IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 September 2022, accepted 25 September 2022, date of publication 30 September 2022, date of current version 14 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3211070

RESEARCH ARTICLE

An Explainable Model for Identifying At-Risk Student at Higher Education

SARAH ALWARTHAN, NIDA ASLAM, AND IRFAN ULLAH KHAN
Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia
Corresponding author: Sarah Alwarthan (saalwarthan@iau.edu.sa)

ABSTRACT Nowadays, researchers from various fields have shown great interest in improving the quality of learning in educational institutes in order to improve student achievement and learning outcomes. The main objective of this study was to predict the at-risk student of failing the preparatory year at an early stage. This study applies several educational data mining algorithms including RF, ANN, and SVM to build three classification models to meet the objectives of this study. Moreover, different features selection methods namely RFE, and GA have been examined to find the best subset of the highest influential features. Furthermore, several sampling approaches are applied to balance the dataset used in this study, including SMOTE, and SMOTE-Tomek Link. Three datasets related to the preparatory year student from the humanities track at IAU were used in this study. The collected datasets are imbalanced datasets, SMOTE-Tomek Link technique has been used to balance the three proposed datasets. The results showed that RF outperformed other techniques as it records the highest performance for building the models. Moreover, RFE with Mutual Information finds the best subset of features to build the first model. Finally, this study not only developed several classification models to identify at-risk students, but it also went a step further by employing XAI techniques such as LIME, SHAP, and the global surrogate model to explain the proposed prediction models, explaining the output and highlighting the reasons for the student failure.

INDEX TERMS Educational data mining, identifying at-risk student, LIME, preparatory year, random forest, SHAP.

I. INTRODUCTION

In recent years, there has been a growing interest in using data mining to investigate several research questions addressed by educational research. Applying data mining techniques on educational data is known as Educational Data Mining (EDM). The most common problems in the EDM field are how to predict correctly whether the student is likely to complete the course or program and identify the student's performance level at an early stage in a specific course or academic year. EDM aimed to leverage the stored educational data to improve the institutes' learning environment and detect valuable information from the massive amount of educational stored data [1]. The educational data is not limited to academic grades (course grade and GPA), but it includes various data such as data collected from online platforms (Learning Management System (LMS)), demographic data (age, nationality, gender), as well as admission data (entry test and high school grade) [2].

The term "at-risk student" is commonly used in the field of education to describe a student who has a high risk of academic failure and frequently requires the support and intervention of instructors to achieve academic success. Increasing the number of at-risk students is a serious concern and identifying students who may be at risk of failing a course at an early stage is of interest to many educational researchers (instructors and institutions). Increasing the number of at-risk students, course failure and student attrition rates are major factors influencing the universities' ranking. Therefore, many universities have applied EDM techniques and take the advantage of stored data to predict students' performance at an early stage and provide the necessary support to at-risk students [3].

Nowadays, predicting student achievement has been one of the major interesting research subjects, due to its great impact on improving the students' academic level by applying

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

different educational data mining techniques to provide the necessary support for the struggling students. The first academic year is a crucial year for undergraduate students so, the lack of necessary support during this year may lead to student frustration and cause dropout and failure. It has been found that students face some difficulties in universities that apply the preparatory year system. Where students are under pressure to obtain a high Cumulative Grade Point Average (CGPA) to fulfil the admission requirements to join one of the academic programs. In addition, the number of failed students is increased in some preparatory year tracks, especially the humanities track. However, existing studies in higher education at Saudi universities are mainly focused on predicting the student academic achievement at the graduation level, and they are very limited to a computer science major. This study used educational data mining techniques to propose several models that assist and support undergraduates in their first academic year. However, identifying at-risk students at an early stage could help instructors and academic advisors to monitor student achievement and progress.

This study used three datasets related to the preparatory year student from the humanities track at Imam Abdulrahman bin Faisal University (IAU). There is a vast variety of classification algorithms in the literature but choosing the optimal one is challenging as they differ in many ways, such as learning rate and the dataset used in training [4]. Three frequently used algorithms, Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) have been evaluated in this study to build the three classification models. The main objective of this study is to classify the CGPA level of the preparatory year student. So that the student at risk of failing the preparatory can be identified. Moreover, this study enhances the explanation of the prediction model by applying XAI techniques. For the successful completion of the preparatory year program at IAU, students must have to pass all the courses and achieved a CGPA ≥ 3 [5]. To accomplish the aim of this study, three classification models were constructed.

In order to identify the at-risk student of failing the preparatory year at an early stage, the first model was constructed to find the most influence features (courses) that identify the at-risk student where the model classifies the student into two categories At-risk and Not At-risk. The second and third prediction models were used to classify whether or not the student will fail the course, as it has been found from the first model that the most influential courses for identifying the at-risk student are ENGL 103 and ENGL 114. Finally, the Local Interpretable Model-agnostic Explanations method (LIME), SHapley Additive exPlanations (SHAP), and global surrogate model were used to explain the predictions of the complex black box classification models used for educational decision-making.

A. CONTRIBUTION

This study contributes to the literature in different ways. First, three real-world datasets for the preparatory year students at IAU were analyzed. Second, several data mining classification algorithms including RF, ANN, and SVM in addition to different pre-processing techniques were evaluated for classifying the CGPA and identifying the at-risk student. Third, three classification models using data mining techniques were developed to predict at-risk students at an early stage. However, the performance of the proposed classification models exceeds the prediction models in the literature that had similar targeted objectives. Fourth, this study differs from previous studies in the way that it used datasets collected about the humanities track students from the deanship of preparatory year and supporting studies at IAU, located in KSA. Also, this study is not only built several classification models to identify at-risk students, but also it went deeper than that, where this work applied explainable artificial intelligence techniques including LIME, SHAP, and global surrogate model to explain the proposed prediction models (black box models), to explain the output and highlight the reasons behind the failure. Finally, classifying student CGPA, finding the most influential features on the CGPA, and identifying at-risk students at an early stage will contribute to solve the study problem. As a result of using the proposed models, instructors and academic advisors will help students to increase their GPA and advise them to put more effort into the most influenced courses on the CGPA as well as help them to pass the preparatory year.

Moreover, the universities aim to graduate highly qualified students that meet the needs of the labor market. The findings from this study will help decision-makers to provide deserving students with additional guidance. Moreover, this model will help preparatory year administration to find the reasons behind the failure.

This study is organized into seven sections. The first section is the introduction where it highlights the addressed problem, the main objectives, and the contribution of this study. Section 2 discusses the previous studies that predict student academic performance. In section 3, the adopted methodology and common techniques used in the main stages of this study are explained. Section 4 explains the empirical study that covers the experimental setup and the description of the datasets. Section 5 discusses the results of all experiments conducted to achieve the aim of this study and presents the explanation of the black box prediction model by using different XAI methods. Section 6 discusses the findings. Finally, the conclusion of this study and the future work are presented in section 7.

II. RELATED WORK

When Many studies related to educational data mining have been conducted during the last several years. This section focused on reviewing published studies that aim to predict student performance in higher education using EDM techniques.

Authors in [6], [7], [3], and [8] applied several educational data mining techniques to predict student achievement on the course level. A study carried out by [6] identified students at risk of failing the course at an early stage. For the earliest prediction, the student performance during the course was divided into six stages: 0%, 20%, 40%, 60, 80%, and 100% of the course length. In this study, the authors used a public dataset named Open University Learning Analytics Dataset (OULAD) that contains 32,593 student records and 31 attributes, including student demographics, course assessments' scores, and student online interaction. They applied six machine learning techniques, including K-Nearest Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM), ExtraTree (ET), AdaBoost, Gradient boosting classifier, and Artificial Neural Network (ANN) to build the prediction model at various stages of the course length. The proposed model classifies the student into one of the four classes: withdrawn, fail, pass, and distinction. The feature engineering technique was used to enhance the performance of the prediction model. The multi-classes problem was converted into a binary classes problem by grouping the withdrawn and fail labels to form the fail class label and combining the pass and distinction labels to create the pass class label. Finally, they found that RF scored the highest performance compared with other classifiers, where it recorded 79% for precision, recall, f-score, and accuracy at 20% of the course duration. Besides, RF improved the model's performance to 88% precision, recall, f-score, and accuracy at 60% of the course duration. At 100% of the course duration, the RF scored 92% precision and 91% for recall, f-score, and accuracy.

Similarly, another study [7] applied the Bagging ensemble method and eight machine learning techniques: KNN, RF, SVM, Logistic Regression (LR), Naïve bayes (NB), and three different topologies of ANN to identify at-risk students who may fail the course. Two datasets for two courses were utilized in this study to classify the students into one of the three classes called Good, Fair, and Weak. The first dataset was collected from the e-learning platform about the assessments' marks for a group of 52 engineering students in one course, while the second dataset contains different tasks grades, including assignments, quizzes, and exams for 486 science students. The results showed that the Bagging ensemble model outperformed other techniques where the first dataset's Bagging model achieved an accuracy of 66.7% when considering 20% or 50% of the coursework, while the second dataset's bagging model scored 88.2% and 93.1% accuracy when considering 20% and 50% of the coursework, respectively.

The study [3] identified at-risk students of failing the final exam of introductory programming course at an early stage using the first two weeks of formative assessment tasks (exercise and homework) grades. The predictive model was deployed using the RF ensemble technique to classify the students whether they are at risk of failing the final exam or not. The proposed RF model was trained and tested using the dataset, including two predictors for 289 students registered in a programming course. The proposed classification model overfitted where it achieved 72.73% accuracy for training and 59.64 % accuracy for testing.

Moreover, the authors in [8] applied ANN, Decision Tree (DT) (J48), SVM, and NB to predict the student that may fail in the programming courses. Two different datasets have been analyzed in this work. The first dataset contains 161 student records from traditional learning (face to face), and the second one includes 262 student records from online distance learning. As a result of preparing the datasets and searching for optimal parameters (fine-tuning), SVM got the best performance with an F1-score of 92% and 83% for distance and on-campus datasets. SVM was able to predict the student performance when the student performed at least half of the distance education course duration, unlike on-campus learning that required a student to complete at least a quarter of the course duration.

Several studies were made to predict student achievement at the graduation time [9], [10], [11], [12]. Authors in [9] conducted a study on 339 computer college students at Imam Abdulrahman Bin Faisal University to classify the student CGPA at graduation based on the preparatory year achievement. The proposed model used classification algorithms based on DT, including J48, Random Tree, and REPTree, to classify students into three classes: High, Average, and Below Average. The final prediction model achieved 69.3% accuracy when using the optimal parameter's value for J48 and reducing the number of features to 4 out of 14 attributes. Authors found that the CGPA of the first year, an introductory math course, a computer skills course, and a communication skills course are the most influential factors from the first-year courses on the graduation CGPA.

Moreover, 15 classification techniques were applied in this study [10] to predict the final CGPA for a computer college student. Authors found seven classifiers, namely NB, Hoeffding Tree, SMO, RF, LMT, Simple Logistic, and KNN, achieved accuracy higher than the average accuracy of the 15 classifiers. The proposed classification model was built using a computer college dataset containing 530 records and 64 features, including the final CGPA class. The highest accuracy, 91%, was scored by NB and Hoeffding Tree. The authors specified some courses that significantly impacted the CGPA: Operating Systems, Statistics, General Physics, Computer Programming, and Algorithms.

A study that analyzed 1841 engineering students' data in [11] examined the impact of the first three years GPA on the final CGPA. Several classification techniques, such as ANN, RF, DT, NB, Tree Ensemble, and LR, were used to classify the final CGPA with high accuracy of 89.15% using LR. Moreover, the authors found that the highest influenced feature is the third year's GPA, followed by the second then first year.

Similarly, another study [12] compared several classification techniques to examine whether the pre-admission requirement and the personal information affect the final GPA or not. Two DT (C4.5 and ID3), NB, and KNN classification methods were used to predict the final GPA level. The experiment was applied on 2281 undergraduate students, and the result showed that NB is an efficient algorithm and obtained the highest accuracy of 43.18%. Also, it was found that the pre-admission requirement and the personal student information have the highest impact on the graduation GPA.

Recently, several studies [13], [14], [15], [16] have focused on predicting the student's performance at the end of the academic year. A study in [13] proposed a three predictive model trained and tested using 9652 students' records collected from a Portuguese Higher Institution. The dataset was collected at three different times: entry time, the end of the first semester, and the end of the first year. Four classification algorithms including RF, DT, ANN, and SVM have been used to build three prediction models where each model was constructed using different collection times. The first prediction model used 30 features collected at the entry time, while the second model used 44 features collected at the end of the first semester, whereas the last third model used 68 features collected at the entry end of the first academic year. The three prediction models that classified the first-year students into binary classes are "Failure" and "Success". They found that SVM outperformed other classification algorithms where it achieved 77% and 91% AUC for the first, and second models whereas the RF and SVM achieved equal performance which is 93% for the last third model.

A study carried out by [14] used family information variables to predict the freshmen student performance at the end of the first semester of the first academic year. In this study, the authors used a dataset collected from a university in Taiwan which contains 2407 student records and 18 independent variables of personal information, including demographic features, parents' jobs, family income. They applied four machine learning techniques, including DT (CART), DT (C5.0), RF, and ANN to build the prediction model at different output cases. This study examined different cases of the target output where the first case classified students into five classes namely Excellent, V. Good, Good, Average, and Poor. The second case classified the student into three classes which are Excellent, Normal, and Poor. While that last case classified the student as either Excellent or Poor. Finally, by using the binary class labels to predict student performance the model was achieved the best results. Also, they found that RF and DT (CART) scored the highest performance compared with other classifiers, where DT (CART) recorded 80% accuracy and the RF got 79.9% accuracy. Besides, they found also that the mother's jobs, department, father's jobs, the main source of living expenses, and the admission status are the highest influential features for predicting the students' learning performance at the end of the first semester.

A group of researchers in [15] proposed a prediction model to classify students' performance as pass or fail in the academic program. The proposed model was built by evaluating several machine learning methods, including ANN, KNN, K-Means Clustering, NB, SVM, LR, DT, and Voting ensemble. The collected data from academic institutions in UAE contains 1491 student records and 13 features, including gender, age group, school system, math level, english level, and scholarship. The Voting ensemble model outperformed other models, where it scored 75.9% overall accuracy. Finally, the authors found that the prediction model's accuracy was improved after applying the Synthetic Minority Oversampling Technique (SMOTE) to balance the proposed dataset.

Moreover, authors in [16] developed a new prediction model that integrated several classifiers, including NB, SVM, and DT classifier with Bagging and Stacking ensemble methods to classify student achievement into one of the four classes: Excellent, Good, Average, and Poor. The proposed ensemble model was trained and tested using a dataset that contains 233 instances with 45 attributes categorized into student personal information, learning pattern, behavior, emotional and cognitive factors. The proposed ensemble model achieved the highest accuracy of 97%, followed by bagging, stacking ensemble, NB, then SVM, and finally DT.

A group of studies have examined the relationship between the admission requirements such as the pre-university test and the student achievements. However, researchers have not agreed on whether the admission requirements have a strong relationship with student achievement or not. There are some studies ([38], [39]) conducted in the Kingdom of Saudi Arabia to find out whether the admission requirements affect the student's performance or not. Both studies were conducted on computer science students. However, their findings were different. The first study [38] found that the student's GPA of the secondary school affects the student's performance in higher education, while the pre-university tests such as SAAT and GAT do not affect the student's performance. However, study [39] showed that the admission test, namely SAAT, significantly predicts the student's CGPA. Moreover, most of the previous studies classified student's CGPA at graduation or predicted the student performance in a specific course, but unfortunately there is a minimal number of studies focusing on predicting student's level at the first academic year. So far, the existing studies on classifying student's achievement at the first academic year examined the non-academic factors that may affect student achievement [40] or tried to evaluate admission criteria and their impact on student's performance at the first-year [39], [41]. Additionally, it has been noticed that most of the previous works analyzed student data from Computer Science or STEM (Science, Technology, Engineering, and Math) major. In contrast, a few studies have examined student data in Arts and Humanities major, indicating the need to investigate and analyze the student data in these academic disciplines.

Table 1 summarizes the previous studies that focus on predicting student performance.

III. DESCRIPTION OF THE CHOSEN TECHNIQUES

This section presents the research methodology that was followed to achieve the objectives of this study. The essential phases of the research methodology are illustrated in

TABLE 1. Summary of the related studies on predicting student performance.

Ref	Techniques	Results	Study Sample Size	Findings
[6]	RF	Accuracy (79%-91%)	32,593	The feature engineering technique improved the prediction model where it's achieved more than 80% accuracy
[7]	Ensemble (Bagging)	Accuracy (66.7%) for dataset_1 (93.1%) for dataset_2	Dataset_1: 52 Dataset_2: 486	The highest results were achieved when considering 50% of the coursework
[3]	RF	Accuracy (59.64%)	289	The formative assessment tasks grades were able to predict at-risk students
[8]	SVM	F1-score (92%) for distance dataset F1-score (83%) for on-campus dataset	Online dataset: 262 On-campus dataset: 161	The accuracy of the prediction model has improved when the student performed 50% of distance education course and 25% of on-campus education course.
[9]	DT(J48)	Accuracy (69.3%)	339	The CGPA of the first year, and 3 courses of the first year: introductory math, computer skills, and communication skills are the most influence factors on the graduation CGPA.
[10]	NB and Hoeffding Tree	Accuracy (91%)	530	4 courses have a significant influence on the CGPA: Operating Systems, Statistics, General Physic, Computer Programming, and Algorithms course.
[11]	LR	Accuracy (89.15%)	1841	3rd year GPA is the highest influenced feature on final graduation GPA
[12]	NB	Accuracy (43.18%)	2281	Pre-admission requirement and the personal student information influence the graduation GPA
[13]	SVM for model 1 and model 2 RF and SVM for model 3	Model1: AUC (77%) Model2: AUC (91%) Model3: AUC (93%)	9652	They found that SVM outperformed RF, DT, ANN.
[14]	DT (CART)	Accuracy (80%)	2407	The mother's jobs, department, father's jobs, the main source of living expenses, and the admission status are the highest influential features
[15]	Ensemble	Accuracy (75.9%)	1491	The accuracy of the prediction model was improved after applying SMOTE technique to balance the proposed dataset
[16]	Ensemble	Accuracy (97%)	233	The proposed ensemble model outperformed bagging, stacking, NB, SVM, and DT

Figure 1. The following subsections discuss the main stages of the proposed methodology. Section A presents data pre-processing methods, section B discusses the feature selection methods applied in this study. Section C explains the classification methods that are used, while section D presents the eXplainable Artificial Intelligence (XAI) techniques that were applied in this study.

A. DATA PREPROCESSING

Pre-processing has a critical role in data mining. The primary goal of the data preparation stage is to make the raw data suitable for data mining techniques to be implemented. This section introduces the essential data pre-processing methods that were applied in this study, where the collected dataset contains missing data in some attributes. For the students in the 2018-2019 academic year, the date of birth attribute has missing data. Age is a derived attribute generated from the date of birth. It has been noticed that the age attribute correlates with the school graduation time attribute. Thus, iterative imputer is used to impute the missing values in the age attribute. Iterative imputer builds a regression model by using the attribute containing missing data as a target and the remaining attributes as independent features. The model will be trained using samples of non-missing values, then the missing data will be estimated using the prediction model [21].

1) DATA DISCRETIZATION

Data discretization by binning is used in this study to convert the continuous attribute's values (numeric) to a categorical value (nominal/ordinal). In dataset #1 The target attribute (CGPA) has converted to an ordinal attribute where first class is Not At-Risk for a CGPA greater than or equal to 3, second class is At-risk for a CGPA less than 3.

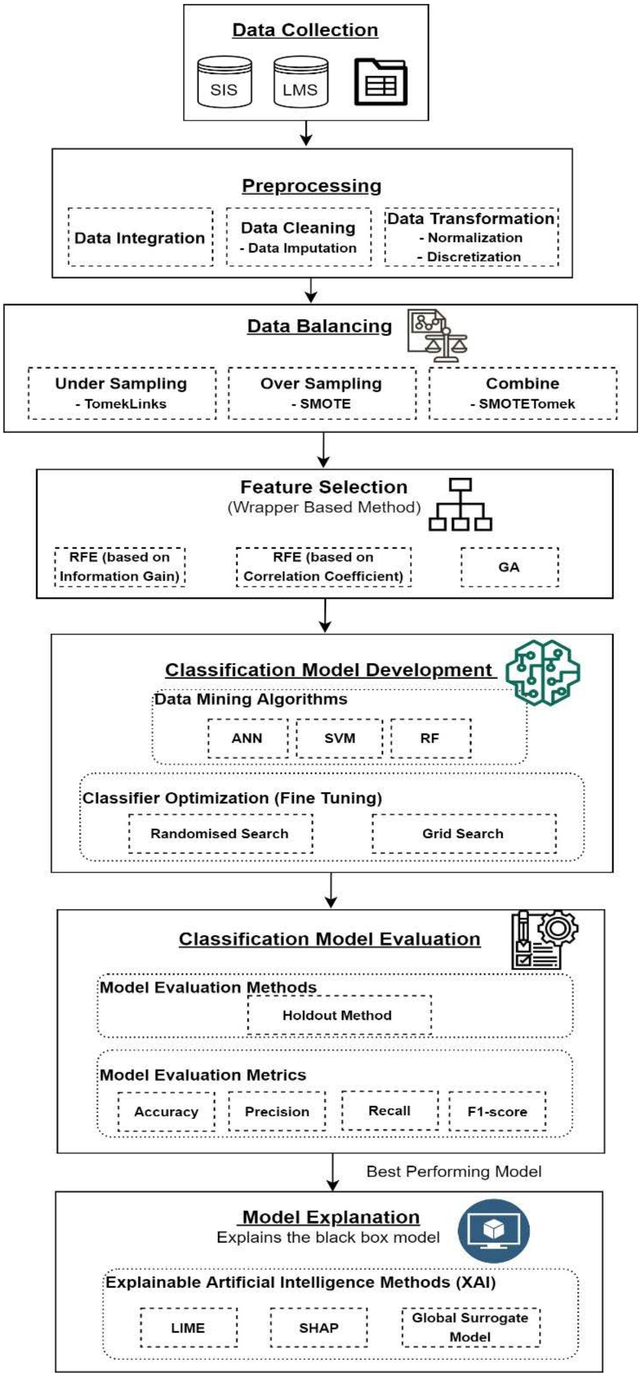


FIGURE 1. The proposed methodology of this study.

In addition, all the courses grades have converted to an ordinal attribute based on the IAU grading system where A+ (final grade >=95), A (90<= final grade <95), B+ (85<= final grade <90), B (80<= final grade<85), C+ (75<= final grade <80), C (70<= final grade <75), D+ (65<= final grade <70), D (60<= final grade <65), and F (final grade <60). Furthermore, the target class in dataset #2 and dataset #3 is the final course grade (out of 100). The target class has been converted into two categories which are Pass and Fail. The first class (Fail) is the positive class for a course grade less than or equal to 60, whereas the second class (Pass) is for a course grade greater than 60.

2) DATA ENCODING AND STANDARDIZATION

Ordinal encoding and label encoding were used in this study to map the categorical features to numerical data for further processing. StandardScaler technique (z-score normalization) was applied in this study to standardize features by rescaling the numeric attribute to have 0 for the attribute's mean and 1 for the attribute's standard deviation.

3) DATA BALANCING

The collected datasets are imbalanced datasets that have unequal class distribution. To overcome the class imbalance problem, SMOTE [22], and SMOTE-Tomek Link [23] sampling methods were employed to balance the datasets used in this study.

- SMOTE

SMOTE [22] stands for Synthetic Minority Oversampling Technique. SMOTE is an over-sampling technique that generates synthetic minority class samples by selecting a random minority class sample X and producing new minority class samples along the lines connecting X and each nearest minority neighbor.

The steps of SMOTE algorithm are

- Select a random minority class sample X from the minority class samples M.

- Find the k- nearest neighbors from the minority class samples M for sample X.

- Place a synthetic sample along the lines between minority class sample X and its k- nearest neighbors from M.

- Repeat previous steps until the dataset is balanced.

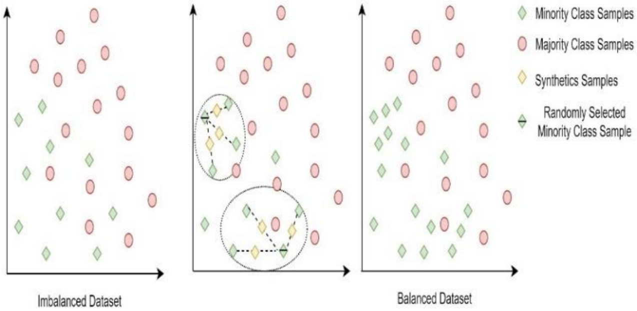


FIGURE 2. SMOTE algorithm.

- Tomek Links

Tomek Links algorithm [42] is an under-sampling technique that finds the nearest majority class sample (y) for every minority class sample (x) to generate a pair of samples from different classes, then the majority class sample (y) from the pair will be deleted.

The steps of the Tomek Links algorithm are:

- The closet sample for sample x is y.

- The closet sample for sample y is x.

- Sample x and y belong to opposite classes. Where x belongs to the minority and y belongs to the majority class.
- Delete sample y.

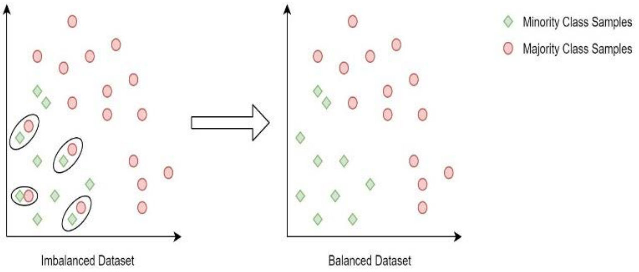


FIGURE 3. **Tomek link algorithm.**

- SMOTE-Tomek Link

SMOTE-Tomek Link [23] is a hybrid-sampling method that combines over-sampling and under-sampling techniques, which are SMOTE and Tomek Link algorithms. This algorithm starts by applying the SMOTE to generate synthetic samples for the minority class then the duplicated samples are cleaned by using the Tomek Link algorithm.

B. FEATURE SELECTION

Feature selection methods remove the irrelevant or redundant features from the dataset, where a small subset of relevant features will be used to build the final model. It has been observed that the wrapper feature selection methods including Recursive Feature Elimination (RFE) [24], [9], and Genetic Algorithm (GA) [25] have proven their success. However, in this study RFE with Pearson correlation coefficient, RFE with mutual information and GA techniques are applied to find the best relevant features for constructing the final prediction model.

C. CLASSIFICATION METHODS

The main objective of this study is to employ supervised learning techniques in order to build prediction models to identify at-risk students of failing the preparatory year. Artificial Neural Network (ANN), Support Vector Machine (SVM), and Random Forest (RF) are the classification techniques used in this study to predict the students' performance. This section presents the classification algorithms that were used in this study to build the prediction models.

1) ARTIFICIAL NEURAL NETWORK (ANN)

Artificial Neural Network gains wide attention in different areas, including the educational field. It has proven great success in analyzing complex large-sized datasets and detecting the nonlinear relationship between the features and the target attribute [26]. The architecture of ANN is inspired by the working of the human brain structure to simulate the human brain's learning process [27]. ANN structure consists of a group of connected artificial neurons where each link has an associated weight. Multilayer Perceptron (MLP) is the most common feed-forward neural network that feeds the information forward from the input to the output layer's neurons. The network neurons are organized into three layers, namely the input, hidden, and output layers, as shown in Figure 4. The input layer neurons receive the input values, where the number of neurons in the input layer is associated with the number of independent variables. In the ANN structure, at least one hidden layer computes the weighted sum of the input values then the activation function will be applied to produce the value of the output layer. Backpropagation is the learning algorithm used to adjust and modify the weight to get different better results [28], [29].

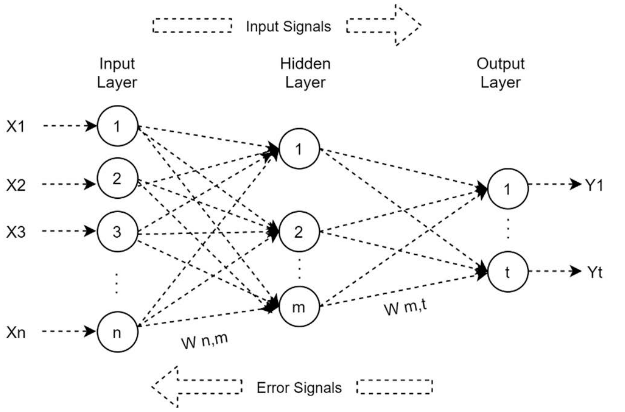


FIGURE 4. **Two layers MLP-ANN structure.**

2) SUPPORT VECTOR MACHINE (SVM)

Support vector machine is a supervised learning technique used for classification and regression to classify linear and non-linear data. SVM performs well when using a small training dataset. Moreover, SVM can handle the high dimensionality data in which the number of features in the dataset is greater and also if the number of attributes is greater than the number of samples.

In the case of the training data is linearly separable, SVM uses a mapping function called kernel to transfer the training data to a new higher dimensional space in which the data can be separable. Linear, polynomial, radial basis function (RBF), and sigmoid are some of the popular kernel functions in SVM [30]. SVM was developed to deal with the binary classification problem, where the optimal hyperplane is used to separate the data into two classes [31]. SVM finds the optimal hyperplane using the support vectors (training samples that touch the hyperplane's margins) and the margins as shown in Figure 5, where the optimal hyperplane produces the lowest classification error [32]. SVM can be used to classify multi-classes problems where a group of two-classes SVM is combined [33].

3) RANDOM FOREST (RF)

Random forest was introduced by L. Breiman in 2001 [34]. RF is a supervised machine learning algorithm that is used

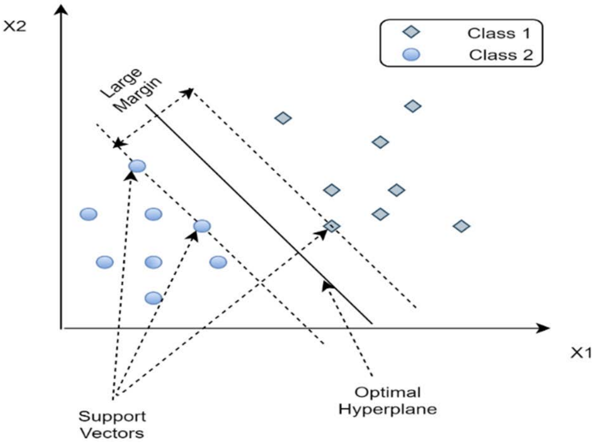


FIGURE 5. Optimal support vector machine.

for both classification and regression problems [35]. The RF has proven to be highly extremely successful in predicting student performance as it outperformed other classification techniques such as KNN, SVM, NB, and LR [34], [8], [3], [4], [36].

RF is an ensemble model that combines a group of decision trees as a base learner to get a powerful prediction model that decreases the overfitting of the training dataset. The main advantage of the RF is that it has less training time. The RF algorithm is relatively easy to learn, but it takes a long time to make predictions once it has been learned. Also, for more accurate forecasting, more trees are required, resulting in a slower and more complex model. Moreover, using multiple decision trees decrease the risk of overfitting. Furthermore, RF has the ability to estimate the missing data and it combines a group of decision trees to have a more accurate and stable prediction model. Decision trees are extremely sensitive to the training data, which might lead to a high variation. As a result, the DT model may be unable to be generalized. However, RF which is a group of multiple decision trees that is used randomness in the model construction is less sensitive to the training data.

D. EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

Recently, the field of XAI has become increasingly significant with the increase of AI applications around the world. Machine learning is one of the most popular applications that build high-performance predictive models. Although the prediction models are high-performance, some models are complex and difficult to explain, therefore the need for XAI techniques to develop high-performance and explainable models has increased. Even though the terms (explainable and interpretable) are sometimes used interchangeably, the expert analysts highlight the major distinction between the two, as the humans can understand what a model has done when it is interpretable. Whereas humans should be able to explain and figure out why and which features have a significant influence on the prediction outcome if the model is explainable [37].

Tree-based data mining methods are the most often used non-linear models nowadays. A variety of data mining models were utilized in this study to detect at-risk students at an early stage. In this study, the RF which is a tree-based approach, performed the best comparing to other data mining techniques. In educational data mining applications, such as identifying at-risk students, it is frequently important to have models that are both accurate and interpretable, where interpretable means that the model is understandable so that the user can understand how the input features were used to arrive at the model prediction.

1) LIME

LIME [17] stands for Local Interpretable Model-Agnostic Explanations where it explains the model at the individual level where it considers the outcome of a single prediction instead of the entire dataset (local interpretability). Moreover, LIME can be applied to any ML model (model-agnostic) after the model had been trained (post-hoc). To explain global ML models (the black box models), local surrogate models which are interpretable models are utilized. LIME trains surrogate model to approximate the original model's predictions by building local surrogate model that explains individual predictions rather than a global surrogate model. Figure 6(A) shows the global model or a black box model that differentiate between two classes blue and red. LIME method builds local surrogate model that explain the predictions of sample X1 as shown in Figure 6(B). A local dataset around X1 is generated by using random sampling. The labels of the local dataset will be defined by using the global prediction model. Finally, the local surrogate model used to fit the local dataset.

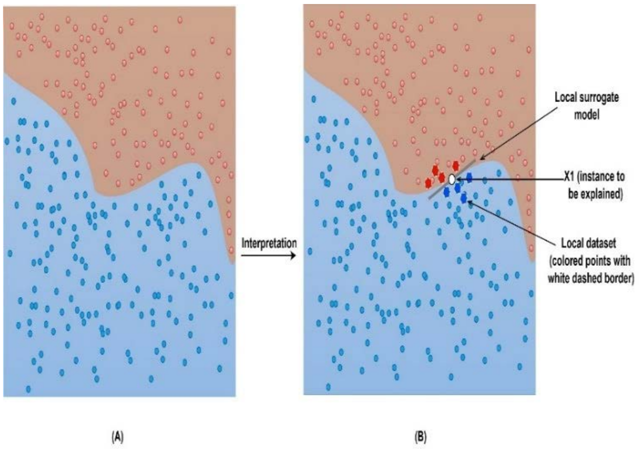

Interpretation
Local surrogate model
X1 (instance to be explained)
Local dataset (colored points with white dashed border)
(A)
(B)

FIGURE 6. Local interpretability using LIME method.

2) SHAP

SHAP [18], which stands for SHapley Additive exPlanations, is a model-agnostic method. SHAP is based on the shapley value from game theory, which focuses on how much each participant in a game (player) contributes to the pay-out, the shapley value attempting to fairly distribute the pay-out among the players assuming each player contributes differently to the game. Shapley value for player x is the amount of money that this player will receive, and it is the player's marginal contribution (which is calculated by running this game with and without this player for every possible subset of players). In ML, the shapley value indicates how much each feature contributes to the final prediction, where each feature is a player, and the prediction value is the pay-out. SHAP works for both local and global interpretability. In local interpretability SHAP answers how much each feature value of an instance contributes to the prediction output such as force plot and waterfall plot. While for the global interpretability, the summary plot presents the most significant features and their impact on the black box model.

3) GLOBAL SURROGATE MODEL

A global surrogate model is an interpretable model that approximates a complex (black box) model, where the global surrogate model could be any explainable model (decision tree, linear regression, etc.) that is trained using the same dataset of the black-box model and the black-box model's prediction output are used as a target variable. The global surrogate model is primarily used to explain the global behavior of a ‘‘black box’’ model.

IV. EMPIRICAL STUDIES

A. EXPERIMENT SETUP

The experiments were set up on Intel(R) Core (TM) i7-1065G7 CPU@ 1.50 GHz, 16 GB RAM, 64-bit Windows 10 OS using Anaconda and Jupyter notebook. Anaconda is an open-source R and Python distribution that includes a variety of Python modules and packages to make package management and deployment easier. However, Jupyter is an open-source interactive web tool, that allows researchers to combine program code, computational output, and visualizations into a single document. Jupyter is named by the programming languages supported by this tool: Julia, Python, and R. This tool is used in many applications including machine learning, data mining, data cleaning, and data visualization [19]. The experiment is carried out using Python 3.7.6 scripts for constructing the predictive models. The Python libraries used are pandas, numpy, imblearn, sklearn, genetic_selection, mlxtend, matplotlib, missingno, shap, lime, and dalex.

B. THE DESCRIPTION OF THE DATASET

This study uses three real datasets related to the preparatory year student from the humanities track at Imam Abdulrahman bin Faisal University (IAU). The humanity track is the track offered by the Deanship of the Preparatory Year and Supporting Studies at IAU. The proposed datasets were collected from several data sources, including Student Information System (SIS), Learning Management System (Blackboard), and Deanship of Preparatory Year and Supporting Studies. Students at humanity track study six courses in the first semester and seven courses in the second semester [20]. The collected data covers student information, including demographic data, pre-university data, courses grades, Cumulative Grade Points Average (CGPA), and assessments marks (detailed grades) for each course. The demographics data (date of birth, nationality, gender), pre-university data (high school grade, SAT1, SAT2), and university data (CGPA, the final grade of the 13 preparatory year courses) were extracted from the SIS databases. While the detailed grades for each course were extracted from the Deanship of the Preparatory Year and Supporting Studies' databases. On the other hand, some assessments were carried out on the Blackboard system, including online testing and assignment, and this resulted in some data being extracted from the Blackboard system.

C. DATA MINING IMPLEMENTATION

In order to identify the significant features that differentiate between the performance of the preparatory year students; feature selection methods including recursive features elimination (RFE), and genetic algorithm (GA), and three data mining classification algorithms (SVM, ANN, RF) were compared. This study went through several phases to build the classification models to meet the research objectives. Three different classification models were developed in this study, the first model used dataset #1 to classify the student as at-risk of failing the preparatory year or not and find the most influential features that were used to identify the at-risk student. After finding the highest influence courses, two models were constructed using dataset #2 and dataset #3 to predict the performance of the student in the highest influence two courses. In this study, each model was constructed by evaluating several data mining models and using different feature selection methods. Moreover, all the models were evaluated using the 70:30 holdout method (70% of the dataset used for training and validation and 30% of the dataset used for testing). Furthermore, accuracy, precision, recall, and F1-score are the performance metrics that are used in this study to evaluate the classification models.

V. RESULTS AND DISCUSSION

A. RESULTS OF THE FIRST MODEL

In order to identify the at-risk students at an early stage, some experiments were conducted to find the most influential courses for identifying whether the student is At-risk of failing the preparatory year or not. The first model aims to find the most influential features for identifying the at-risk student. Thus, the classification problem has two classes, namely At-risk, and Not At-risk. Each model went through several stages and the result of each stage is presented in the following sections.

1) RESULT OF USING THE BALANCING METHODS

To address the problem of class imbalance, two sampling approaches were evaluated to balance the dataset used in this study, including SMOTE, and SMOTE-Tomek Link. The testing set results of each classification algorithm using different sampling methods on dataset #1 are presented in Table 2. As shown in the results, it has been observed that the performance of the model increased after handling the imbalanced data using oversampling (SMOTE) and combined sampling (SMOTE-TomekLinks) methods where the highest performance of all models was achieved when using SMOTE-TomekLinks method followed by SMOTE.

TABLE 2. **Model performance of the testing set after applying balancing methods on dataset #1.**

Algorithms	Balancing Methods	Accuracy	Precision	F1	Recall
RF	Original data	99.323%	99.332%	99.323%	99.323%
	SMOTE	99.493%	99.498%	99.493%	99.493%
	SMOTE-Tomeklinks	99.590%	99.748%	99.328%	98.921%
SVM	Original data	91.940%	88.692%	86.077%	84.009%
	SMOTE	94.003%	94.307%	93.993%	94.003%
	SMOTE-Tomeklinks	94.585%	94.834%	94.578%	94.585%
ANN	Original data	95.902%	93.142%	93.376%	93.615%
	SMOTE	98.057%	98.074%	98.057%	98.057%
	SMOTE-Tomeklinks	98.223%	98.254%	98.223%	98.223%

2) RESULT OF APPLYING FEATURE SELECTION METHODS

In this study, three feature selection methods have been evaluated including RFE with Pearson correlation coefficient, RFE with mutual information, and GA. RF got the highest performance by using the RFE with mutual information as a feature selection method where the number of features decreased from 21 to 2 features. The results show that the most influential courses that were used to predict whether the student At-risk of failing the preparatory year or not are the English language courses: english language 1 (ENGL104), and english language 2 (ENGL113).

3) RESULT OF OPTIMIZATION TECHNIQUES

To find the best value of hyperparameter for each classifier, grid search and randomized search have been used. After balancing the dataset and finding the best subset of features, fine-tuning is applied to enhance the performance of the models. In this stage, the 5-fold cross-validation is applied on the training set to search for the optimal values for each model. The optimized hyperparameters' values for each algorithm are presented in the following tables.

As shown in Table 4, each classification model was built using the balanced dataset and the best subset of features, as well as the optimized hyperparameters. Finally, the highest performance model is RF where it achieved the accuracy of 99.662%. Furthermore, the most influential courses that were used to predict whether the student At-risk of failing the preparatory year or not are the English language courses: english language 1 (ENGL104), and english language 2 (ENGL113). Therefore, the assessments of these two courses will be used to identify the at-risk student at an early stage.

TABLE 3. **The optimal hyperparameters values for the SVM model.**

Classification Techniques	Hyperparameter	Optimized value
SVM	kernel	RBF
	C	21
	gamma	scale
RF	n_estimators	500
	max_features	auto
	criterion	gini
ANN	hidden_layer_sizes	(100,)
	activation	relu
	solver	adam
	learning_rate	constant (0.005)

TABLE 4. **The final model performance for the testing set of dataset #1.**

Experiment Characteristic	Algorithm	Accuracy	Precision	F1	Recall
Balanced dataset (SMOTE-Tomek Link), Best subset of features (# features = 2), Optimal hyper-parameters	RF	99.662%	100%	99.661 %	99.324%
	SVM	98.731%	100%	98.715 %	97.462%
	ANN	98.816%	100%	98.801 %	97.631%

B. RESULTS OF THE SECOND MODEL AND THIRD PMODEL

After finding the most influential features that identify the at-risk student, two more models have been constructed in order to predict the at-risk student at an early stage by using different percentages of the course assessment grades. From the first model, it has been noticed that ENGL104 and ENGL113 are the most influential features in identifying the at-risk student. This section presents the results of the two courses' models where the second model of this study used dataset #2 to predict whether the student is at risk of failing ENGL104 or not. Whereas the third model used dataset #3 to predict whether the student is at risk of failing ENGL113 or not.

1) RESULT OF USING THE BALANCING METHODS

The collected two datasets (dataset #2 and dataset #3) are unbalanced datasets with binary classes. SMOTE, and SMOTE-Tomek Link were used to handle the imbalanced dataset. As shown in Table 5, using the RF classification technique with SMOTE-TomekLinks in order to build the classification model has a great impact on the model performance, where the model achieved the highest accuracy (91.715%). Furthermore, Table 6 compares the testing set results of each classification algorithm using different sampling methods on dataset #3. As shown in Table 5 using the RF classification technique with SMOTE-TomekLinks in order to build the classification model has a great impact on the model performance, where the model achieved the highest accuracy (95.363%), highest precision (93.322%), highest recall (97.727%), and the highest F1-score (95.474%). Finally, SMOTE-TomekLinks was used as a balancing method with RF, ANN, and SVM classification methods to handle the unbalanced dataset and build the prediction model.

TABLE 5. Model performance of the testing set after applying balancing methods on dataset #2.

Algorithms	Balancing Methods	Accuracy	Precision	F1	Recall
RF	Original data	91.255%	84.066%	83.152%	82.258%
	SMOTE	91.595%	89.068%	91.867%	94.847%
	SMOTE-Tomeklinks	91.715%	88.592%	92.037%	95.761%
SVM	Original data	89.140%	82.249%	78.310%	74.731%
	SMOTE	89.780%	85.641%	90.352%	95.611%
	SMOTE-Tomeklinks	90.559%	87.389%	90.943%	94.798%
ANN	Original data	88.293%	79.769%	76.880%	74.194%
	SMOTE	90.162%	86.736%	90.611%	94.847%
	SMOTE-Tomeklinks	90.462%	88.182%	90.739%	93.449%

TABLE 6. Model performance of the testing set after applying balancing methods on dataset #3.

Algorithms	Balancing Methods	Accuracy	Precision	F1	Recall
RF	Original data	92.748%	89.076%	80.303%	73.103%
	SMOTE	95.013%	92.422%	95.165%	98.077%
	SMOTE-Tomeklinks	95.363%	93.322%	95.474%	97.727%
SVM	Original data	93.026%	89.256%	81.203%	74.483%
	SMOTE	91.076%	88.525%	91.371%	94.406%
	SMOTE-Tomeklinks	93.991%	91.540%	94.169%	96.953%
ANN	Original data	92.469%	82.270%	81.119%	0.800%
	SMOTE	92.563%	90.516%	92.754%	95.105%
	SMOTE-Tomeklinks	92.913%	89.789%	93.188%	96.853%

TABLE 7. Model performance of the testing set using different percentage of the ENGL 104 course length (dataset #2).

Model	Selected Features and Timeline	# Features	Accuracy	Precision	F1	Recall
RF	Admission	8	75.43%	75.19%	75.55%	75.92%
	Admission +Week 2	9	76.01%	75.19%	76.40%	77.65%
	Admission +Week 4	11	80.83%	80.08%	81.07%	82.08%
	Admission +Week 5	12	81.41%	79.74%	81.91%	84.20%
	Admission +Week 6	13	90.08%	87.68%	90.38%	93.26%
	Admission +Week 7	14	89.21%	86.67%	89.57%	92.68%
	Admission +Week 9	15	90.17%	87.98%	90.45%	93.06%
	Admission +Week 10	16	91.91%	89.33%	92.16%	95.18%
	Admission +Week 12	17	92.20%	89.11%	92.49%	96.15%
	Admission +Week 15	19	91.71%	88.05%	92.10%	96.53%
SVM	Admission	8	66.47%	64.72%	68.36%	72.45%
	Admission +Week 2	9	68.30%	66.27%	70.17%	74.57%
	Admission +Week 4	11	70.62%	69.04%	71.79%	74.76%
	Admission +Week 5	12	69.94%	67.81%	71.64%	75.92%
	Admission +Week 6	13	86.80%	83.99%	87.33%	90.94%
	Admission +Week 7	14	86.99%	84.41%	87.47%	90.75%
	Admission +Week 9	15	87.76%	85.90%	88.08%	90.37%
	Admission +Week 10	16	90.08%	87.41%	90.42%	93.64%
	Admission +Week 12	17	90.66%	87.81%	90.99%	94.41%
	Admission +Week 15	19	90.56%	87.39%	90.94%	94.80%
ANN	Admission	8	67.92%	66.79%	68.97	71.29%
	Admission +Week 2	9	69.56%	67.12%	71.58	76.69%
	Admission +Week 4	11	73.31%	71.61%	74.33	77.26%
	Admission +Week 5	12	73.12%	72.14%	73.70	75.34%
	Admission +Week 6	13	87.48%	85.95%	87.74	89.60%
	Admission +Week 7	14	88.54%	87.31%	88.72	90.17%
	Admission +Week 9	15	88.73%	87.22%	88.95	90.75%
	Admission +Week 10	16	89.11%	86.78%	89.45	92.29%
	Admission +Week 12	17	90.46%	88.75%	90.67	92.68%
	Admission +Week 15	19	90.94%	88.22%	90.94	93.83%

2) RESULT OF USING DIFFERENT PERCENTAGES OF THE COURSE ASSESSMENT GRADES

In this study, several prediction models were constructed using several classification algorithms and considering different percentages of the course module length. This model aimed to identify the at-risk student of failing the course at an early stage (before the final exam). However, the model predicts the student's final grade of the course using the student's assessment grades over multiple periods of the course length. Based on the course syllabus used at the deanship of the preparatory year and supporting studies, several prediction models were built based on different assessment grades during the academic semester (16 weeks before the final exam). Table 7 presents the prediction models for the ENGL104 course using different percentages of the course length (as shown in appendix) and the best sampling method of each classification algorithm discussed in the previous section. Many experiments have been done for each classification algorithm where the highest performance was achieved by RF followed by SVM, and ANN. RF model achieved the highest accuracy (92.20%), and the highest F1-score (92.49%) when considering 17 features including the admission features (*SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age*), and 40% of the assessments' grades (the assessments up to week 12: *WP 1, WP 2, E-Learning, Practice test, Midterm 1, WP 3, WP 4, Midterm 2, WP 5).* Whereas SVM and ANN obtained approximately the same accuracy as the SVM model achieved 90.66% when considering the admission features and 40% of the assessments' grades (the assessments up to week 12: *WP 1, WP 2, E-Learning, Practice test, Midterm 1, WP 3, WP 4, Midterm 2, WP 5),* while the ANN got 90.94% when considering the admission features and 50% of the assessments' grades (the assessments up to week 15: *WP 1, WP 2, E-Learning, Practice test, Midterm 1, WP 3, WP 4, Midterm 2, WP 5, Attendance, Participation.)*

In addition, Table 8 presents the prediction models for the ENGL113 course using different percentages of the course length (as shown in the appendix) and the best sampling method of each classification algorithm discussed in the previous section. Many experiments have been done for each classification algorithm where the highest performance was achieved by RF followed by ANN, and SVM. RF model achieved the highest accuracy (95.36%), the highest precision (93.91%), and the highest F1-score (95.44%) when considering 18 features including the admission features (*SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age*), First semester GPA (*GPA 1*), and 60% of the assessments' grades (the assessments up to week 15: *WP 1, E-Learning, Midterm 1, WP 2, Midterm 2, WP 3, Speaking, Attendance, Participation*). Whereas ANN and SVM obtained approximately the same accuracy as the ANN model achieved 93.70% while the SVM got 93.09% when considering 16 features including the admission features (*SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age*), First semester GPA (*GPA 1*), and 50% of the assessments' grades (the assessments up to week 14: *WP 1, E-Learning, Midterm 1, WP 2, Midterm 2, WP 3, Speaking*).

TABLE 8. Model performance of the testing set using different percentage of the ENGL 113 course length (dataset #3).

Model	Selected Features and Timeline	# Features	Accuracy	Precision	F1	Recall
RF	Admission	8	80.84%	79.86%	81.17%	82.52%
	Admission + GPA1	9	89.50%	87.05%	89.85%	92.83%
	Admission + GPA1 + Week 4	11	89.85%	86.77%	90.27%	94.06%
	Admission + GPA1 + Week 6	12	93.88%	92.98%	93.94%	94.93%
	Admission + GPA1 + Week 10	14	94.23%	92.17%	94.37%	96.68%
	Admission + GPA1 + Week 14	16	94.93%	92.98%	95.04%	97.20%
	Admission + GPA1 + Week 15	18	95.36%	93.91%	95.44%	97.03%
SVM	Admission	8	72.53%	68.53%	75.24%	83.39%
	Admission + GPA1	9	83.99%	81.83%	84.53%	87.41%
	Admission + GPA1 + Week 4	11	85.21%	82.76%	85.76%	88.99%
	Admission + GPA1 + Week 6	12	89.50%	87.05%	89.85%	92.83%
	Admission + GPA1 + Week 10	14	90.81%	88.47%	91.09%	93.88%
	Admission + GPA1 + Week 14	16	93.09%	90.08%	93.34%	96.85%
	Admission + GPA1 + Week 15	18	92.83%	89.39%	93.13%	97.20%
ANN	Admission	8	72.79%	69.57%	74.90%	81.12%
	Admission + GPA1	9	84.78%	81.99%	85.43%	89.16%
	Admission + GPA1 + Week 4	11	86.00%	84.33%	86.35%	88.46%
	Admission + GPA1 + Week 6	12	89.41%	87.27%	89.72%	92.31%
	Admission + GPA1 + Week 10	14	91.16%	88.67%	91.45%	94.41%
	Admission + GPA1 + Week 14	16	93.70%	91.39%	93.88%	96.50%
	Admission + GPA1 + Week 15	18	93.35%	91.20%	93.53%	95.98%

3) RESULT OF OPTIMIZATION TECHNIQUES

Grid search and randomized search were used to find the optimal hyperparameters' value for each classifier. After handling the imbalanced dataset and finding the best percentage of the course length, the optimization is applied to enhance the performance of the models. In this stage, the 5-fold cross-validation is applied on the training set to search for the optimal values for each model. Table 9 shows the performance of the models after applying the optimization technique where the optimal hyperparameters' values that used to build the prediction models for ENGL104 and ENGL 113 datasets are presented in Table 10. As shown in Table 9, the performance of the ANN model was improved after applying the fine-tuning technique where the accuracy increased from 90.94%% to 92.00%, the precision from 88.29% to 89.07%, the F1 from 91.25% to 92.29%, and the recall from 94.41% to 95.76%. Moreover, Table 9 presents the performance of the models for the ENGL113 course where the performance of RF, SVM, and ANN was improved after applying the fine-tuning technique.

TABLE 9. The testing set model performance using the optimized hyperparameters values for the ENGL104 and ENGL113 models.

Model		Accuracy	Precision	F1	Recall
ENGL104 Model	RF	92.20%	89.11%	92.49%	96.15%
	SVM	90.66%	87.81%	90.99%	94.41%
	ANN	92.00%	89.07%	92.29%	95.76%
ENGL113 Model	RF	95.36%	93.47%	95.47%	97.55%
	SVM	93.61%	90.83%	93.83%	97.03%
	ANN	94.49%	92.35%	94.63%	97.03%

TABLE 10. The optimized hyperparameters' values for the ENGL104 and ENGL113 models.

Classification Techniques	Optimized value for ENGL104 Model	Optimized value for ENGL113 Model
SVM	kernel = RBF	kernel = RBF
	C=1	C = 21
	gamma=scale	gamma = Auto
		coef0 = 2.001
RF	n_estimators =100	n_estimators =1500
	max_features =auto	max_depth =18
	criterion =gini	criterion = gini
		bootstrap = False
ANN	hidden_layer_sizes = (100,)	hidden_layer_sizes = (700,)
	solver = adam	solver = adam
	alpha = 0.0005	alpha = 0.0005
	Max_iter = 3000	Max_iter = 3000
	activation =tanh	

C. RESULTS OF EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) TECHNIQUES

1) LIME EXPLANATIONS

As mentioned earlier LIME explains the prediction at the instance level where it explains the prediction outcome of a single instance. LIME generates a series of explanations that show how each feature value contributes to the prediction result. LIME offers a local explanations as well as determining which feature will have the highest effect on the prediction outcome. Figure 7, and Figure 9 show a visual explanation created by LIME for Pass and Fail data samples. The left graph of LIME visualization shows the prediction probabilities for binary class label (Fail and Pass), whereas the right table shows the actual value for each feature, with only features utilized in the explanation shown. The most significant features are returned in the center graph. Two colors represent the binary classification problem, orange and blue. The orange bars show the features support the Pass class, whereas the features support the Fail class are shown in blue. The relative significance of each feature is represented by float point numbers on the colored bars.

Figure 7 shows the results of the sample explanation for the Fail sample of the ENGL 104 course dataset (dataset #2). As shown in Figure 7, the prediction probability for the class to be Fail in this instance is 0.87. Therefore, the highest three features that support the prediction (Fail class) and have a negative influence on the prediction are Midterm 2, Midterm 1, and Nationality. The Fail class is supported by Midterm 2 = 3.98, Midterm 1 = 4.39, and Nationality = 0 (Saudi). Midterm 2 = 3.98 has a coefficient value of −0.31, Midterm 1=4.39 has a coefficient value of −0.19, and Nationality = 0 has a coefficient value of −0.17, showing that they have an impact on the final predictive outcome. Additionally, Figure 8 presents the coefficient, which depicts the impact of each feature on the final prediction class label. Moreover, the explanations of Figure 8 are presented in Table 11.

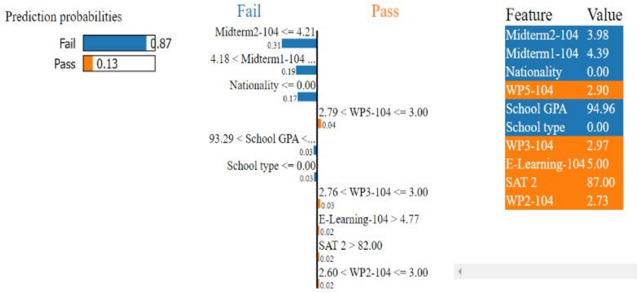
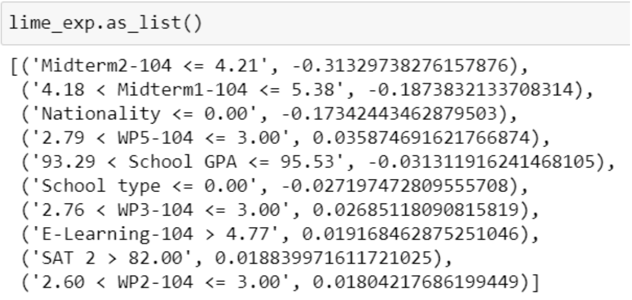


FIGURE 7. Results of LIME explanation for RF model applied to first Fail class sample of the ENGL 104 dataset (dataset #2).



```
lime_exp.as_list()

[('Midterm2-104 <= 4.21', -0.31329738276157876),
 ('4.18 < Midterm1-104 <= 5.38', -0.1873832133708314),
 ('Nationality <= 0.00', -0.17342443462879503),
 ('2.79 < WP5-104 <= 3.00', 0.035874691621766874),
 ('93.29 < School GPA <= 95.53', -0.031311916241468105),
 ('School type <= 0.00', -0.027197472809555708),
 ('2.76 < WP3-104 <= 3.00', 0.02685118090815819),
 ('E-Learning-104 > 4.77', 0.019168462875251046),
 ('SAT 2 > 82.00', 0.018839971611721025),
 ('2.60 < WP2-104 <= 3.00', 0.01804217686199449)]
```

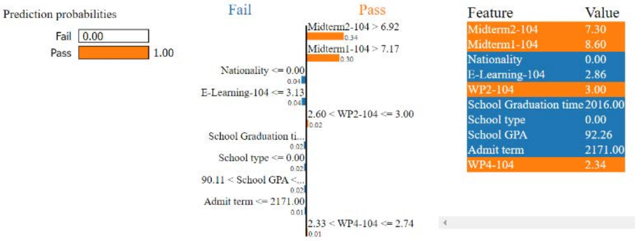
FIGURE 8. Explanation and coefficient of the fail sample.

Figure 9 shows the results of the sample explanation for the Pass sample of the ENGL 104 course dataset (dataset #2). As shown in Figure 9, the prediction probability for the class to be Pass for this instance is 1. In this case, the

TABLE 11. Descriptive explanation for the top 5 features of Fail sample.

Feature Name	Feature value	Explanation
Midterm 2	3.98	The student got low grade (3.98) which was <= 4.21 in midterm 2. Midterm 2 was the highest influential feature that support the Fail class prediction. It has a negative impact where the impact of this feature value on the prediction result is 31%.
Midterm 1	4.39	The student got low grade (4.39) in midterm 1 which was fall between 4.18 and 5.38. Midterm 1 was the second highest influential feature that support the Fail class prediction. It has a negative impact where the impact of this feature value on the prediction result is 18%.
Nationality	0 (Saudi)	The student Nationality = Saudi was the third highest influential feature that support the Fail class prediction. It has a negative impact where the impact of this feature value on the prediction result is 17%.
WP 5	2.90	The student got high grade (2.90 out of 3) in WP 5 assignment which was fall between 2.79 and 3. WP5 was the fourth highest influential feature that support the Pass class prediction. It has a positive impact where the impact of this feature value on the prediction result is 3%.
School GPA	94.96	The student got School GPA (94.96) which was fall between 94.29 and 95.53 School GPA was the fifth highest influential feature that support the Fail class prediction. It has a negative impact where the impact of this feature value on the prediction result is 3%.

highest three features that support the prediction to be Pass and have a positive influence on the prediction value are Midterm 2, Midterm 1, and WP 2. The Pass class is supported by Midterm 2 = 7.30, Midterm 1 = 8.60, and WP 2 = 3. Midterm 2 = 7.30 has a coefficient value of 0.34, Midterm 1 = 8.60 has a coefficient value of 0.30, and WP 2 = 3 has a coefficient value of 0.2, showing that they have an impact on the final predictive outcome. Additionally, Figure 10 presents the coefficient, which depicts the impact of each feature on the final prediction class label. Moreover, the explanations of Figure 10 are presented in Table 12.

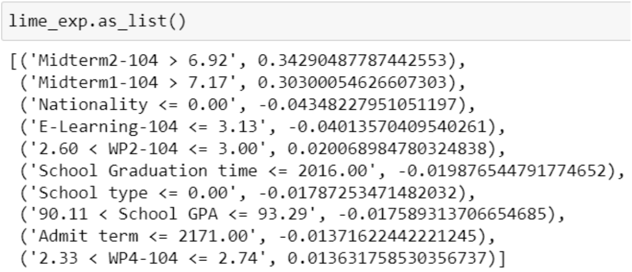


Feature	Value
Midterm2-104	7.30
Midterm1-104	8.60
Nationality	0.00
E-Learning-104	2.86
WP2-104	3.00
School Graduation time	2016.00
School type	0.00
School GPA	92.26
Admit term	2171.00
WP4-104	2.34

FIGURE 9. Results of LIME explanation for RF model applied to first pass class sample of the ENGL 104 dataset (dataset #2).

2) SHAP EXPLANATIONS

SHAP explains the ML model using the shapley value, as this value is used to measure the contribution of each feature to the predictive model. SHAP is useful for both local and global explainability since it can explain predictions for a single data sample as well as for the entire dataset.



```
lime_exp.as_list()
```

```
[('Midterm2-104 > 6.92', 0.34290487787442553),
 ('Midterm1-104 > 7.17', 0.30300054626607303),
 ('Nationality <= 0.00', -0.04348227951051197),
 ('E-Learning-104 <= 3.13', -0.04013570409540261),
 ('2.60 < WP2-104 <= 3.00', 0.020068984780324838),
 ('School Graduation time <= 2016.00', -0.019876544791774652),
 ('School type <= 0.00', -0.01787253471482032),
 ('90.11 < School GPA <= 93.29', -0.017589313706654685),
 ('Admit term <= 2171.00', -0.01371622442221245),
 ('2.33 < WP4-104 <= 2.74', 0.013631758530356737)]
```

FIGURE 10. Explanation and coefficient of the pass sample.

TABLE 12. Descriptive explanation for the top 5 features of pass sample.

Feature Name	Feature value	Explanation
Midterm 2	7.30	The student got a grade of 7.30 out of 10 which was > 6.92 in midterm 2. Midterm 2 was the highest influential feature that support the Pass class prediction. It has a positive impact where the impact of this feature value on the prediction result is 34%.
Midterm 1	8.60	The student got high grade (8.60) in Midterm 1 which was > 7.17. Midterm 1 was the second highest influential feature that support the Pass class prediction. It has a positive impact where the impact of this feature value on the prediction result is 30%.
Nationality	0 (Saudi)	The student Nationality = Saudi was the third highest influential feature that support the Fail class prediction. It has a negative impact where the impact of this feature value on the prediction result is 4%.
eLearning	2.86	The student got low grade (2.86 out of 5) in eLearning which was <= 3.13. eLearning was the fourth highest influential feature that support the Fail class prediction. It has a negative impact where the impact of this feature value on the prediction result is 4%.
WP 2	3	The student got high grade (3 out of 3) in WP 2 assignment satisfy the following condition: 2.60 < WP 2 <= 3. WP 2 was the fifth highest influential feature that support the Pass class prediction. It has a positive impact where the impact of this feature value on the prediction result is 2%.

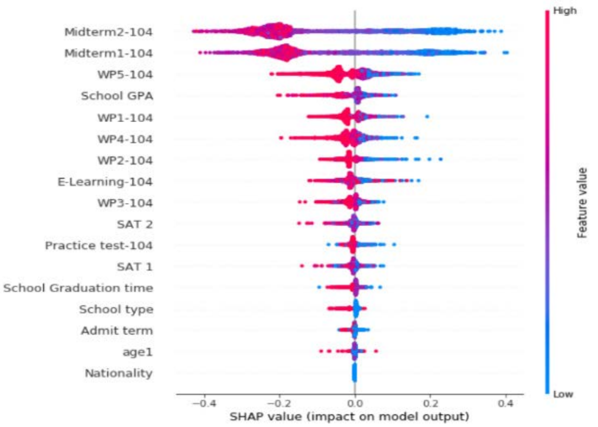


FIGURE 11. SHAP summary plot for entire ENGL 104 dataset explanation (dataset #2).

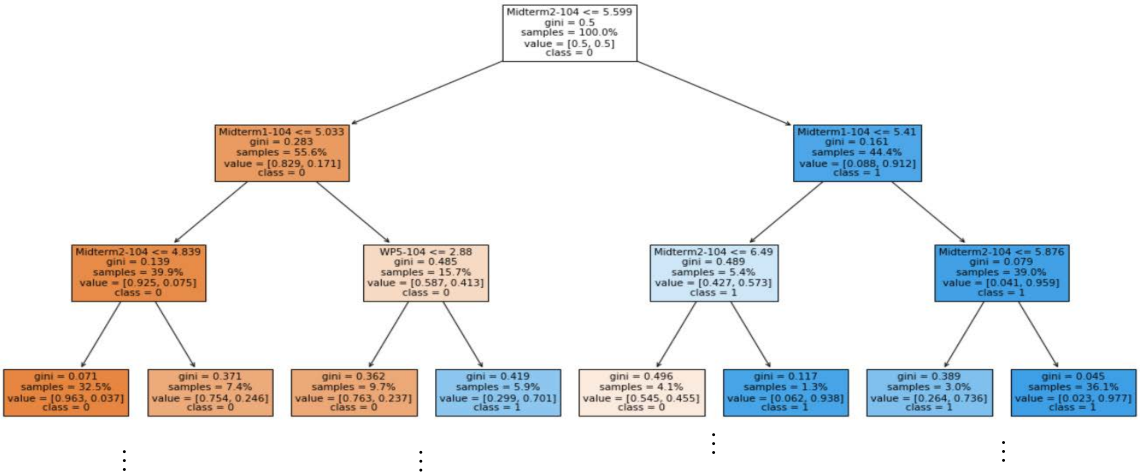


FIGURE 12. The first three layers of the global surrogate DT model for the complex RF model of ENGL 104 (dataset #2).

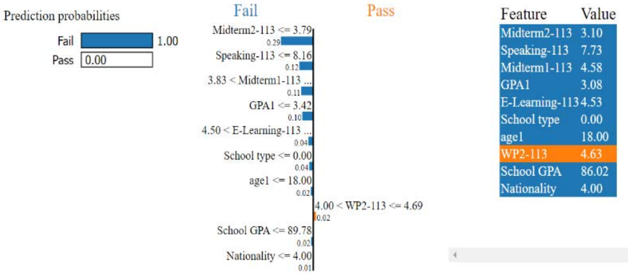
SHAP, as previously stated, displays the global contribution or importance of features across the entire dataset. All data points were shown on the summary plot with colored dots, as illustrated in Figure 11. The shapley value is shown on the x-axis, and the features are sorted using the absolute sum of the shapley values of all samples on the y-axis. In addition, the color of the dot indicates whether the data point has a higher or lower value (red denoting a higher value and blue denoting a lower value). Lower values of Midterm 2 and Midterm 1 have a positive influence on the prediction in the ENGL 104 dataset, whereas higher values have a negative impact, since Fail is the positive class in this study. Moreover, school GPA which is a pre-admission feature has the fourth highest influential feature on the prediction model. However, the remaining pre-admission features including age, nationality, SAT 1, admit term, and SAT 2 scored the lowest impact and contribution on the prediction model.

3) GLOBAL SURROGATE MODEL

Figure 12 depicts surrogate decision tree models that approximate the complicated RF model of the ENGL 104 dataset. The RF model achieved 92.20% accuracy as shown in Table 9. Figure 12 shows the first three layers of the global surrogate model that achieved 91.84 accuracy when using 5 variables as maximum depth and at most 5 variables to train the surrogate model.

Table 13 summarizes the extracted rules from the surrogate DT model in Figure12. Table 13 presents the top 12 rules that are based on a large number of samples.

As shown in Table 13, Midterm 2, and Midterm 1 can determine the CGPA level of the preparatory year students. From Table 13, rules 1, 4, 8, and 9 show that most of the Pass students achieved greater than 5.599 in Midterm 2 and at least 5.033 in Midterm 1 while the Fail students achieved in Midterm 2 less than 5.599 and less than 5 in Midterm 1 as shown in rules 3, 6, and 7. In addition, a student's score of more than 5 in Midterm 1 is not a sufficient indication of his success, as shown in rule 12.



Feature	Value
Midterm2-113	3.10
Speaking-113	7.73
Midterm1-113	4.58
GPA1	3.08
E-Learning-113	4.53
School type	0.00
age1	18.00
WP2-113	4.63
School GPA	86.02
Nationality	4.00

FIGURE 13. Results of LIME explanation for RF model applied to first fail class sample of the ENGL 113 dataset (dataset #3).

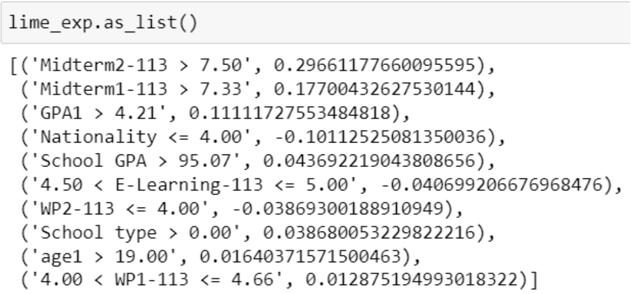
The same explanations used for explaining the RF model (black box model) of ENGL 104 dataset have been applied to final prediction model of the ENGL 113 dataset (dataset #3). Figure 13, Figure 15 show the result of applying LIME explanation for the RF prediction model that was built using ENGL 114 dataset (dataset #3). However, the explanation and the coefficient of each LIME explanation are presented in Figure 14, Figure 16.

SHAP, as previously mentioned, shows the global contribution of each feature across the entire dataset to the prediction model as illustrated in Figure 17.

Figure 18 depicts surrogate decision tree models that approximate the complicated RF model of the ENGL 113 dataset. The RF model achieved 95.36% accuracy as shown in Table 9. Figure 18 shows the first three layers of the global surrogate model that achieved 92.56% accuracy when using 5

TABLE 13. Extracted rules from surrogate DT of ENGL 104 (dataset #2).

No.	Rules	Based on
1	**IF** (Midterm2> 5.599) and (Midterm1 > 5.41) and (E-Learning> 0.05) **then** Pass (100.0%)	704 samples
2	**IF** (Midterm2<= 4.839) and (Midterm1<= 4.276) and (School GPA <= 99.1) **then** Fail (98.53%)	545 samples
3	**IF** (Midterm2 <= 5.599) and (Midterm1<= 5.831) and (WP5 <= 2.88) **then** Fail (87.8%)	164 samples
4	**IF** (Midterm2> 5.599) and (Midterm1> 5.41) and (E-Learning> 1.985) **then** Pass (92.86%)	154 samples
5	**IF** (Midterm2 > 4.244) and (Midterm1 > 5.033) and (WP5 > 2.88) and (E-Learning <= 4.953) **then** Pass (86.21%)	87 samples
6	**IF** (Midterm2<= 5.599) and (Midterm1<= 5.033) and (WP5<= 2.757) and (School GPA > 89.772) **then** Fail (95.77%)	71 samples
7	**IF** (Midterm2<= 5.599) and (Midterm1<= 5.033) and (WP5> 2.757) and (E-Learning<= 4.98) **then** Fail (50.7%)	71 samples
8	**IF** (Midterm2> 5.599) and (Midterm1> 5.41) and (WP5> 2.108) **then** Pass (85.19%)	54 samples
9	**IF** (Midterm2 >= 5.599) and (Midterm1> 5.831) and (WP5 >2.07) **then** Pass (71.05%)	38 samples
10	**IF** (5.599<Midterm2<=6.49) and (4.553<Midterm1<= 5.41) and (WP5> 2.546) **then** Pass (73.53%)	34 samples
11	**IF** (Midterm2<= 5.599) and (Midterm1> 5.033) and (WP5> 2.99) and (E-Learning > 4.953) **then** Pass (62.96%)	27 samples
12	**IF** (Midterm2<= 4.244) and (Midterm1> 5.033) and (WP > 2.88) and (E-Learning<= 4.953) **then** Fail (53.33%)	15 samples



```
lime_exp.as_list()
```

```
[('Midterm2-113 > 7.50', 0.29661177660095595),
 ('Midterm1-113 > 7.33', 0.17700432627530144),
 ('GPA1 > 4.21', 0.11111727553484818),
 ('Nationality <= 4.00', -0.10112525081350036),
 ('School GPA > 95.07', 0.043692219043808656),
 ('4.50 < E-Learning-113 <= 5.00', -0.040699206676968476),
 ('WP2-113 <= 4.00', -0.03869300188910949),
 ('School type > 0.00', 0.038680053229822216),
 ('age1 > 19.00', 0.016403715715004463),
 ('4.00 < WP1-113 <= 4.66', 0.012875194993018322)]
```

FIGURE 14. Explanation and coefficient of the fail sample.

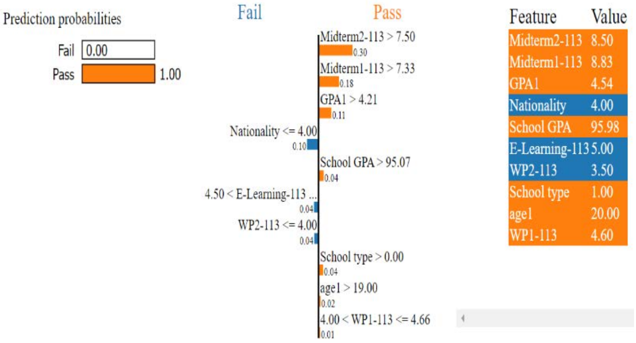
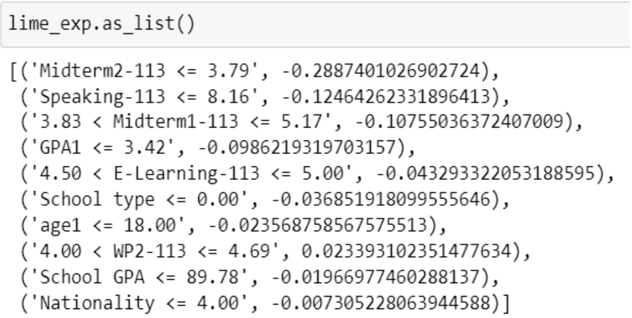


FIGURE 15. Results of LIME explanation for RF model applied to first pass class sample of the ENGL 113 dataset (dataset #3).



```
lime_exp.as_list()
```

```
[('Midterm2-113 <= 3.79', -0.2887401026902724),
 ('Speaking-113 <= 8.16', -0.12464262331896413),
 ('3.83 < Midterm1-113 <= 5.17', -0.10755036372407009),
 ('GPA1 <= 3.42', -0.0986219319703157),
 ('4.50 < E-Learning-113 <= 5.00', -0.043293322053188595),
 ('School type <= 0.00', -0.036851918099555646),
 ('age1 <= 18.00', -0.023568758567575513),
 ('4.00 < WP2-113 <= 4.69', 0.023393102351477634),
 ('School GPA <= 89.78', -0.019669774602881 37),
 ('Nationality <= 4.00', -0.007305228063944588)]
```

FIGURE 16. Explanation and coefficient of the pass sample.

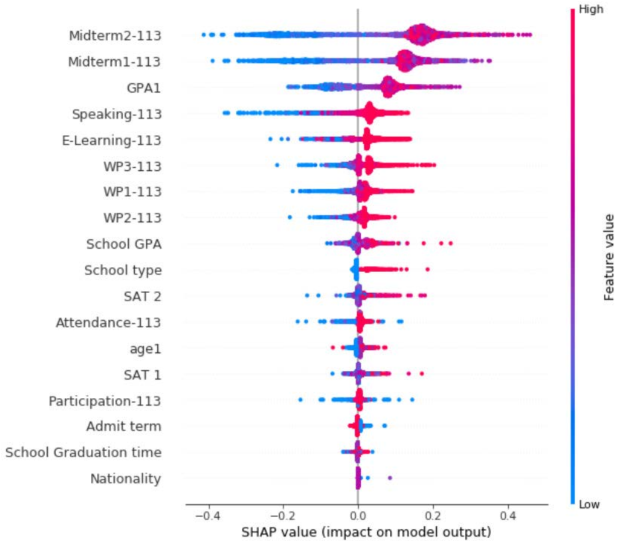


FIGURE 17. SHAP summary plot for entire ENGL 113 dataset explanation (dataset #3).

variables as maximum depth and at most 5 variables to train the surrogate model.

Table 14 summarizes the extracted rules from the surrogate DT model in Figure 18. Table 14 presents the top 12 rules that are based on a large number of samples.

As shown in Table 14, Midterm 1, and the Midterm 2 can determine the CGPA level of the preparatory year students. From Table 14, it has been noticed that most of the Pass students achieved at least 5 out of 10 in Midterm 1 while the Fail students achieved in Midterm 1 less than 4.9 as shown in Rule 2, 6, 8, 9, and 10. In addition, from rules 1, 4, and 11, the Pass students achieved greater than 5.164 in Midterm 2 while the Fail students achieved less than 5.164 as shown in rules 3, and 9.

VI. FINDINGS AND DISCUSSION

In recent years, researchers have shown great interest in providing a solution for existing academic problems that were

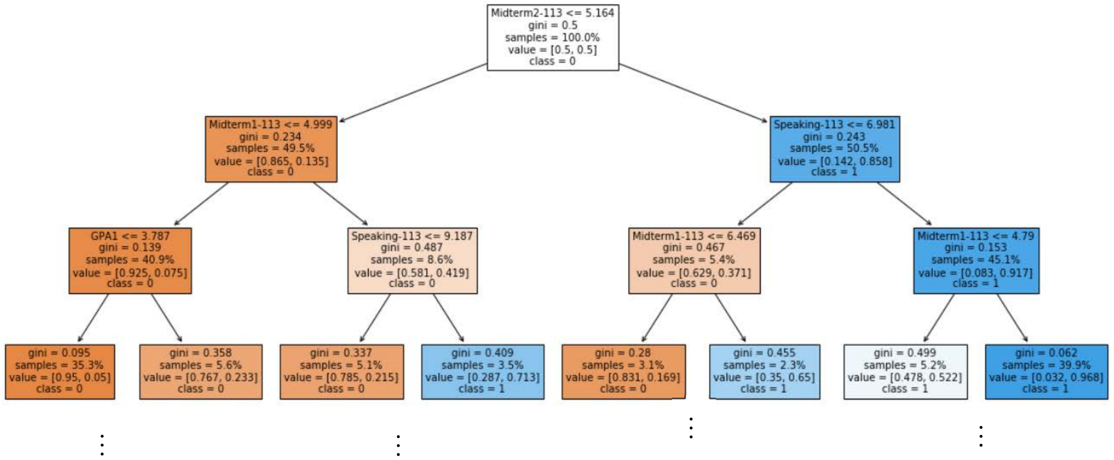


FIGURE 18. The first three layers of the global surrogate DT model for the complex RF model of ENGL 113 (dataset #3).

TABLE 14. Extracted rules from surrogate DT of ENGL 113 (dataset #3).

No.	Rules	Based on
1	**IF** (Midterm2 > 5.164) and (Speaking > 6.981) and (Midterm1 > 6.294) and (GPA1 > 2.896) **then** Pass (99.31%)	867 samples
2	**IF** (Midterm2<= 4.99) and (Midterm1 <= 4.999) and (GPA1 <= 3.787) and (E-Learning<= 5.0) **then** Fail (98.73%)	550 samples
3	**IF** (Midterm2<= 5.164) and (Midterm1<= 4.488) and (GPA1 <= 3.787) **then** Fail (92.56%)	336 samples
4	**IF** (Midterm2 > 5.164) and (Speaking > 6.981) and (GPA1 > 2.896) and (4.79 <Midterm1<= 6.294) **then** Pass (87.3%)	189 samples
5	**IF** (Midterm2<= 6.051) and (Speaking > 6.981) and (Midterm1 <= 4.79) **then** Fail (76.56%)	64 samples
6	IF (Midterm2<= 4.631) and (Midterm1<= 4.999) and (GPA1 > 3.787) **then** Fail (75.93%)	54 samples
7	**IF** (2.917< Midterm2<= 5.164) and (Midterm1> 4.999) and (Speaking> 9.187) and (GPA1 > 3.667)) **then** Pass (92.45%)	53 samples
8	**IF** (Midterm2<= 3.787) and (Midterm1<= 4.999) and (GPA1 > 3.787) **then** Fail (94.34%)	53 samples
9	**IF** (Midterm2 <= 5.164) and (Midterm1<= 4.999) and (GPA1 <= 3.787) **then** Fail (70.21%)	47 samples
10	**IF** (Midterm2 <= 4.323) and (Midterm1<= 4.999) and (Speaking<=9.187) **then** Fail (87.5%)	40 samples
11	**IF** (Midterm2 > 5.164) and (Speaking > 6.981) and (Midterm1>= 4.79) **then** Pass (97.22%)	36 samples
12	**IF** (Midterm2 > 4.323) and (Midterm1 > 4.999) and (Speaking <= 7.784) **then** Pass (60.61%)	33 samples

faced by students such as failure, low academic achievement, and dropout. As mentioned before, one of the problems that students faced in most Saudi universities that apply preparatory year system is getting a low CGPA that does not qualify them to enroll in the college/major of their interest. Similarly, the number of failed students is increased in some preparatory year tracks, especially the humanities track. Thus, identifying the at-risk preparatory year student at an early stage will contribute to solve the previous problems. Moreover, this study aims to classify the CGPA of the preparatory year student and identify the influencing features that affect the CGPA level of the preparatory year students.

In this study, three prediction models have been developed by applying several EDM techniques to achieve the objectives of this study. The study was conducted using three datasets of preparatory year students at the humanity track enrolled in IAU. The students' first-year courses and grades were used to construct three classifiers; where the first model classifies the student into two classes: At-risk and Not At-risk of failing the preparatory year. The second and third models predict whether the student will pass the course or not, however the second and third models are constructed using the assessment and the detailed grades of ENGL104 and ENGL113, respectively.

Moreover, three well-known data mining algorithms, namely: RF, ANN, and SVM have been evaluated. The results show that RF scored the highest performance when constructing the three models. Furthermore, to show the most influential courses for identifying whether the student is at-risk of failing the preparatory year or not, several feature selection methods including RFE with Pearson correlation coefficient, RFE with information gain, GA have been used. Moreover, the results from the first model shows that the most influential courses for identifying whether the student is at-risk of failing the preparatory year or not are ENGL104 and ENGL113. However, by using the highest impacted courses selected by RFE with mutual information and the RF model, the first model was able to classify the at-risk student with 99.662 % accuracy.

Additionally, the second and third models were built to identify at-risk students of failing the preparatory year at an early stage, where they were able to identify the at-risk student of failing the preparatory year by using 20% of the course assessments grades, where the RF achieved an accuracy of 90.1% after applying the first midterm of ENGL104 course in week 6, and accuracy of 93.9% after applying the first midterm in week 6 of the ENGL 113 course. Furthermore, the outcomes of this study revealed that pre-admission requirements could be used to predict the CGPA level of preparatory year students, indicating their value in forecasting the CGPA level of preparatory year students and identifying at-risk students at an early stage. RF and the best subset of pre-admission features including SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age were able to predict at-risk students of failing ENGL 104, ENGL 113 course at an early stage with reasonable accuracy 75.43% % and 80.84%, respectively.

Finally, to enhance the explainability of the prediction model, the explainable AI techniques including LIME, SHAP, and global surrogate model were applied in this study to explain the complex prediction models, explain the prediction output and highlight the reasons behind the failure.

VII. CONCLUSION AND FUTURE WORK

In conclusion, the findings of this study show that using pre-admission data, university grades, and demographic data were able to train DM algorithms including RF, SVM, and ANN to identify at-risk students of failing the preparatory year at an early stage. The RF outperformed other classification methods. Moreover, applying the feature selection technique was also improve the prediction results. Additionally, using the balancing methods have a significant impact on enhancing prediction accuracy. Finally, the proposed models identify the student that has a low CGPA at early stage and providing early warnings to student at risk. The result of this study will help decision makers to provide the student with an additional guidance for deserving students. Moreover, this model will help preparatory year administration to explore the reasons behind failure and identify the courses that mostly affect the CGPA.

The main limitation of this work is that the dataset was collected from IAU automated system (SIS), thus some features such as student personality, parent education, and parent job were not considered in this study. As mentioned before, all the classification models were built by using the datasets collected from a single Saudi Arabian university (IAU). Therefore, the results are not generalizable. As future work, researchers should include data from several universities to further verify and validate these outcomes. Furthermore, there is a need to consider more factors that could affect student performance including student personality, family income, parent jobs, and parents' education level, and student eLearning activities such as student clicks, and student participation in discussion room that could be collected using the LMS (Blackboard) and questionnaire.

APPENDIX

	Features Category	Grades Percentage Used	Feature Name
ENGL104 Model (Dataset #2)	Admission	0%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age.
	Admission +Week 2	3%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, WP 1.
	Admission +Week 4	9%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, WP 1, WP 2, E-Learning.
	Admission +Week 5	10%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, WP 1, WP 2, E-Learning, Practice test.
	Admission +Week 6	20%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, WP 1, WP 2, E-Learning, Practice test, Midterm 1.
	Admission +Week 7	25%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, WP 1, WP 2, E-Learning, Practice test, Midterm 1, WP 3.
	Admission +Week 9	28%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, WP 1, WP 2, E-Learning, Practice test, Midterm 1, WP 3, WP 4.
	Admission +Week 10	38%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, WP 1, WP 2, E-Learning, Practice test, Midterm 1, WP 3, WP 4, Midterm 2.
	Admission +Week 12	40%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, WP 1, WP 2, E-Learning, Practice test, Midterm 1, WP 3, WP 4, Midterm 2, WP 5.
	Admission +Week 15	50%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, WP 1, WP 2, E-Learning, Practice test, Midterm 1, WP 3, WP 4, Midterm 2,WP5, Attendance, Participation.
ENGL113 Model (Dataset #3)	Admission	0%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age.
	Admission + GPA1	0%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, GPA 1.
	Admission + GPA1 + Week 4	10%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, GPA 1, WP 1, E-Learning.
	Admission + GPA1 + Week 6	20%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, GPA 1, WP 1, E-Learning, Midterm 1.
	Admission + GPA1 + Week 10	30%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, GPA 1, WP 1, E-Learning, Midterm 1, WP 2, Midterm 2.
	Admission + GPA1 + Week 14	50%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, GPA 1, WP 1, E-Learning, Midterm 1, WP 2, Midterm 2, WP 3, Speaking.
	Admission + GPA1 + Week 15	60%	SAT 1, SAT 2, School GPA, School Graduation time, Nationality, School type, Admit term, age, GPA 1, WP 1, E-Learning, Midterm 1, WP 2, Midterm 2, WP 3, Speaking, Attendance, Participation.

REFERENCES

[1] B. Rienties, H. K. Simonsen, and C. Herodotou, ''Defining the boundaries between artificial intelligence in education, computer-supported collaborative learning, educational data mining, and learning analytics: A need for coherence,'' *Frontiers Educ.*, vol. 5, p. 128, Jul. 2020, doi: 10.3389/feduc.2020.00128.

[2] N. M. Seel, *Encyclopedia of the Sciences of Learning*. New York, NY, USA: Springer, 2011.

[3] A. K. Veerasamy, D. D'Souza, M.-V. Apiola, M.-J. Laakso, and T. Salakoski, ''Using early assessment performance as early warning signs to identify at-risk Students in programming courses,'' in *Proc. IEEE Frontiers Educ. Conf. (FIE)*, Oct. 2020, pp. 1–9, doi: 10.1109/FIE44824.2020.9274277.

[4] G. Ramaswami, T. Susnjak, A. Mathrani, J. Lim, and P. Garcia, ''Using educational data mining techniques to increase the prediction accuracy of Student academic performance,'' *Inf. Learn. Sci.*, vol. 120, nos. 7–8, pp. 451–467, Jul. 2019, doi: 10.1108/ILS-03-2019-0017.

[5] *FAQ About Preparatory Year At IAU*. Accessed: Aug. 19, 2020. [Online]. Available: https://www.iau.edu.sa/en/administration/deanships/deanship-of-preparatory-year-and-supporting-studies/faq

[6] 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, doi: 10.1109/access.2021.3049446.

[7] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, ''Multi-split optimized bagging ensemble model selection for multi-class educational data mining,'' *Int. J. Speech Technol.*, vol. 50, no. 12, pp. 4506–4528, Dec. 2020, doi: 10.1007/s10489-020-01776-3.

[8] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, ''Evaluating the effectiveness of educational data mining techniques for early prediction of Students academic failure in introductory programming courses,'' *Comput. Hum. Behav.*, vol. 73, pp. 247–256, Aug. 2017, doi: 10.1016/j.chb.2017.01.047.

[9] E. Alyahyan and D. Dusteaor, ''Decision trees for very early prediction of Student's achievement,'' in *Proc. 2nd Int. Conf. Comput. Inf. Sci. (ICCIS)*, Oct. 2020, pp. 1–7, doi: 10.1109/ICCIS49240.2020.9257646.

[10] N. Alangari and R. Alturki, ''Predicting Students final GPA using 15 classification algorithms,'' *Romanian J. Inf. Sci. Technol.*, vol. 23, no. 3, pp. 238–249, 2020.

[11] A. I. Adekitan and O. Salau, ''The impact of engineering Students' performance in the first three years on their graduation result using educational data mining,'' *Heliyon*, vol. 5, no. 2, Feb. 2019, Art. no. e01250, doi: 10.1016/j.heliyon.2019.e01250.

[12] N. Putpuek, N. Rojanaprasert, K. Atchariyachanvanich, and T. Thamrongthanyawong, ''Comparative study of prediction models for final GPA score: A case study of Rajabhat Rajanagarindra University,'' in *Proc. IEEE/ACIS 17th Int. Conf. Comput. Inf. Sci.*, Jun. 2018, pp. 92–97, doi: 10.1109/ICIS.2018.8466475.

[13] P. D. Gil, S. D. C. Martins, S. Moro, and J. M. Costa, ''A data-driven approach to predict first-year Students' academic success in higher education institutions,'' *Educ. Inf. Technol.*, vol. 26, no. 2, pp. 2165–2190, Mar. 2021, doi: 10.1007/s10639-020-10346-6.

[14] L. Chen, T.-T. Huynh-Cam, and H. Le, ''Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university Students learning performance,'' *Algorithms*, vol. 14, no. 1, p. 318, 2021, doi: 10.3390/a14110318.

[15] 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, doi: 10.1016/j.compeleceng.2020.106903.

[16] R. Vidhya and G. Vadivu, ''Towards developing an ensemble based two-level Student classification model (ESCM) using advanced learning patterns and analytics,'' *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 7, pp. 7095–7105, 2020, doi: 10.1007/s12652-020-02375-3.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, '''Why should i trust you?': Explaining the predictions of any classifier,'' in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

[18] S. M. Lundberg and S. Lee, ''A unified approach to interpreting model predictions Scott,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[19] J. M. Perkel, ''Why Jupyter is data scientists computational notebook of choice,'' *Nature*, vol. 563, pp. 145–146, 2018, doi: 10.1038/d41586-018-07196-1.

[20] *Academic Plan for the Humanities Track* Accessed: Aug. 18, 2020. [Online]. Available: https://www.iau.edu.sa/en/node/15957

[21] *Sklearn.impute.IterativeImputer*. Accessed: Aug. 13, 2021. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ''SMOTE: Synthetic minority over-sampling technique,'' *J. Artif. Intell. Res.*, vol. 16, no. 28, pp. 321–357, Jun. 2006, doi: 10.1613/jair.953.

[23] G. E. Batista, R. C. Prati, and M. C. Monard, ''A study of the behavior of several methods for balancing machine learning training data,'' *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.

[24] A. K. Hamoud, A. S. Hashim, and W. A. Awadh, ''Predicting Student performance in higher education institutions using decision tree analysis,'' *Int. J. Interact. Multimedia Artif. Intell.*, vol. 5, no. 2, p. 26, 2018, doi: 10.9781/ijimai.2018.02.004.

[25] G. A. S. Santos, K. T. Belloze, L. Tarrataca, D. B. Haddad, A. L. Bordignon, and D. N. Brandao, ''EvolveDTree: Analyzing Student dropout in universities,'' in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Niterói, Brazil, Jul. 2020, pp. 173–178, doi: 10.1109/IWSSIP48289.2020.9145203.

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

[27] S. N. Latifah, R. Andreswari, and M. A. Hasibuan, ''Prediction analysis of Student specialization suitability using artificial neural network algorithm,'' in *Proc. Int. Conf. Sustain. Eng. Creative Comput. (ICSECC)*, Aug. 2019, pp. 355–359, doi: 10.1109/ICSECC.2019.8907173.

[28] Y. S. Alsalman, N. K. A. Halemah, E. S. AlNagi, and W. Salameh, ''Using decision tree and artificial neural network to predict Students academic performance,'' in *Proc. 10th Int. Conf. Inf. Commun. Syst. (ICICS)*, Jun. 2019, pp. 104–109, doi: 10.1109/IACS.2019.8809106.

[29] E. Heidari, M. A. Sobati, and S. Movahedirad, ''Accurate prediction of nanofluid viscosity using a multilayer perceptron artificial neural network (MLP-ANN),'' *Chemometric Intell. Lab. Syst.*, vol. 155, pp. 73–85, Jul. 2016, doi: 10.1016/j.chemolab.2016.03.031.

[30] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, ''A comprehensive survey on support vector machine classification: Applications, challenges and trends,'' *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[31] P. Kamal and S. Ahuja, ''An ensemble-based model for prediction of academic performance of Students in undergrad professional course,'' *J. Eng., Des. Technol.*, vol. 17, no. 4, pp. 769–781, Aug. 2019, doi: 10.1108/JEDT-11-2018-0204.

[32] J. Nalepa and M. Kawulok, ''Selecting training sets for support vector machines: A review,'' *Artif. Intell. Rev.*, vol. 52, pp. 857–900, Jan. 2018, doi: 10.1007/s10462-017-9611-1.

[33] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2012.

[34] C. Beaulac and J. S. Rosenthal, ''Predicting university Students academic success and major using random forests,'' *Res. Higher Educ.*, vol. 60, no. 7, pp. 1048–1064, Nov. 2019, doi: 10.1007/s11162-019-09546-y.

[35] G. Biau and E. Scornet, ''A random forest guided tour,'' *Test*, vol. 25, no. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.

[36] W. Nuankaew and J. Thongkam, ''Improving Student academic performance prediction models using feature selection,'' in *Proc. 17th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol. (ECTI-CON)*, 2020, pp. 392–395, doi: 10.1109/ECTI-CON49241.2020.9158286.

[37] N. Burkart and M. F. Huber, ''A survey on the explainability of supervised machine learning,'' *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, Jan. 2021.

[38] S. M. Hassan and M. S. Al-Razgan, ''Pre-university exams effect on Students GPA: A case study in IT department,'' *Proc. Comput. Sci.*, vol. 82, pp. 127–131, Jan. 2016, doi: 10.1016/j.procs.2016.04.018.

[39] 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, doi: 10.1109/ACCESS.2020.2981905.

[40] J. Willems, L. Coertjens, B. Tambuyzer, and V. Donche, ''Identifying science Students at risk in the first year of higher education: The incremental value of non-cognitive variables in predicting early academic achievement,'' *Eur. J. Psychol. Educ.*, vol. 34, no. 4, pp. 847–872, Oct. 2019, doi: 10.1007/s10212-018-0399-4.

[41] A. I. Adekitan and E. Noma-Osaghae, ''Data mining approach to predicting the performance of first year Student in a university using the admission requirements,'' *Educ. Inf. Technol.*, vol. 24, no. 2, pp. 1527–1543, 2019, doi: 10.1007/s10639-018-9839-7.

[42] I. Tomek, ''Tomek link: Two modifications of CNN,'' *IEEE Trans. Syst. Man Cybern.*, vol. SMC-6, no. 11, pp. 769–772, Nov. 1976.

SARAH ALWARTHAN received the B.S. degree in computer science from Dammam University, Dammam, Saudi Arabia, in 2012. She is currently pursuing the M.S. degree in computer science with the College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam.

She is currently a Teaching Assistant with the Computer Department, Deanship of the Preparatory Year and Supporting Studies, Imam Abdulrahman Bin Faisal University. Her research interests include artificial intelligence, data mining, and machine learning.

NIDA ASLAM received the Ph.D. degree from Middlesex University, London. She is currently an Assistant Professor with the College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University (IAU), Saudi Arabia. Her research interests include machine learning, data mining, image processing, computer vision, and specifically the application of AI in health.

IRFAN ULLAH KHAN received the Ph.D. degree from Middlesex University, London. He is currently an Assistant Professor with the College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University (IAU), Saudi Arabia. His research interests include machine learning, data mining, big data processing, image processing, computer vision, and specifically the application of AI in health.