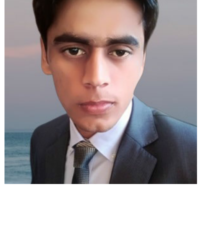
[44] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. Choi, ''Tweets classification on the base of sentiments for U.S. Airline companies,'' *Entropy*, vol. 21, no. 11, p. 1078, Nov. 2019.

[45] A. Mehmood, B.-W. On, I. Lee, I. Ashraf, and G. Sang Choi, ''Spam comments prediction using stacking with ensemble learning,'' *J. Phys., Conf. Ser.*, vol. 933, Jan. 2018, Art. no. 012012.

[46] M. Khalid, I. Ashraf, A. Mehmood, S. Ullah, M. Ahmad, and G. S. Choi, ''GBSVM: Sentiment classification from unstructured reviews using ensemble classifier,'' *Appl. Sci.*, vol. 10, no. 8, p. 2788, Apr. 2020.

[47] M. Umer, I. Ashraf, A. Mehmood, S. Ullah, and G. S. Choi, ''Predicting numeric ratings for Google apps using text features and ensemble learning,'' *ETRI J.*, vol. 43, no. 1, pp. 95–108, Feb. 2021.

[48] H. Zeineddine, U. Braendle, and A. Farah, ''Enhancing prediction of student success: Automated machine learning approach,'' *Comput. Electr. Eng.*, vol. 89, Jan. 2021, Art. no. 106903.

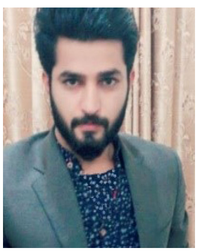


ADNAN RAFIQUE received the M.C.S. degree from the Department of Computer Science and IT, The Islamia University of Bahawalpur, Pakistan, in 2017. He is currently pursuing the master's degree in computer science with the Department of Computer Science, COMSATS University Islamabad (CUI), Lahore, Pakistan. His recent research interests include data mining, mainly working on machine learning and deep learning-based text mining, natural language processing, interactive learning, and educational technologies.

MUHAMMAD SALMAN KHAN received the M.Sc. degree in computer science from the Punjab University College of IT, Lahore, Pakistan, in 2004, and the Ph.D. degree from the Graz University of Technology, Austria, in 2011. Since 2011, he has been an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad, Lahore Campus. His research interests include interactive learning, e-learning, and NLP.

MUHAMMAD HASAN JAMAL received the B.S. degree in computer and information engineering from International Islamic University Malaysia, in 2005, the M.S. degree in computer engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2008, and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 2015. From 2006 to 2010, he was a Research Associate and a Senior Research Associate with the Al-Khawarizmi Institute of Computer Science, UET, Lahore. He was a Graduate Technical Summer Intern at Sandia National Laboratories, Albuquerque, NM, USA, in Summer 2015. Since 2016, he has been an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad, Lahore Campus. His research interests include parallel and distributed systems, scientific computing, data analytics, and NLP. He was a recipient of the Fulbright Scholarship for his Ph.D. studies.

MAMOONA TASADDUQ received the master's degree from the Department of Computer Science, COMSATS University Islamabad, Lahore Campus, in 2018. She is currently a Lecturer with the Department of Computer Science, COMSATS University Islamabad. Her research interests include educational technology, information sciences, natural language processing (NLP), and learning analytics.



FURQAN RUSTAM received the master's degree in computer science from The Islamia University of Bahawalpur, in 2017, and the M.S. degree from the Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Pakistan, in 2020. He is currently a Research Assistant and a Lecturer at the Fareed Computing and Research Center, KFUEIT. His recent research interests include data mining, mainly working on machine learning and deep learning-based IoT, text mining, and bioinformatics tasks.

ERNESTO LEE is currently working as a Professor at the Department of Computer Science, Broward College, Broward County, FL, USA. His research interests include block-chain, the IoT, data mining, mainly working machine learning and deep learning-based IoT, and text mining tasks.

PATRICK BERNARD WASHINGTON is currently an Associate Professor of finance at the Morehouse College. He is also a philosopher leading the world conversation at the intersection of finance, technology, and public policy. He is a distinguished author and public speaker providing innovative solutions, investing in transformative ideas, and changing people's financial trajectory. As a leading authority on data science in finance, he has integrated deep and machine learning into his research while assisting companies in their implementation of artificial intelligence. His research interests include the intersections between investments, asset pricing, technology, real estate, and corporate governance.

IMRAN ASHRAF received the M.S. degree in computer science from the Blekinge Institute of Technology, Karlskrona, Sweden, in 2010, and the Ph.D. degree in information and communication engineering from Yeungnam University, Gyeongsan, South Korea, in 2018. He has worked as a Postdoctoral Fellow at Yeungnam University, where he is currently working as an Assistant Professor at the Department of Information and Communication Engineering. His research interests include indoor positioning and localization, indoor location-based services in wireless communication, smart sensors (LIDAR) for smart cars, and data mining.