Ghareeb, S, Hussain, A, Al-Jumeily, D, Khan, W, Al Jumeily, R, Baker, T, Al Shammaa, A and Khalaf, M

Evaluating student levelling based on machine learning model's performance

http://researchonline.ljmu.ac.uk/id/eprint/19538/

Article

Discover Internet of Things

Discover

Research

Evaluating student levelling based on machine learning model's performance

Shatha Ghareeb[1] · **Abir Jaafar Hussain**[1,2] · **Dhiya Al-Jumeily**[1] · **Wasiq Khan**[1] · **Rawaa Al-Jumeily**[1] · **Thar Baker**[2] · **Ahmed Al Shammaa**[3] · **Mohammed Khalaf**[4]

Received: 20 December 2021 / Accepted: 29 April 2022
Published online: 30 May 2022

Abstract

In this paper, a novel application of machine learning algorithms is presented for student levelling. In multicultural countries such as UAE, there are various education curriculums where the sector of private schools and quality assurance is supervising various private schools for many nationalities. As there are various education curriculums in United Arab Emirates, specifically Abu Dhabi, to meet expats' needs, there are different requirements for registration and success. In addition, there are different age groups for starting education in each curriculum. Every curriculum follows different education methods such as assessment techniques, reassessment rules, and exam boards. Currently, students who transfer to other curriculums are not correctly placed to their appropriate year group as a result of the start and end dates of each academic year as well as due to their date of birth, in which students who are either younger or older for that year group can create gaps in their learning and performance. In addition, pupils' academic journeys are not stored which create a gap for the schools to track their learning process. In this paper, we propose a computational framework applicable in multicultural countries such as United Arab Emirates in which multi-education systems are implemented. Machine Learning are used to provide the appropriate student' level aiding schools to provide a smooth transition when assigning students to their year groups and provide levelling and differentiation information of pupils for a smooth transition between one education curriculums to another, in which retrieval of their progress is possible. For classification and discriminant analysis of pupils levelling, three machine learning classifiers are utilised including random forest classifier, Artificial Neural Network, and combined classifiers. The simulation results indicated that the proposed machine learning classifiers generated effective performance in terms of accuracy.

Keywords Machine learning · Student levelling datasets · Receiver operating characteristic (ROC) · Performance evaluations · Accuracy · The area under curve (AUC) · School curriculum · Student grade prediction · Artificial intelligence

✉ Shatha Ghareeb, S.R.Ghareeb@2019.ljmu.ac.uk; Abir Jaafar Hussain, A.Hussain@ljmu.ac.uk; Dhiya Al-Jumeily, D.Aljumeily@ljmu.ac.uk; Wasiq Khan, W.Khan@ljmu.ac.uk; Rawaa Al-Jumeily, R.Aljumeily@2019.ljmu.ac.uk; Thar Baker, Tshamsa@sharjah.ac.ae; Ahmed Al Shammaa, alshammaa@sharjah.ac.ae; Mohammed Khalaf, M.i.khalaf@uoa.edu.iq | [1]Faculty of Engineering and Technology, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK. [2]College of Computing and Informatics, University of Sharjah, 27272 Sharjah, United Arab Emirates. [3]College of Engineering, University of Sharjah, 27272 Sharjah, United Arab Emirates. [4]Department of Computer Science, Al-Maarif University College, Ramadi 31001, Iraq.

Discover Internet of Things (2022) 2:3 | https://doi.org/10.1007/s43926-022-00023-0

1 Introduction

Pupil's levelling is a process that is based on age and lacks automation as well as a unified process amongst schools in UAE. In this case, based on the schools' curriculum, they could offer pass/fail or credit/no credit or better grades. To address the issue of family relocation, many countries such as UAE, Saudi Arabia, and Qatar offer a variety of school curriculums including British, American, Arabic to meet expat needs, yet this can create conflict when it comes to pupils levelling who are transferring curriculums. To determine student levelling, Machine Learnings (MLs) can offer an accurate levelling system for schools. Machine learning algorithms have been used in various fields such as health, financial time series prediction, genetics analysis, image analysis and online learning. Machine learning is transforming the way education is delivered and fundamentally changing the method of teaching, learning and research. Education institutes such as schools and colleges are using ML to identify struggling students earlier and develop action plans to improve success and retention [1]. ML is expanding the availability and effectiveness of online learning through localization, transcription, text-to-speech, and personalisation [2].

There are different methods of machine learning which include, supervised learning, unsupervised learning, deep learning, and reinforcement learning. Data preparation and training of a machine learning model is the foundation of the machine learning method, however, it is also vital to consider measuring the performance of the trained models. There are numerous machine learning methods used for classification which include, Random Forest Classifier (RFC), K-nearest Neighbours (KNN), Support Vector Machine (SVC), Naïve Bayes Classifier (NAIVEBC), Levenberg–Marquardt (LEVNN), and Voted Perception Classifier (VPC) [3].

This research aims to apply machine learning algorithms using datasets gathered from several education curricula to predict pupils' grades and levels when they transfer curricula [4]. To predict pupils' grades and levels, datasets with certain features have to be prepared. The datasets used in the study were classified into 3 classes based on student average level.

The reminder of this paper is organised as follows. Section 2 will discuss related work conducted in machine learning in education. Section 2 discusses the several types of machine learning methods such as supervised, unsupervised and reinforcement learning. Section 4 focuses on the method of supervised classifications, and how they have been implemented in this study. Section 5 mentions the model evaluation metrics that are required to quantify model performance. Section 6 of this paper discusses the steps of machine learning implementation and statistical tools. Finally, the paper is summarised with Sect. 7 which discusses what we have learnt from this and what methods and techniques we can implement in the future to improve the study.

2 Related work

In recent years, education sectors have faced many challenges to meet demands of e-Learning. The main motivation for researchers is to create a new system that is able to support education institutes in terms of student levelling and student transition between schools as well as performances [5]. There are a number of machine learning studies conducted on student levelling which is related to our research [6]. In this section we discuss influence gained from different studies and their limitations including online courses preferences, students levelling and students performances.

Masci et al. [7] developed and applied novel machine learning and statistical methods to analyse the reason for students PISA (Program for International Student Assessment) 2015 test scores in nine countries: Australia, Canada, France, Germany, Italy, Japan, Spain, UK and USA. The study was to find out which student characteristics are associated with test scores and which school characteristics are associated with school value-added (measured at school level). To address these issues, they applied a two-stage methodology using flexible tree-based methods. The researchers first run multilevel regression trees in the first stage to estimate school value-added. In the second stage, they relate the estimated school value-added to school-level variables using regression trees and boosting. Results show that while several student and school-level characteristics are significantly associated with students' achievements, there are marked differences across countries. This study has focussed on ways the school has impacted the student performance; however, it doesn't discuss the actual performance of the student just based on the input he/she is putting.

Shabandar et al. [8] researched Machine Learning approaches to predict learning outcomes in Massive Open Online Courses (MOOCs). The research is available within the area of MOOC data analysis, in particular considering the behavioural patterns of users. Based on learner behavioural patterns, two sets of features were compared in terms of their suitability for predicting the course outcome of learners participating in MOOCs. The research discovered that there is a strong correlation between click steam actions and successful learner outcomes. Various machine learning algorithms have been applied to enhance the accuracy of classifier models. Simulation results from the investigation showed that Random Forest achieved viable performance for the prediction problem, obtaining the highest performance of the models tested. This study has some relation to our study in terms of focusing on the student performance. However, the limitation in this study is that the researcher has not put emphasis on the past student performance and applying a prediction based on that which is discussed in this paper.

Hsia et al. [9] conducted a study using data mining techniques to analyse course preference and the completion rates of enrollees in extension education courses in a university-based in Taiwan. Using their collected data, the researcher aimed to improve the target curriculum based on the student needs. The algorithms that the researcher used were Decision Tree Algorithms, Link Analysis Algorithms, and Decision Forest Algorithms. The data collected from the university was in the range of 5 academic years from the year 2000 to 2005, overall, 1408 records were collected. After testing 8 different algorithms and studying their capabilities in classification, prediction, clustering and description, the researcher selected Decision tree, Link Analysis and Decision Forest. Three separate variables were selected taking into consideration the desired outcome of the study. The variables selected were course category, completion status, and enrolee profession. Based on the results gained from the study, the Extension Education Center at CTU can plan for future courses based on the needs of the students. Masci study is closely related to our study because the researcher has used some features related to student needs such as student learning outcomes, their capabilities in different subject areas. Although the study is built around a concept that is similar to our research, yet there is a gap in the study in which a potential to be fulfilled. There's further testing that could have been implemented which would show if a student is suitable for a specific course or not based on past collected data.

Lykourentzou et al. [10] researched the dropout prediction method for e-learning courses based on three popular machine learning techniques and detailed student data. The machine learning techniques used in the study are feed-forward neural networks, support vector machines and probabilistic ensemble simplified fuzzy ARTMAP. Since using a single algorithm may not provide accurate results, three different machine learning techniques were used to predict student dropout for e-learning courses. The method was examined in terms of overall accuracy, sensitivity and precision and its results were found to be significantly better than those reported in the relevant literature. This research provides vital information for the education sector which the dropout prediction, however, they focused on e-learning courses and not on face-to-face education in universities. Using the method conducted in this paper, the student level is predicted based on past and current exam marks in school, using those predictions we can predict if the student is going to be successful or not for different courses just based on his past performance.

In multicultural countries where many expats are found, there are several education systems available to accommodate parents and pupils needs. Switching from one system to another has become a critical issue as differences among systems could create gaps in education levels for students. Hence, it is extremely difficult to assign the right level for students when moving to a new curriculum. The challenge is to provide a smooth transition with minimum effects on pupils' performance. The age groups for each year group varies amongst curriculums. Nevertheless, there is a huge ambiguity in this area, which usually creates conflicts between parents, school, and the Ministry of Education which usually happens because different schools follow different curricula and therefore there's no common levelling guide amongst those schools. Hence machine learning model can provide method for automating the levelling process for students when changing between education systems.

3 Classification algorithms

This section of the study discusses machine learning algorithms and statistical tools. This section takes into consideration information about ML and the deployment of extracting useful information from the student and school data. In this section, we discuss supervised, unsupervised and reinforcement learning.

3.1 Supervised learning

Most practical machine learning uses supervised learning. Supervised learning is where we have an input variable known as (x) and output variable (y), and then we use an algorithm to learn the mapping function from the input to the output [11].

The purpose of supervised learning is to make an approximate of the mapping function close to accurate so that when we introduce a new input data (x), we can predict the output variable (y) for that new data inserted. Supervised learning is anticipated in finding the patterns in the data which can then be applied to an analytics process [11]. The main aim of the training set is to learn from labelled instances in the training set to identify unlabelled instances during the testing phase with high potential accuracy as shown in Fig. 1.

3.2 Unsupervised learning

Unsupervised learning is where we only have input data (X) and without any corresponding output variables. With unsupervised learning there is no correct answer or teacher, algorithms are left alone to discover and present the fascinating structure in the data.

Unsupervised learning can be grouped into two problematic scenarios:

- Clustering: clustering problem is where we want to the inherent groupings in the data, for example using purchasing behaviour of consumers to put them in groups.
- Association: An association rule learning problem is when we intend to find rules that describe large portions of our data.

Unsupervised learning is used when there is a use of large amounts of unlabeled data. An example of large amounts of unlabeled data is social media applications such as Twitter, Instagram and Snapchat. For us to understand the meaning behind this unlabeled data, we require an algorithm that classifies that data based on the patterns or clusters the system finds. Cluster analysis is the most common method in unsupervised learning that is used for exploratory analysis to find groupings or hidden patterns in datasets [12]. The main purpose of applying this kind of method is to discover the smallest group feature subset (clustering) from the datasets according to the chosen criteria [12]. Unsupervised learning conducts a repetitive process of analysing data without the use of human interaction. Unsupervised learning is used in email spam-detecting technology. For analysts to tag unsolicited bulk emails will be complicated because there are too many variables in legitimate and spam emails, therefore, with the use of applied clustering and association, machine learning is able to identify unwanted emails. Figure 2 shows the unsupervised learning workflow.

3.3 Reinforcement learning

Reinforcement learning (RL) is a behavioural learning model where the system learns in an interactive environment by the use of trial and error [13]. The algorithm uses feedback from own actions and experiences, in other words receiving feedback from the data analysis to guide the user to the best outcome. As the system learns through trial and error, having a sequence of successful decisions will help the process be reinforced [13].

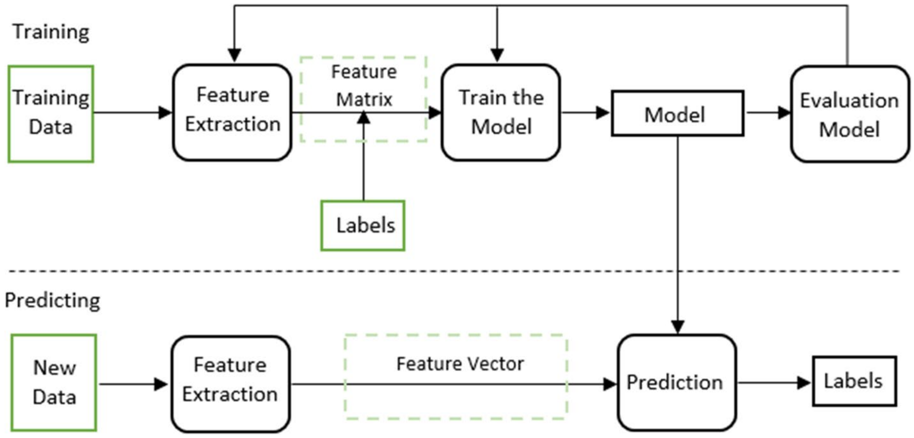

Fig. 1 Supervised learning workflow

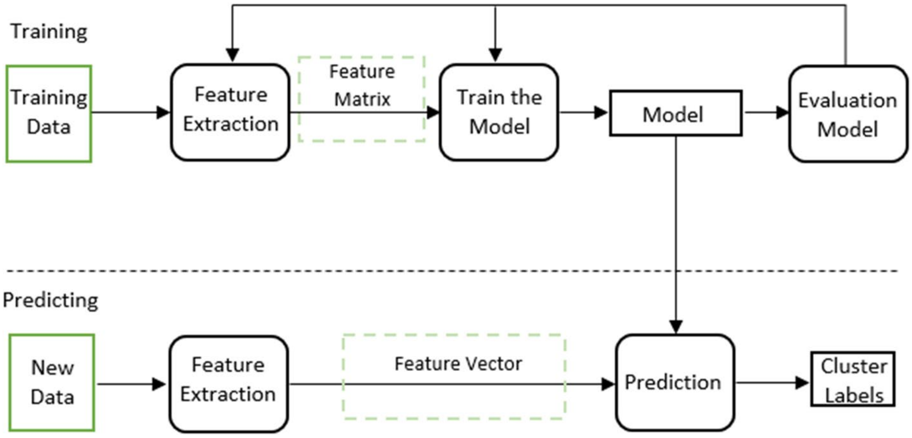

Fig. 2 Unsupervised learning workflow

4 Classification algorithms

This research uses supervised classification because the datasets collected from the schools have been identified with relevant labels. Regression models are aimed at mapping instance (input) values to continue outcome values, while classification procedures aim to map instances (input) into discreet classes. For example, some studies aim to classify students that will drop out of their exams or otherwise won't. Within classification, the aim is to learn a decision-making platform that can correctly map an instance (input) space to an output. Within the education sector, ML researchers have researched ways to implement algorithms such as RFC, SVM, and ANN to predict student performance and requirements. The classification of student and education data has shown positive outcomes for student benefits and education institutes.

In terms of the classification process, object (x) is the input that has a set of features, while (y) is the class label that's assigned alongside x. The classification model is implemented to predict the class label for new samples. Classification techniques in education are important to improve student levelling and enhance education institute decision making in terms of grading and student abilities. There are various methods implemented for classification which are grouped into two, linear and nonlinear classifiers. As represented in Eq. (1), the linear classifier is represented as a linear function (g) of the input (x). From the equation, (w) represents a set of weights, while (T) is the metrics response and (b) refers to the bias. Having two sets of classes k_1 and k_2, therefore input vector x gets assigned to k_1 when $(x) \geq 0$. The boundary among k_1 and k_2 is linear.

$$g(x) = w^T x + b. \tag{1}$$

4.1 Random forest classifier (RFC)

Random forest classifier (RFC) is one of the most successful ensemble learning techniques which has been proven to be an effective technique in pattern recognition and ML for high-dimensional classification and skewed problems. RFC does have a drawback which is tree classifiers have high variance. In practice, it is common for a slight change in the training dataset to be in a different tree. The reason behind this is the hierarchical nature of the tree classifiers.

Equations (2) and (3) describe the method of RFC.

$$f(x) = \frac{1}{m}\sum\nolimits_{i=1}^{m} f\left(x, x_i p\right). \tag{2}$$

In Eq. (2), the variable that has a partial dependence is referred to as x and $x_i p$ takes into consideration any other variable, while other variables in the equation refer to feature from student data.

$$f(x) = \log t_j - \frac{1}{J}\sum\nolimits_{k=1}^{J} \left(\log t_k(\mathrm{y})\right). \tag{3}$$

The number of classes is referred to as J in the equation while j is the class. The proportion for the total votes for the class is t_k.

Training of the RFC can be done by developing several B trees. The training sets for the classifier is $X = \mathrm{x}_1 \ldots \mathrm{x}_\mathrm{n}$ and the classes are $Y = \mathrm{y}_1 \ldots \mathrm{y}_\mathrm{n}$. When replacing B in $b = 1, \ldots, B$: the training samples from X, Y have its place of B, which is then referred to as Xb, Yb, where Y comes from the predicted class. To implement the classifier and select numerous training data sets, $M = \{(\mathrm{X}_1, (\mathrm{X}_\mathrm{n}) \ldots, (\mathrm{Y}_1, \mathrm{Y}_\mathrm{n})$, where $X_i, i = 1.., n$ is the descriptors vectors and depending on the classification outcome, Y_i can be referred to as the activity of interest or corresponding label.

4.2 Artificial neural network (ANN)

Artificial neural network (ANN) is an established numerical computerized system that can model complex relationships between random experimental inputs and their relative outputs [14, 15]. An artificial neural network is based on biological neural networks, which consist of neurons and synapses between them. During learning, each synapse adds weights to the received signal. Subsystems are collections of neurons, and the brain is formed by subsystems. The neurons in an ANN are weighted, and the weights of each neuron are corrected as the ANN learns. A neural network is similar to a brain because the layers of a neural network act as subsystems and neurons play the same role in both networks. ANN generally contains three layers, the input layer, hidden layer, and output layer that output data are recognized to the network [16, 17]. There are different types of ANN models, but for nonlinear functions multilayer perception (MLP) NN are more widely used. MLP neural network is a feed-forward ANN and constructed with three layers an input layer, an output layer and hidden layers. The activation function in the output layer at MLP ANN must be linear and hidden layers can be linear or nonlinear. Most MLP ANN models use a sigmoid function, Tansig or Logsig.

4.3 Combined classifiers

Machine learning uses a training set combined with building a classifier that provides a reliable classification. This study used the multi-class classification problem where many classes are available in the datasets. This research combines multiple classifiers to improve the classification accuracy and performance in comparison to the single model. Study shows that combining different classifiers can generate better output [18]. The total information from multiple models s combined to generate better decisions. Using the bootstrap method, the training set is delivered to each model. Each model generates an outcome using the performance metrics method. To discover the classifiers that generate the highest performance, voting was used to select the classifier with the highest accuracy and performance. This study has taken the approach of combing the final classification results gained using N different feature sets $\left(f_i{}^1(x), \ldots, f_i{}^N(x)\right)$. To construct both classifiers, the training models need to be trained with the use of different feature sets. Where x refers to particular input and each model m^n generates its own output $y^n = (y^n(1), \ldots y^n(Z)$, where z is considered the class label, while $y^n(m)$ corresponds to the probability of c^n. Each classifier i generates L approximations to the probabilities $f_j{}^N(x), j = 1, \ldots, L$. The purpose of combining multiple algorithms is to create better results. The research used the stacked and voting method. The stacked method involves a set of models that lead to the same space for them to be combined. Each classifier receives training from the same training set which receives 70% of the datasets, the validation receives 10%, while testing receives 20%.

5 Model evaluation metrics

Model evaluation metrics are required to quantify model performance. The choice of our evaluation was dependent on the task our model is required to perform.

5.1 Predictive models

When discussing about predictive models, two models which are regression model (continues output) and classification model (nominal or binary output) are considered. The evaluation metrics that are used in each model varies.

In our case with regards to classification, two types of algorithms can be used:

- Class output: includes algorithms like SVM and KNN.

- Probability output: includes algorithms like Logistics Regression, Random Forest etc.
- Confusion Matrix involves:
- True positives (TP): Predicted positive and are positive.
- False positives (FP): Predicted positive and are negative.
- True negatives (TN): Predicted negative and are negative.
- False negatives (FN): Predicted negative and are positive.
- Sensitivity is the percentage of positive instances out of the *total actual positive* instances.

$$\frac{TP}{TP + FN}$$

- Specificity is the percentage of negative instances out of the *total actual negative* instances.

$$\frac{TN}{TN + FP}$$

Precision is the percentage of positive instances out of the *total predicted positive* instances.

$$\frac{TP}{TP + FP}$$

- F-Measure is a mean of precision and recall, therefore the higher F1 score received the better the accuracy

$$\frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \times precision \times recall}{precision + recall}$$

- J score measures the ratio of True Negatives against the total predicted negatives (whether true or false ones).

$$\frac{TN}{TN + FN}$$

- Accuracy is the most commonly used metric to judge a model, however, accuracy is not a clear indicator of the performance of the model, specifically when the classes are imbalanced.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

- The Receiver Operator Characteristics (ROC) curve is an evaluation metric for binary classification problems. It is the probability curve that plots the TPR against FPR at various threshold values and essentially separates the signal from the noise.
- The area under ROC Curve is a performance metric for measuring the ability of a binary classifier to discriminate between positive and negative classes.

6 Machine learning implementation and statistical tools

Machine learning models have been utilised to classify student academic levels and e-learning. Using machine learning models, schools can predict students' current levels and predict their academic path in the future. The main benefit of this model is to make use of recent technological development in ML algorithms to assist schools and teachers to develop their students more effectively with the use of past records. The aim is to propose the prediction of the student level using classification methods based on the data sets collected previously from schools.

The proposed implementation comprises of many steps; data collection, data pre-processing, and then the data differentiates into three parts which are, building the model based on the training data, evaluating the model based on the testing sets, and then selecting the relevant model. The data collection process starts when the student is first registered into the school. To apply the ML model, the collected data must be cleaned (removing unwanted data and filling missing data), thereafter, various ML models are selected to evaluate the data sets. The holdout method is used to split the datasets into training, validation, and testing as shown in Fig. 3.

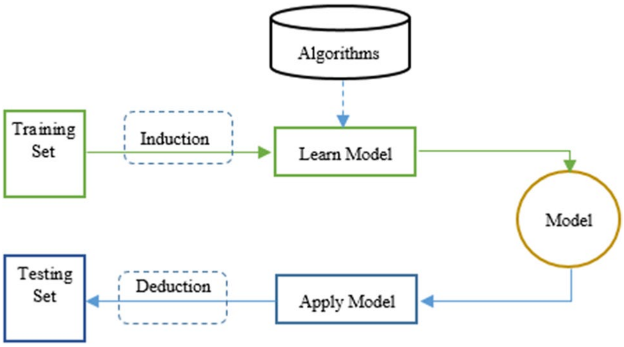

Fig. 3 The methodology process

6.1 Data collection

The dataset presented comprising students' records for two academic years that included math, English, and science for 3 terms [4]. The selection of subject areas and some terms were based on influence from other researchers in a similar subject matter [8, 19]. The dataset comprises novel aspects specifically, in terms of student grading in diverse educational cultures within multiple countries—Researchers and other education sectors will be able to see the impact of having varied curriculums in a country. Furthermore, Dataset compares different levelling cases when students transfer from curriculum to curriculum and the unreliable levelling criteria currently set by the international schools. Figure 4 demonstrates the data collection process used in this study. Table 1 shows the distribution of records collected for the research from the schools among the three classes.

6.2 Data preparation

Data preparation is an essential part when handling data which helps to check for correctness, meaningfulness, and the security of data to be used. Therefore, the data was prepared by unifying the outputs, filling in missing data, eliminating non-related features, and preparing headers. The data was cleaned using excel filtering, equalising the instances according to their categories. The date of birth for the students was converted to age for categorisation as per 2017. New attributes of age groups were created according to the age of the student and the standard age of the class into three categories as explained in Table 2. A new attribute is then created which is the average marks for the 2019/2020 academic year for three subjects over 3 terms. Then an additional attribute is created called class, which divides the average marks of the students for the academic year 2019/2020 for 3 subjects over 3 terms. Three classes are created, Class 1 is the average marks from 85 to 100%, class 2 is the average marks from 75% to 84.99%, and class 3 is the average marks below 75%. Any data that was deemed confidential is removed such as student name and ID. The selection of a suitable classifier involves trial-and-error processes, but the use of statistical validation can guide that process [20].

6.3 Feature selection

Feature selection is one of the essential pre-processing steps in data mining [21]. The feature selection is to select a subset by disregarding irrelevant features and unwanted information from the student levelling dataset. The feature extraction technique can help generate accurate results and remove any negative influence on the learning models [22]. It is an effective dimensionality reduction technique to remove noise features. In general, the basic idea of a feature selection algorithm is to search through all possible combinations of attributes in the data to find which subset of features works best for prediction. Thus, the attribute vectors can be reduced by which the most meaningful ones are kept, and the irrelevant or redundant ones are removed and deleted [21].

Feature Subset selection has two approaches: Filter and Rapper. The filter approach applies data with an examining property, in general, to calculate the goodness of the feature subset except for a learning algorithm that evaluates the quality of the feature subsets [23]. Using this technique in the research, unnecessary features were reduced. Isolating irrelevant features helped improve the performance of the models and the results generated from the dataset. Over-fitting can harm the performance of the models; therefore, this technique can reduce those risks [24]. Feature selection decreases the search space determined throughout all the features, and consequently, the models can

Fig. 4 Data collection process

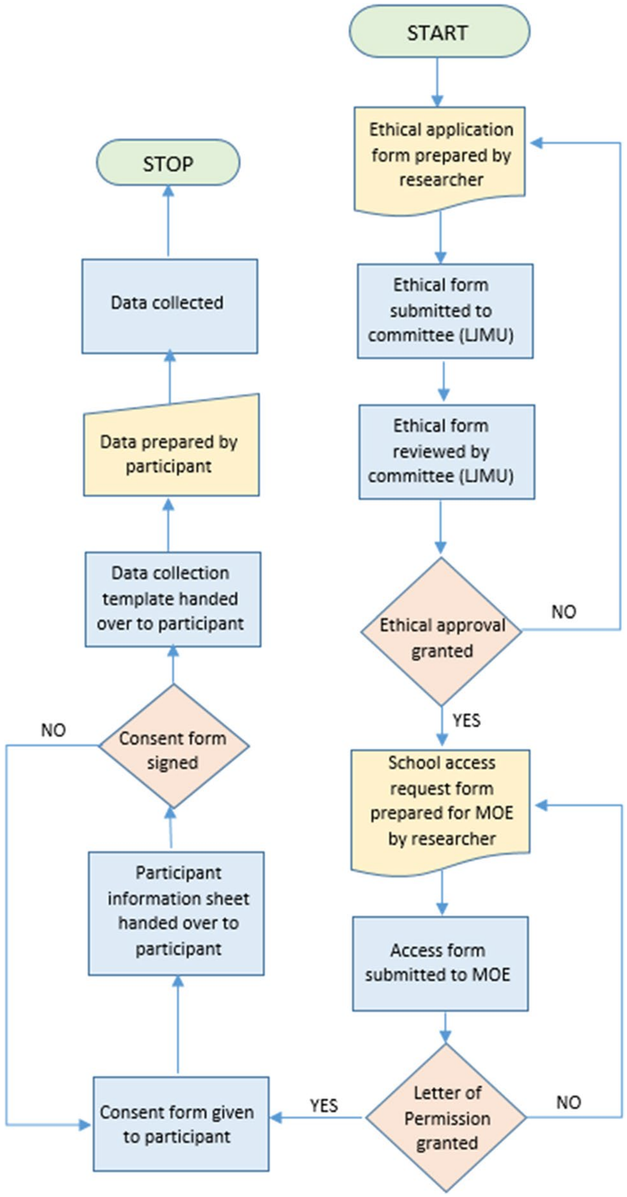
START
Ethical application form prepared by researcher
Ethical form submitted to committee (LJMU)
Ethical form reviewed by committee (LJMU)
Ethical approval granted
NO
YES
School access request form prepared for MOE by researcher
Access form submitted to MOE
Letter of Permission granted
NO
YES
Consent form given to participant
Participant information sheet handed over to participant
Consent form signed
NO
Data collection template handed over to participant
Data prepared by participant
Data collected
STOP

Table 1 Sample of student dataset [4]

Type	Number	Description
Records	1550	Data collection for 2 academic years
Classes	3	Class 1 is 85% ≤ 100% Class 2 is 75% ≤ 84.99% Class 3 is < 75%
Attributes	25	Gender, age, proposed year/grade, year of admission, etc

process the data faster with less memory consumption [5, 26]. The irrelevant and redundant features can confuse the learning when dealing with a few training examples, leading to overfitting and high dimensionality [25]. The high dimensionality of the extracted features will be reduced using feature selection methods [27]. This method is achieved by identifying spaces with lower dimensions than the actual data.

The process of feature selection is divided into two forms [28]:

Table 2 Collected data description

Attribute name	Value	Description
Input attributes		
Gender	Male/Female represented by 0 and 1	Gender of the student
Student age (as of 2017/18)	6, 7, 8, 9,…..etc	Age of the student calculated from 2017/18 academic year
Age group	− 1, 0, 1	Based on the age of the student as per 2017; -1 group means 1 year below accepted age; 0 group is on the accepted age; and 1 group is 1 year above accepted age
Year of admission	Old BBS Student, Old GEMS Student, New Admission 18/19	The data collected is for two or more academic years; 2017/18 and before academic years + 2018/19 academic year
Current year (17/18)	Foundation Stage 1, Foundation Stage 2, Year 1–12/Grade 1–11	This is the year or grade group assigned to the student by the school
Proposed year/grade (18/19)	Foundation Stage 1, Foundation 2, Year 1–13/Grade 1–12	This is the year or grade group assigned to the student by the school
Year of admission	Old BBS Student, Old GEMS Student, New Admission 18/19	The data collected is for two or more academic years; 2017/18 and before academic years + 2018/19 academic year
Previous school (17/18)	Many schools in UAE	Previous schools that the student was in before this study
Previous curriculum	British/American/MOE/Canadian/Indian/Australian/CBSE/German	The curriculum the student transferred from
Current school	GEMS, Belvedere	Name of the school that the data has been collected
Current curriculum	0, 1	The curriculum the student transferred to, 0 = American, 1 = British
Math-exam	Mark out of 100	Exam marks for school entry exam in math
Science-exam	Mark out of 100	Exam marks for school entry exam science
English-exam	Mark out of 100	Exam marks for school entry exam English
Math19-1	Percentage out of 100%	Term 1 student Maths Exam marks during academic year 2018/19
Science19-1	Percentage out of 100%	Term 1 student science Exam marks during academic year 2018/19
English19-1	Percentage out of 100%	Term 1 student English Exam marks during academic year 2018/19
Math19-2	Percentage out of 100%	Term 2 student Maths Exam marks during academic year 2018/19
Science19-2	Percentage out of 100%	Term 2 student science Exam marks during academic year 2018/19
English19-2	Percentage out of 100%	Term 2 student English Exam marks during academic year 2018/19
Math19-3	Percentage out of 100%	Term 3 student Maths Exam marks during academic year 2018/19
Science19-3	Percentage out of 100%	Term 3 student science Exam marks during academic year 2018/19
English19-3	Percentage out of 100%	Term 3 student English Exam marks during academic year 2018/19
Math20-1	Percentage out of 100%	Term 1 student Maths Exam marks during academic year 2019/20
Science20-1	Percentage out of 100%	Term 1 student science Exam marks during academic year 2019/20
English20-1	Percentage out of 100%	Term 1 student English Exam marks during academic year 2019/20
Math20-2	Percentage out of 100%	Term 2 student Maths Exam marks during academic year 2019/20
Science20-2	Percentage out of 100%	Term 2 student science Exam marks during academic year 2019/20
English20-2	Percentage out of 100%	Term 2 student English Exam marks during academic year 2019/20
Math20-3	Percentage out of 100%	Term 3 student Maths Exam marks during academic year 2019/20
Science20-3	Percentage out of 100%	Term 3 student science Exam marks during academic year 2019/20
English20-3	Percentage out of 100%	Term 3 student English Exam marks during academic year 2019/20

Table 2 (continued)

Attribute name	Value	Description
Output attributes (only class as it is a result of a formula applied to the average attributes)		
Average 19/20	Percentage out of 100%	The average of 2019/2020 academic year extracted from the marks of the 3 subjects among the 3 terms
Class	1, 2, and 3	3 classes extracted from the Average value; class 1 from 85 to 100%; class 2 from 75% to 84.99%; and class 3 is below 75%

i. Feature transformations—dealing with lower dimensional space such as independent component analysis (ICA) and principal component analysis (PCA).
ii. Select some features for a given pattern based on the mean or standard deviation of the feature values.

Feature selection can reduce both the data and the computational complexity [29]. Various feature selection methods are available such as Information Gain (IG), Symmetric Uncertainty (SU) and Correlation-based feature selector (Cfs).

In this research, names and student IDs should be removed from the dataset to avoid disclosing students' personal information as per Ethical Approval. Then, the date of birth needs to be converted to age for categorisation. A new attribute of age group having three categories: − 1, 0, and 1, was created according to the student's age and legal age of the class the student was assigned to. This process will allow the Age Group feature to be used instead of the Date of Birth, so Date of Birth and Age features will be excluded from this research.

The Previous School attribute has been removed because it has the same effect as the Previous Curriculum attribute. Nevertheless, the Previous Curriculum attribute will be replaced by three columns as below:

1. PrevCurrUS: 1 for the US curriculum, and all other curricula are represented by 0.
2. PrevCurrUK: 1 for the UK curriculum, and all other curricula are represented by 0.
3. PrevCurrOther: 1 for all curricula except the US and UK (e.g., Asia, UAE, Pakistan, Iraq etc.), and the US and UK curricula would be 0.

Hint: There should be only one value of "1" among the three features in all cases, and the remaining attributes must remain zeroes. In other words, if a student has 0 0 1 respectively, they had neither British nor American previous curriculum. While having 1 0 0 will indicate that they had an American curriculum. And so, if they have 0 1 0, they were in a British curriculum.

In addition, the Average attribute has been created using all the students' marks for the 2019/2020 academic year. Then the class label feature has been built called "Class", which divides the average into three categories: 1, 2, and 3. Class one from 85 and 100%, class two from 75% to 84.99%, and class three is below 75%. Finally, the attributes used to create other features are excluded from the dataset, such as 2019/2020 academic year marks and their average. As a result, ML algorithms will be applied on 24 attributes, as shown in Table 3.

Cfs Subset Evaluator and the best-first search method have been used in this study to get the final feature set because they give good results for the students' performance dataset. Thus, Cfs Subset Evaluator and best-first search are applied as the feature selection algorithm in this research as shown in Fig. 5 [23]. Tables 3 and 4 represents the features selected using the Cfs Subset Evaluator and best-first search.

6.4 Experimental procedure

The experimental part of the research covers the design of the test environment plus the formation of each model. In this research, the holdout method is used to evaluate how statistical analysis can transform into a dataset. The datasets are split into training, validation, and testing sets. Using this method allows us to find the average percentage of correct and incorrect classifications, therefore the training set receives 70%, validation 10%, and testing is 20%.

The algorithms used in this study is comprised of Random Forest Classifier (RFC), Artificial Neural Network (ANN), and combined classifiers. The main purpose of selecting those models is because they can deal with high-performance classifiers whilst also dealing with strong non-linear classifiers.

To obtain performance results for each algorithm used, the simulation was repeated 50 times and then the mean was calculated. A description of the models used in this study is described in Table 5.

6.5 Random forest classifier (RFC)

The performance evaluation techniques were achieved using the collected student levelling dataset of 1550 samples. The imperial study in RFC was performed using random forest models. The classification performance was evaluated using the evaluation metrics discussed in previous chapters. During simulation, both training and testing datasets were selected randomly whilst repeating in every test run. The RFC models were applied to the 16 features (including the class label). The results gained from this experiment produced reasonable values, as shown in Table 6. Training and testing

Table 3 Updated features

#	Features selected
1	Gender
2	Age group
3	Current class
4	Proposed class
5	Year of admission
6	Prev_US_curr
7	Prev_UK_curr
8	Prev_Other_curr
9	Current school
10	Current curriculum
11	Previous system
12	Math-exam
13	Science-exam
14	English-exam
15	Math19-1
16	Science19-1
17	English19-1
18	Math19-2
19	Science19-2
20	English19-2
21	Math19-3
22	Science19-3
23	English19-3
24	Class Label

stages were applied to the dataset and measured by the performance measurements as shown in the below tables. The proposed method's activity has also been evaluated in visual performance evaluation with ROC and AUC charts, as shown in Figs. 6 and 7.

The RFC model was built and trained using multiple trees prepared using the student levelling dataset; 50, 100, 200, 400, 500 and 1000 trees. RFC100 had generated the best Accuracy result during the training process with an average of 0.69 for the three classes, but the AUC average was 0.68, the third-highest result. Although the AUC was higher using the RFC50 method (0.72), it had a lower 0.66 Accuracy than the RFC100. The Sensitivity of RFC200 performed better than the other methods with 0.661. As shown in Fig. 6, the RFC50 demonstrates the best results of the ROC curve among all other models.

Random forest classification combines the basics of decision trees with additional flexible parameters, which increases the model's Accuracy [30]. The bootstrapped method allows multiple times a selection of critical samples. Once the bootstrapped datasets are made, a random subset of variables is used to develop the random forest [31]. Fifteen input features and the class label from the students' dataset are considered for every step. Viewing the subset of variables for each step, a new bootstrapped dataset is developed alongside several RF trees, and this process was completed and repeated several times. Once the student data is ingested into all the trees in RF, a calculation is done to discover which model received the highest number of votes.

In the training stage of the model, RFC50 had the highest results among all the RFC models, for classes 2 and 3, as shown in Fig. 7. However, the average of classes produced a specificity of 0.674, Precision 0.466, F1 score of 0.473, j score 0.306, and AUC 0.72. The AUC for RFC100 and RFC200 are 0.68 and 0.69, respectively, while RFC1000 has the lowest AUC performance.

The RFC models' performance during the testing phase is slightly lower than the results during the training stage. As shown in Table 7, the RFC50 generated the highest AUC with an average of 0.587, whereas RFC100 generated the highest Accuracy of 0.634. On the other hand, RFC200 generated the lowest AUC compared to all other RFC models. RFC50 yielded the most heightened Sensitivity of 0.709 too, whilst RFC500 scored 0.621 to be the second-highest one in terms of Sensitivity. RFC50 had the second-lowest specificity of 0.483, whilst RFC100 generated the highest specificity of 0.656.

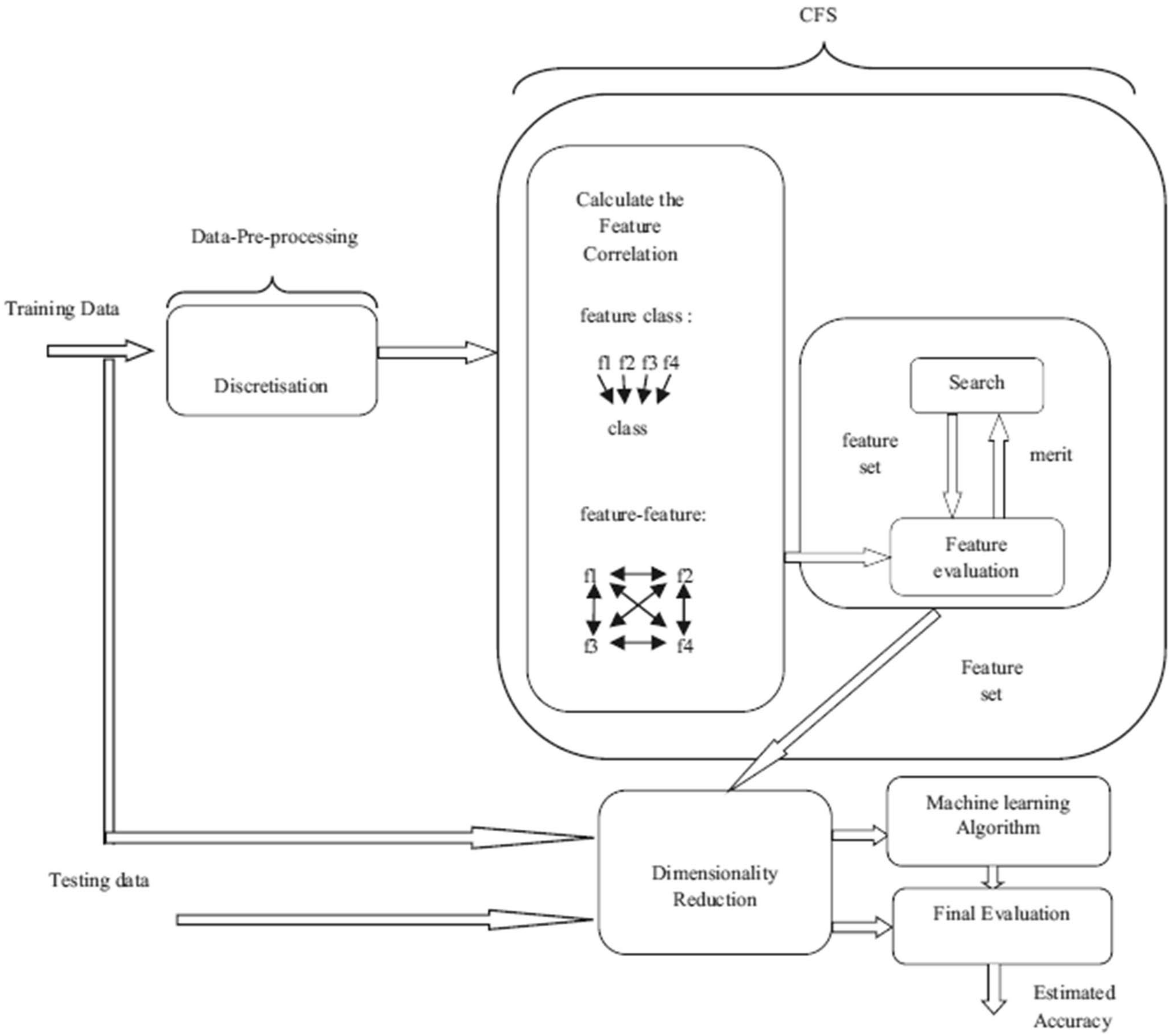

Fig. 5 Correlation-based feature selector (CFS)

As shown in Fig. 8, the RFC50 has given the best performance of ROC in testing and training stages as per the AUC readings. Figures 6 and 8 show the visual representations of the ROC curve for training and testing stages, respectively.

As implied from Fig. 9, the models' 1st class performance fluctuated highly. In contrast, 2nd and 3rd Classes had shown closer results amongst these models.

6.6 Artificial neural network (ANN)

The ANN can perform multiple classifications, clustering, and dimensionality reduction [32]. The purpose of implementing NN is to measure the network with generalisation capability using the weigh unit throughout the construction of the NN network and compare the model's performance with other classifiers.

The evaluation of the model is achieved using the holdout method implemented on 1550 samples. The model received three sets of datasets: 70% for training, 20% for testing, and 10% for validation. This research utilised two methods of NN models: Levenberg Neural Network (LEVNN) and Backpropagation Network Classifier (BPXNC). Although LEVNN obtained a higher accuracy average than BPXNC, they both generated the same AUC value of 0.962, as shown in Table 8. Overall,

Table 4 Features selected using the Cfs subset evaluator and best first search

#	Features selected
1	Gender
2	Age group
3	Current class
4	Proposed class
5	Year of admission
6	Prev_UK
7	Prev_Other Curric
8	Current school
9	Current curriculum
10	Previous system
11	Math-exam
12	English-exam
13	English19-2
14	Math19-3
15	English19-3

all the parameters have shown that LEVNN caused better results than BPXNC. The ROC curve visualisations for LEVNN and BPXNC are shown in Figs. 10 and 11.

Despite receiving high accuracy results during training for both LEVNN and BPXNC, the testing results performance for both NN models was slightly lower than the training process. For LEVNN, the hidden layers were amended using backpropagation links, whereas BPXNC were altered using feedback from the output layer. Table 9 shows the classification performance evaluation for LEVNN and BPXNC during testing. Although BPXNC demonstrates the slightly better performance of Precision (0.571) and AUC (0.767) than the LEVNN model, generally, LEVNN has produced higher scores for most of the parameters with 0.801, 0.815, 0.602, 0.616, and 0.816 for the Sensitivity, Specificity, F1, J, and Accuracy, respectively. Figure 12 shows the ROC curve illustration of LEVNN and BPXNC, implying that these models perform well. Figure 13 shows the AUC of an average of the three classes for LEVNN and BPXNC.

6.7 Combined classifiers

A combined classifier is a method that combines two or more classifiers to improve the performance and accuracy of the models [33]. This model of machine learning algorithms is found to generate good results when combining specific classifiers. Implementing a combined classification in this study allowed the models to learn from non-linear components and yield good results. The combined classifiers use a pattern recognition system combined with a bootstrap aggregating approach to enhance the selected models [34].

The proposed study implemented several classifiers combined which included NN Com, LEVNN Com, and NN and RF Com1 and Com2. The classification performance evaluation method is based on 16 elements and three classes of student leveling datasets. Table 10 shows the outcome for combined classifiers in training. The prediction of the student levelling dataset during the training phase is achieved by taking the majority vote for RFC for several cycles. Alternatively, NN generated the predictions after several processes by taking a weighted vote.

The overall AUC for all the combined classifiers was 1 for training as shown in Table 10, which is impressive results when comparing them to ANN and RFC classifiers. Although the precision for NN Com and LEVNN Com was 1 compared to 0.999 for NN combined with RF, yet still the overall performance for NN combined with RF was performing better. A strong combination between NN and RF is highlighted, which indicates that there's vital information is available in our selected datasets. NN Com and LEVNN Com were both capable of handling the datasets showing accuracy, ROC and AUC of 1 as shown in Figs. 14 and 15.

During testing of combined classifiers, the performance of the algorithm was lower than that of training as shown in Table 11. NN and RF obtained a sensitivity of 0.755, specificity 0.798, precision 0.565, F1 0.585, J score 0.552, accuracy 0.799, and AUC 0.824. We have stated previously that NN combined with RF performed excellently during training generating the highest performance, yet still, during testing, it didn't perform as well but it generated the highest AUC as

Table 5 Description of the used models

Model	Description	Architecture	Training algorithm	Parameters	Role
RFC	Random forest	14 inputs, 50, 3 outputs	Random feature bagging	Number of generated decision trees 50; Size of feature: 1	Non-linear comparison model
BPXNC	The feed-forward neural network algorithm	Units: one context unit for each output unit	They are Trained with mapping a set of input data to generate a computational modification for the whole weights	Momentum coefficients between 0.01 and 1.0 Sigmoid function $f(x) = 1/(1 + e^{-x})$ Learning rate between 0.25 and 0.9 Performance: 0.0932	Non-linear comparison model
LEVNN	Multilayer perception, Levenberg–Marquardt algorithm	Units 29–2-4 transit activations	Levenberg–Marquardt	Initialisation: Nguyen Windrow Adaptive learning rate settings: initial value: 0.001 coefficient for increasing LR: 10 coefficients for decreasing LR: 0.1 maximum learning rate: 1e10	Non-linear comparison model

Table 6 RFC performance with average of 3 classes (training)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
RFC50/1	0.632	0.674	0.466	0.473	0.306	0.66	0.72
RFC100/2	0.592	0.673	0.434	0.460	0.265	0.69	0.68
RFC200/3	0.661	0.617	0.416	0.443	0.278	0.63	0.69
RFC400/4	0.524	0.657	0.405	0.416	0.181	0.65	0.61
RFC500/5	0.579	0.585	0.371	0.404	0.163	0.61	0.61

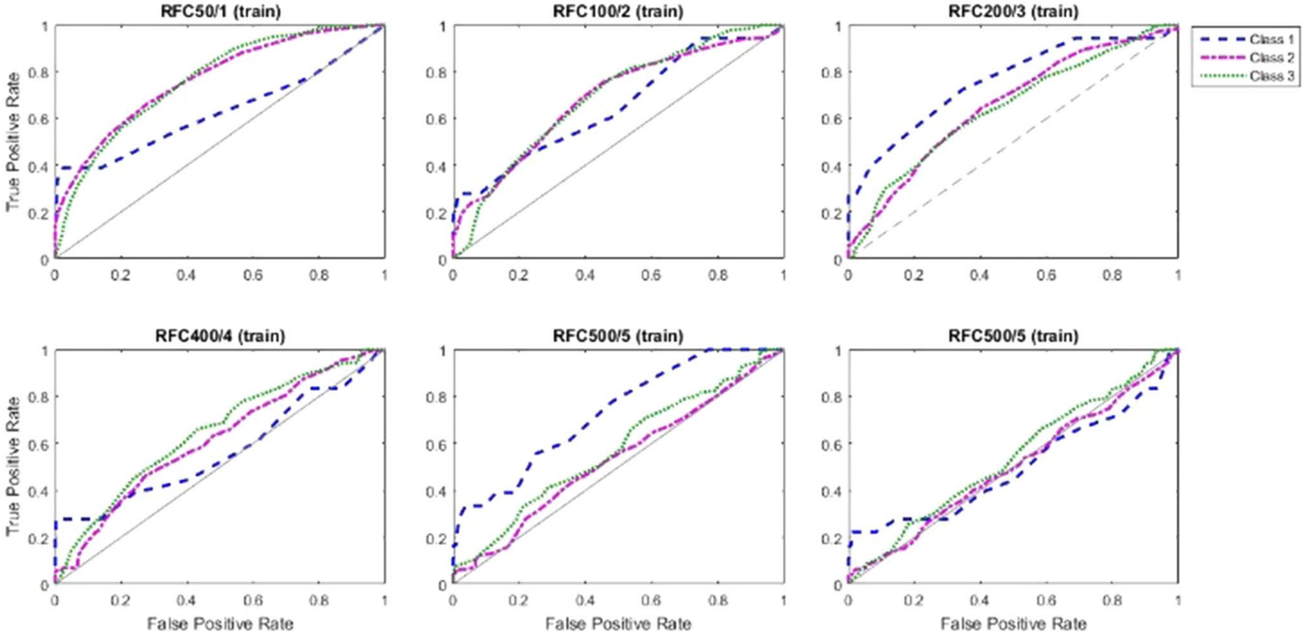

Fig. 6 ROC curve (training) for RFC

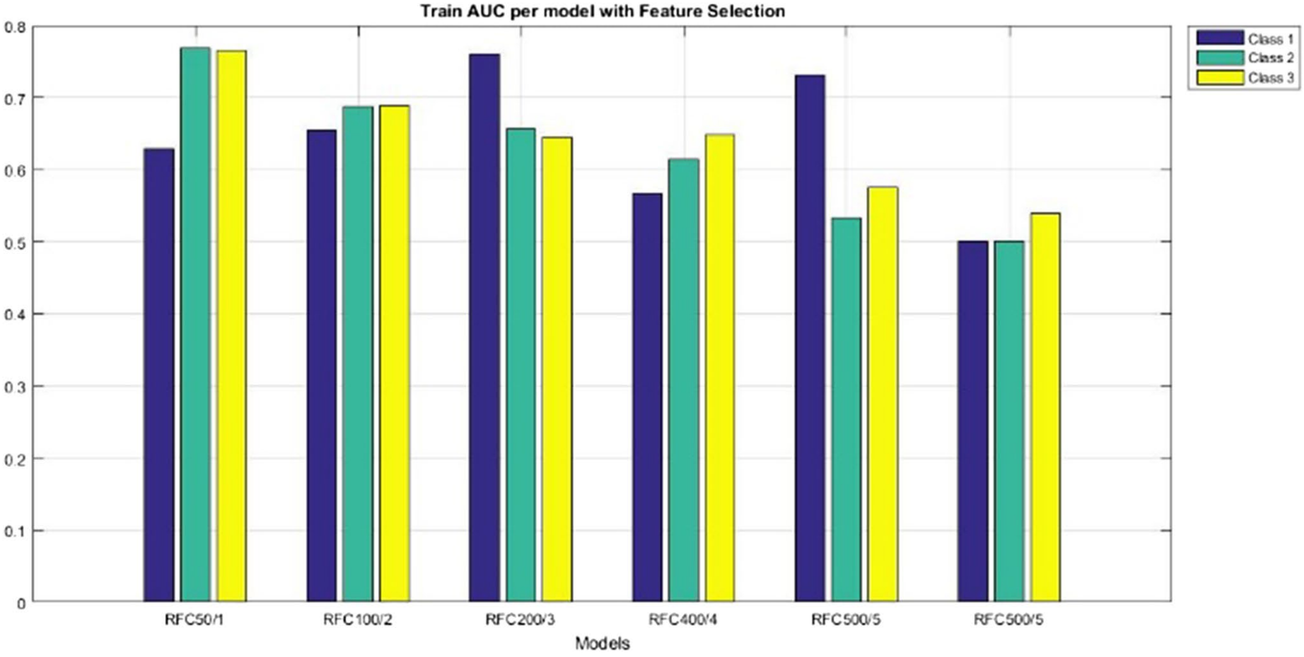

Fig. 7 AUC (training) for RFC

Table 7 RFC average with 3 classes (testing)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
RFC50/1	0.709	0.483	0.417	0.439	0.192	0.505	0.587
RFC100/2	0.493	0.656	0.388	0.393	0.149	0.634	0.554
RFC200/3	0.422	0.608	0.393	0.379	0.196	0.573	0.463
RFC400/4	0.508	0.55	0.368	0.381	0.03	0.556	0.527
RFC500/5	0.621	0.446	0.35	0.395	− 0.52	0.515	0.531

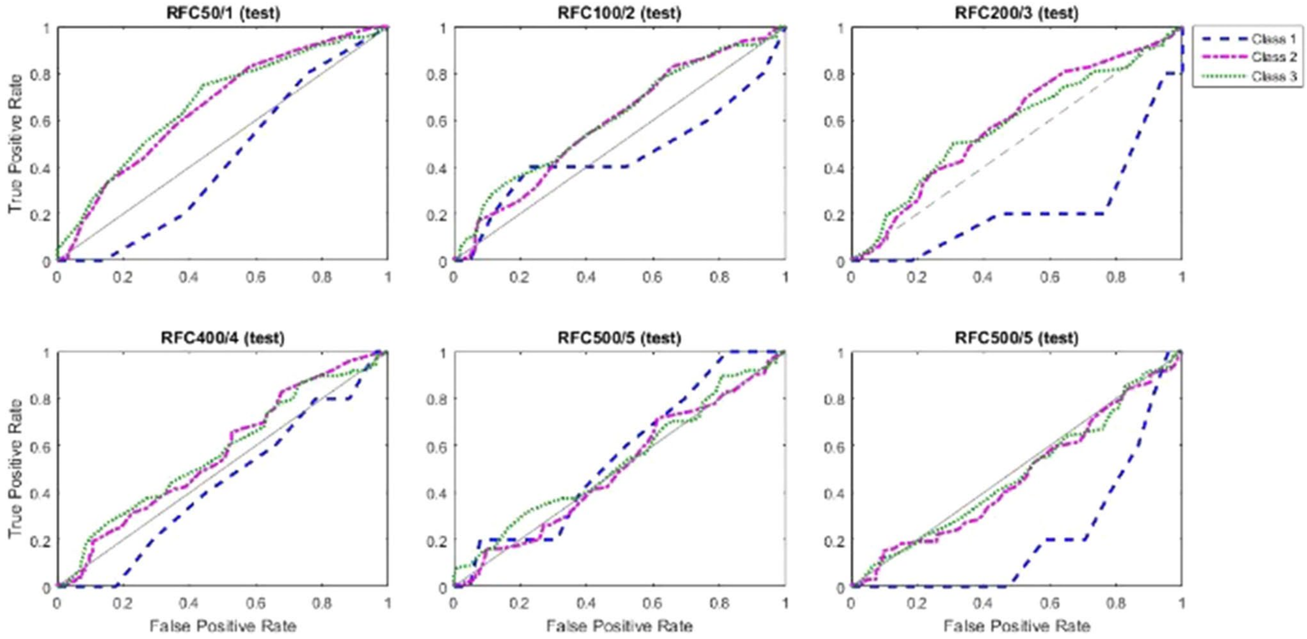

Fig. 8 ROC curve (testing) for RFC

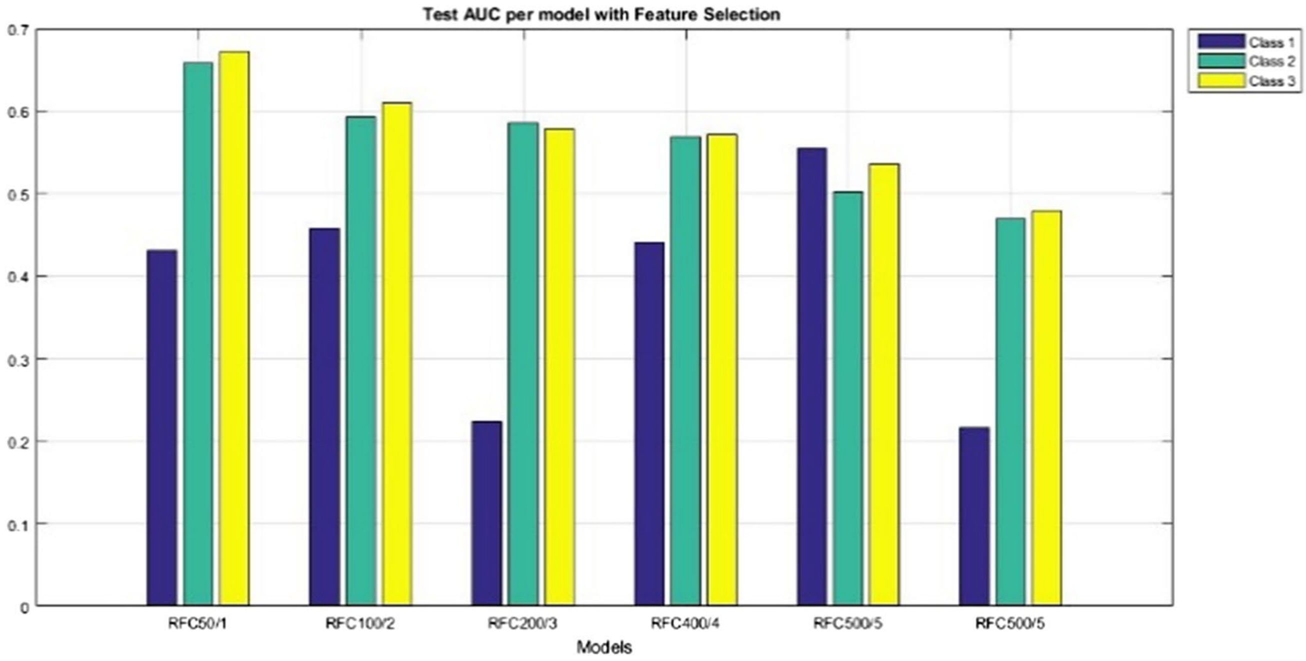

Fig. 9 AUC (testing) for RFC

Table 8 ANN performance (training)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
LEVNN	0.952	0.929	0.697	0.73	0.881	0.936	0.962
BPXNC	0.935	0.92	0.692	0.718	0.855	0.925	0.962

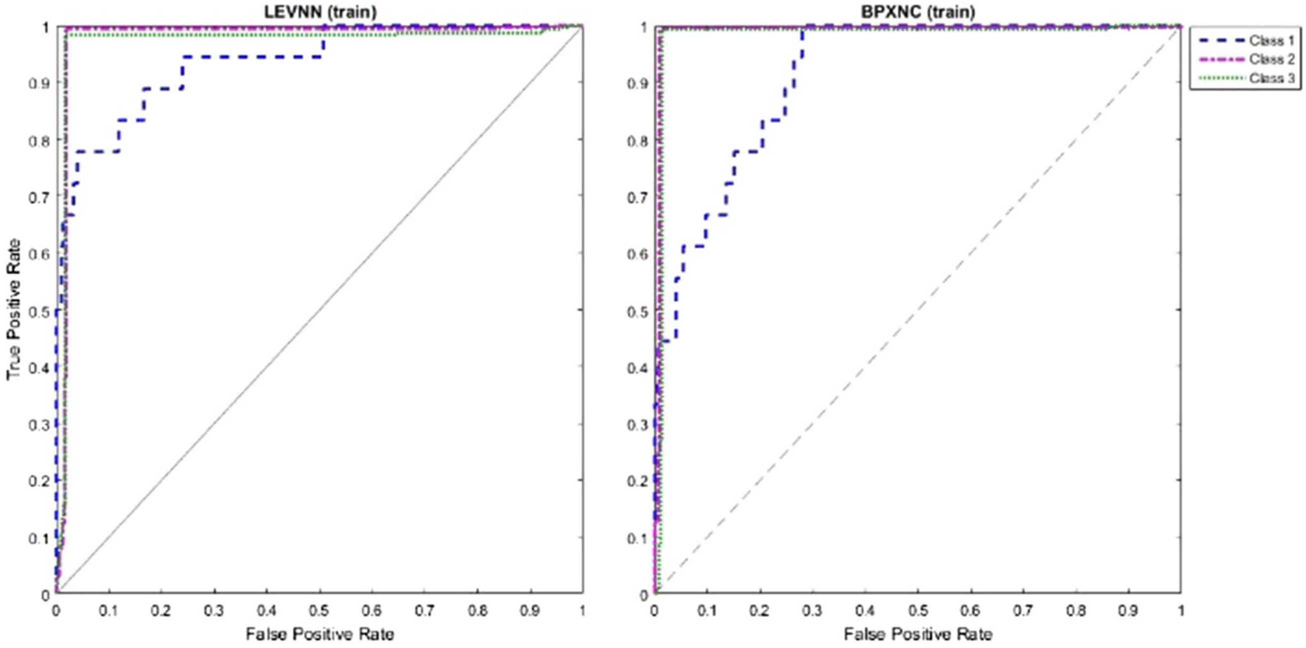

Fig. 10 ROC curve (training) for LEVNN and BPXNC

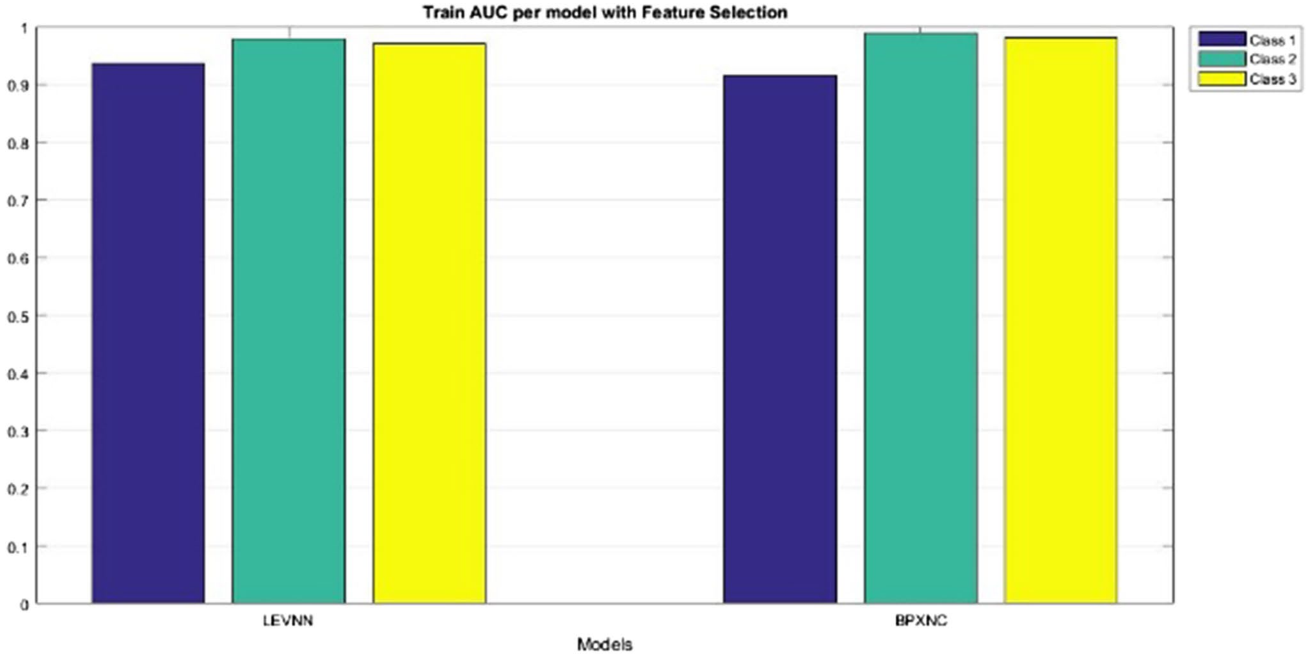

Fig. 11 AUC (training) for LEVNN and BPXNC

Table 9 ANN performance (testing)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
LEVNN	0.801	0.815	0.568	0.602	0.616	0.816	0.732
BPXNC	0.792	0.789	0.571	0.577	0.581	0.774	0.767

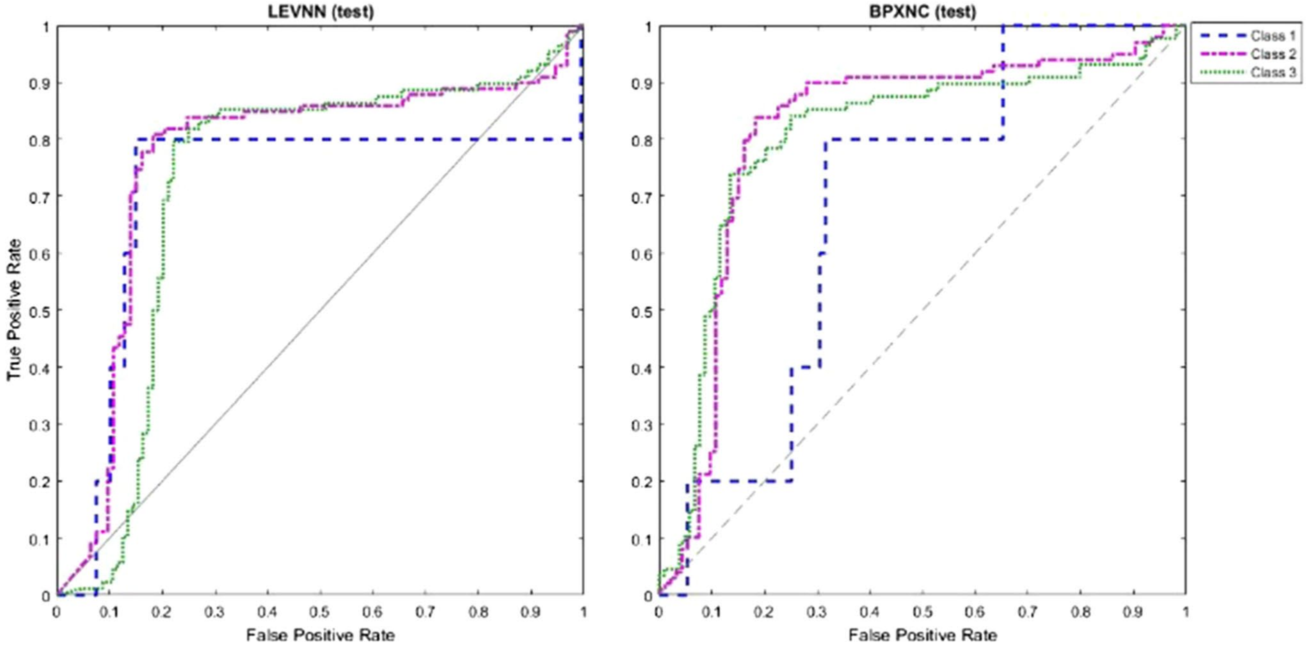

Fig. 12 ROC curve for ANN (testing)

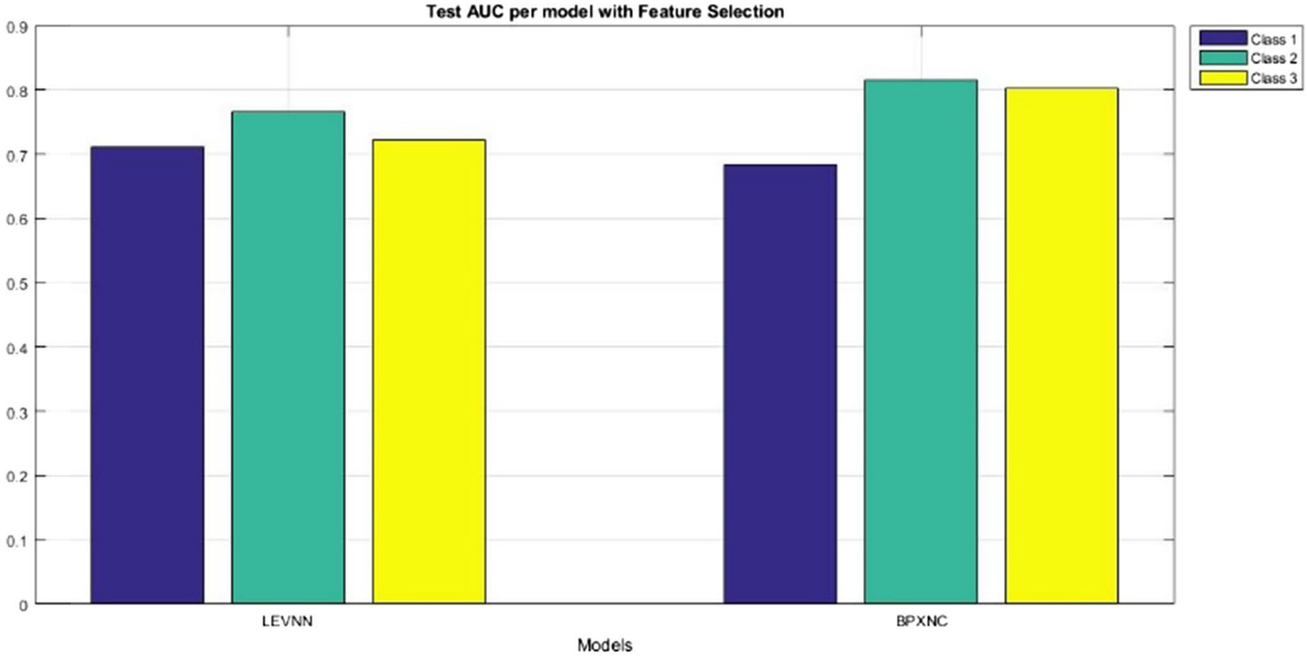

Fig. 13 AUC plot for ANN (testing)

Table 10 Combined classifiers (training)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
NN Com	0.997	0.995	1	0.997	0.992	1	1
LEVNN Com	0.997	0.995	1	0.997	0.992	1	1
NN and RF Com 1	0.999	0.999	0.999	0.999	0.998	0.999	1
NN and RF Com 2	0.999	0.999	0.999	0.999	0.998	0.999	1

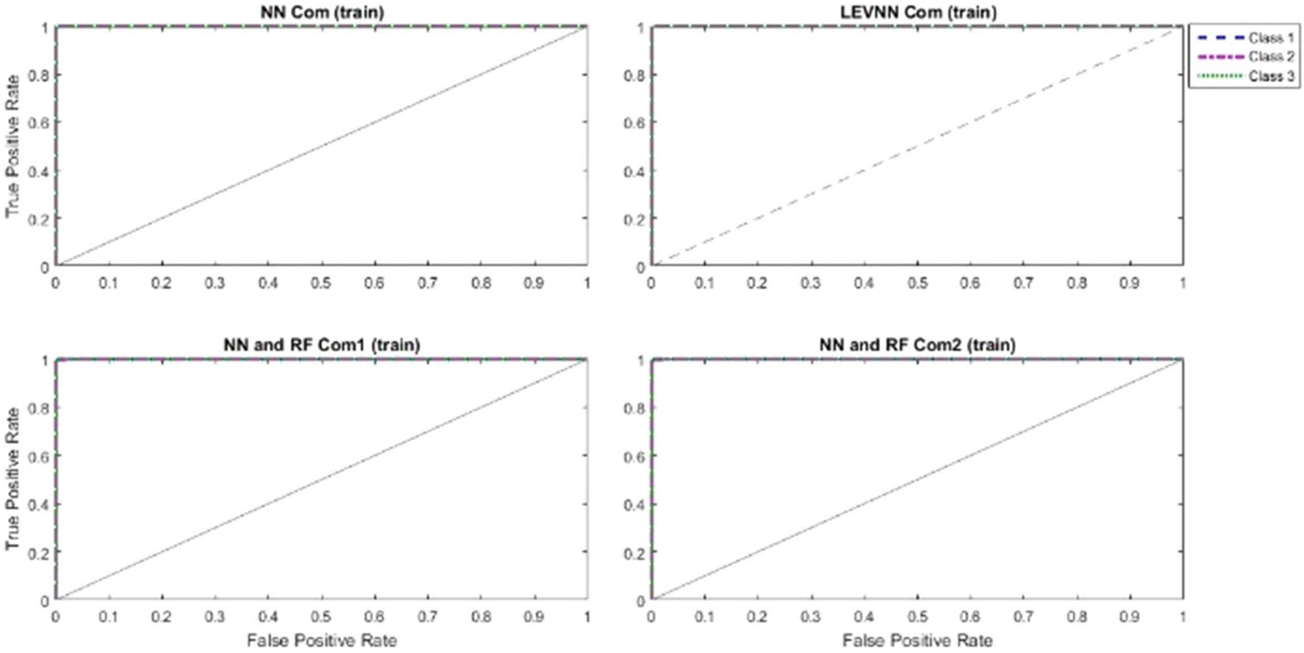

Fig. 14 ROC curve for the combined classifier (training)

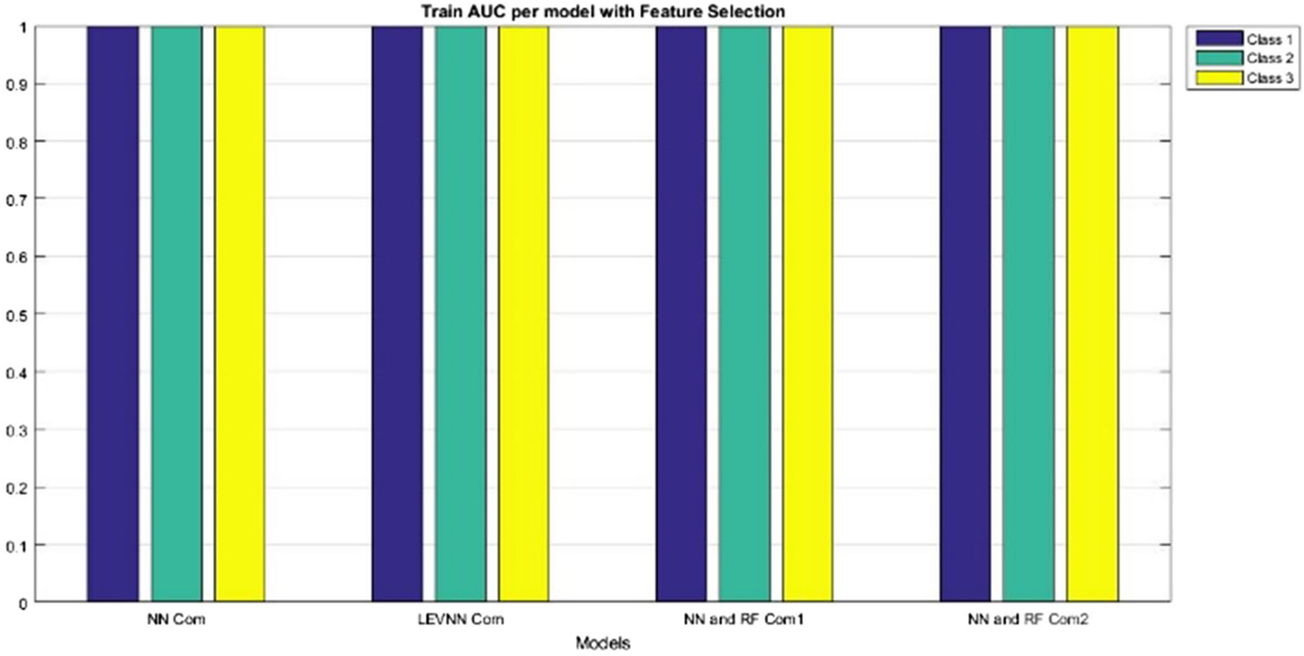

Fig. 15 AUC plot for the combined classifier (training)

Table 11 Combined classifiers (testing)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
NN Com	0.769	0.76	0.538	0.549	0.529	0.748	0.712
LEVNN Com	0.773	0.656	0.528	0.533	0.429	0.649	0.663
NN and RF Com 1	0.755	0.798	0.565	0.585	0.552	0.799	0.824
NN and RF Com 2	0.776	0.722	0.542	0.547	0.498	0.712	0.787

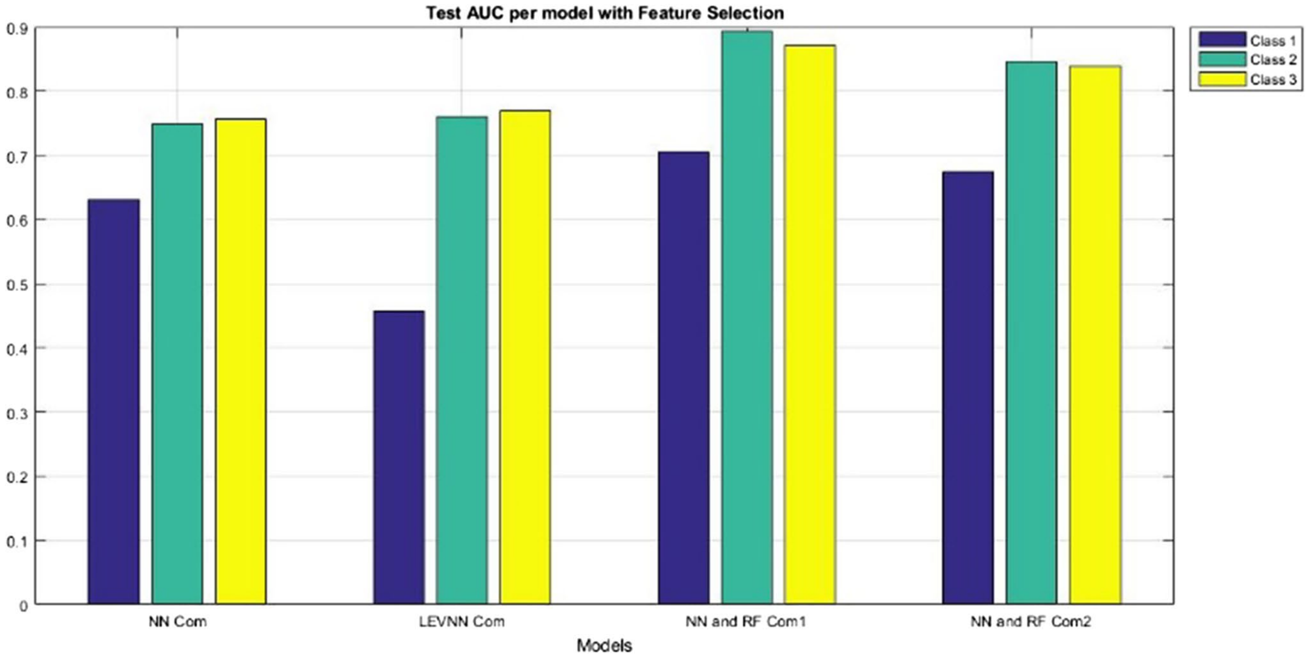

Fig. 16 AUC for the combined classifier (testing)

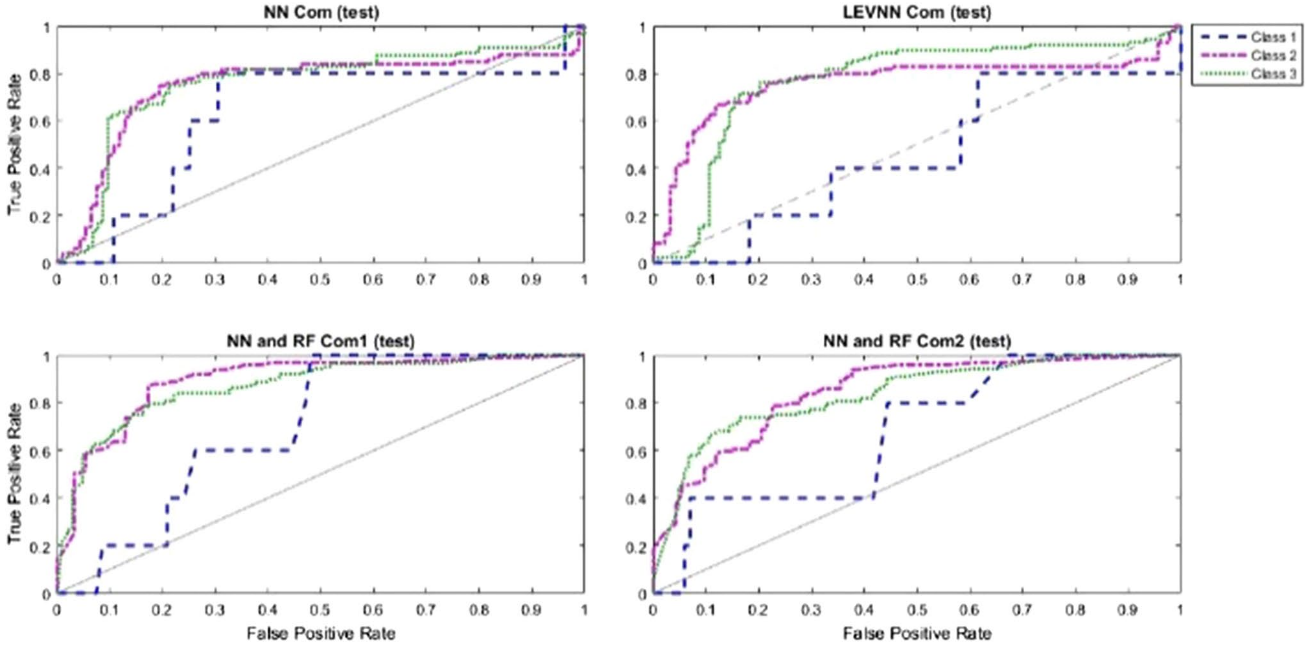

Fig. 17 ROC curve for the combined classifier (testing)

shown in Fig. 16. As shown in Fig. 17, the ROC generated for classes 2 and 3 were performing much better than class 1 compared between a true positive rate of 100% and a false positive rate of 0%.

7 Conclusion and future work

In this research, a data science methodology combines 16 features extracted from 1550 records to predict the student levelling. The chosen student dataset demonstrates non-leaner relationships, which creates a real challenge to the classifiers. The RFC classifiers performance is poor compared to the other classifiers, showing that RFC does not have the capabilities to handle the training data and unseen examples. However, the NN and RF Com1 performed well during the training stage, generating the highest performance. Yet, it did not perform as well during the testing process as the results were way lower than the training ones, but it generated the highest AUC of 0.824 over the rest of the classifiers.

LEVNN model produced the best performance results amongst all the models applied in this research during the testing process except the Precision and AUC. The highest figures of Sensitivity, Specificity, F1, J, and Accuracy are 0.801, 0.815, 0.602, 0.616, and 0.816, respectively. In contrast, the best results of AUC were achieved by NN and RF Com1 of 0.824. The AUC for the combined classifier (NN Com, LEVNN Com, NN and RF Com) generated 1 for the training sets, while testing generated AUC of 0.824 for the NN and RF COM1, which is the best result of all the algorithms used in this study. Compared with the other tested machine learning models, RFC did not show high Accuracy results.

Because the LEVNN method achieved the best performance outcomes during the testing phases, it has been considered the best classifier in this research, which is more important than training.

Overall, the type of results gained highlights the potential of Multi-National Schools' data for classifying the student levelling. The choice of model is essential when needing accepted results, as shown in the classification performance tables for training and testing in this research. The LEVNN classifier responded well to the students' data and has potential use in education.

The LEVNN is a powerful model for analysing the students' datasets, and it has proven in this domain that it presented substantial prediction accuracy and performance compared to other classifiers. A good relationship between input features and target values is discovered during the development process. The datasets are moderate in size, with 20% of the input features randomly selected for testing and the remaining percentages of 70% and 10% used for training and validation, respectively. In this context, the test set errors are averaged, and the procedure was repeated several times.

Sikder et al. [35] also used LEVNN in their study for predicting 120 students' performance with a training dataset of 70% from the total dataset, a testing dataset of 15%, and a validation dataset of 15%. They used the following features: Class Test Mark, Family Education, Class Performance, Living Area, Home, Class Attendance, Social Media Interaction, Assignment, Extra-Curricular Activity, Lab Performance, Drug Addiction, Study Time, Affair, Previous Result, and Year Final Result, living place and social interactions. In case of the success of their study, they will prove that students' performance can be affected by their lifestyle. Their study has shown good accuracy results for both outstanding and poor GPA performance of students with better accuracy for the poor performance giving an average of the accuracy of 97.2%. They claimed that a previous study used the ANN algorithm to predict students' performance using multilayer perception trained by static backpropagation. It produced less precision than the first study, with an average of 74%.

As shown in Tables 12 and 13, LEVNN is the best model for predicting students' performance for levelling usage in the testing stage, given a dataset of previously mentioned features. It produced accepted prediction accuracy and performance compared to other classifiers.

Table 12 Classification performance (training stage)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
RFC50/1	0.632	0.674	0.466	0.473	0.306	0.66	0.72
RFC100/2	0.592	0.673	0.434	0.460	0.265	0.69	0.68
RFC200/3	0.661	0.617	0.416	0.443	0.278	0.63	0.69
RFC400/4	0.524	0.657	0.405	0.416	0.181	0.65	0.61
RFC500/5	0.579	0.585	0.371	0.404	0.163	0.61	0.61
LEVNN	0.952	0.929	0.697	0.73	0.881	0.936	0.962
BPXNC	0.935	0.92	0.692	0.718	0.855	0.925	0.962
NN Com	0.997	0.995	1	0.997	0.992	1	1

Table 13 Classification performance (testing stage)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
RFC50/1	0.709	0.483	0.417	0.439	0.192	0.505	0.587
RFC100/2	0.493	0.656	0.388	0.393	0.149	0.634	0.554
RFC200/3	0.422	0.608	0.393	0.379	0.196	0.573	0.463
RFC400/4	0.508	0.55	0.368	0.381	0.03	0.556	0.527
RFC500/5	0.621	0.446	0.35	0.395	− 0.52	0.515	0.531
LEVNN	0.801	0.815	0.568	0.602	0.616	0.816	0.732
BPXNC	0.792	0.789	0.571	0.577	0.581	0.774	0.767
NN Com	0.769	0.76	0.538	0.549	0.529	0.748	0.712

Latest studies revealed that machine learning is found to be effective for pre-processing of student datasets to predict student suitability for certain careers and their capacity to learn in certain subjects. Our data can be analysed effectively with machine learning models when introduced to the classification of student past levelling marks and generate the student level in future. For future improvements, we need to consider using other machine learning techniques and increase the number of features used such as learning style of students, their motivation and interest, concentration level, family background, personality type, information processing ability and the way they attempt exams, however, the limitation is that the study has not considered students' performance in exams in their previous and current academic stage which can be an improvement to our study.

Author contributions All authors in this research have contributed equally to the study conception and design. Thar Baker, Rawaa Al-Jumeily, and Dhiya Al-Jumeily contributed to the research the background of the research as well as the introduction. Abir Jaafar Hussain, Wasiq Khan, and Ahmed Al Shammaa contributed to the related work of the study, classification algorithms, and model evaluation metrics. Mohammed Khalaf and all authors contributed to implementation of machine learning and statistical tools. All authors read and approved the final manuscript.

Data availability The data set generated during and analysed during the current study are available in the Mendeley Data repository, https://data.mendeley.com/datasets/3g8dtwbjjy/1 (https://doi.org/10.17632/3g8dtwbjjy.1).

Declarations

Competing interests The authors declare no competing interests.

References

1. Albreiki B, Zaki N, Alashwal H. A systematic literature review of student' performance prediction using machine learning techniques. Educ Sci. 2021;11(9):552. https://doi.org/10.3390/educsci11090552.
2. Lee MW, Chen SY, Chrysostomou K, Liu X. Mining students' behavior in web-based learning programs. Exp Syst Appl. 2009;36(2):3459–64. https://doi.org/10.1016/j.eswa.2008.02.054.
3. AL, Wiener M. Classification and regression by randomForest. R News 2. 2003;3:18–22.
4. Ghareeb S, Hussain A, Khan W, Al-Jumeily D, Baker T, Al-Jumeily R. Dataset of student level prediction in UAE. Data Brief. 2021;35: 106908. https://doi.org/10.1016/j.dib.2021.106908.
5. Kučak D, Juričić V, Đambić G. Machine learning in education—a survey of current research trends. Ann DAAAM Proc Int DAAAM Symp. 2018;29(1):0406–10. https://doi.org/10.2507/29th.daaam.proceedings.059.
6. Ghareeb AS, Al-jumeily R, Baker T. A machine learning based framework for education levelling in multicultural countries: UAE as a case study. 2020;14(3).
7. Masci C, Johnes G, Agasisti T. Student and school performance across countries: a machine learning approach. Eur J Oper Res. 2018;269(3):1072–85. https://doi.org/10.1016/j.ejor.2018.02.031.

8. Al-Shabandar R, Hussain A, Laws A, Keight R, Lunn J, Radi N. Machine learning approaches to predict learning outcomes in Massive open online courses. Proc Int Jt Conf Neural Netw. 2017;713–720:2017. https://doi.org/10.1109/IJCNN.2017.7965922.
9. Hsia TC, Shie AJ, Chen LC. Course planning of extension education to meet market demand by using data mining techniques—an example of Chinkuo technology university in Taiwan. Expert Syst Appl. 2008;34(1):596–602. https://doi.org/10.1016/j.eswa.2006.09.025.
10. Lykourentzou I, Giannoukos I, Nikolopoulos V, Mpardis G, Loumos V. Dropout prediction in e-learning courses through the combination of machine learning techniques. Comput Educ. 2009;53(3):950–65. https://doi.org/10.1016/j.compedu.2009.05.010.
11. Nath V, Levinson SE. Machine learning. 2014. https://doi.org/10.1007/978-3-319-05606-7_6.
12. Tong JC. Cross-validation. Encyclopedia of Systems Biology. 2013;508–508. https://doi.org/10.1007/978-1-4419-9863-7_941.
13. Bertsekas DP. Dynamic programming and optimal control. 4th edn. 2012.
14. Celik AN. A techno-economic analysis of wind energy in Southern Turkey. Int J Green Energy. 2007;4(3):233–47. https://doi.org/10.1080/15435070701338358.
15. Zhuang X, Zhang W, Wu Y, Zhao Z. Comprehensive prediction method for die-roll height of fine-blanking components. Int J Adv Manuf Technol. 2018;98(9–12):2819–29. https://doi.org/10.1007/s00170-018-2430-y.
16. Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J Pharm Biomed Anal. 2000;22(5):717–27. https://doi.org/10.1016/S0731-7085(99)00272-1.
17. Haghbakhsh R, Adib H, Keshavarz P, Koolivand M, Keshtkari S. Development of an artificial neural network model for the prediction of hydrocarbon density at high-pressure, high-temperature conditions. Thermochim Acta. 2013;551:124–30. https://doi.org/10.1016/j.tca.2012.10.022.
18. Jain A, Solanki S. An efficient approach for multiclass student performance prediction based upon machine learning. 2019;1457–1462.
19. Kotsiantis SB. Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. Artif Intell Rev. 2012;37(4):331–44. https://doi.org/10.1007/s10462-011-9234-x.
20. Khalaf M, et al. Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models. Neurocomputing. 2017;228:154–64. https://doi.org/10.1016/j.neucom.2016.10.043.
21. Ting SL, Ip WH, Tsang AHC. Is Naïve Bayes a Good Classifier for Document Classification? 2011. [Online]. Available: https://www.researchgate.net/publication/266463703.
22. Su RKRB. Linear feature extraction and description. 1980.
23. Thomas Rincy N, Gupta R. An efficient feature subset selection approach for machine learning. Multimed Tools Appl. 2021;80(8):12737–830. https://doi.org/10.1007/s11042-020-10011-7.
24. Dietterich T. Overfitting and Undercomputing in Machine Learning.
25. Hall MA. Correlation-based feature selection for machine learning. 1999.
26. Vafaie H, De Jong K. Genetic algorithms as a tool for feature selection in machine learning. In: Proceedings International Conference on Tools with Artificial Intelligence, ICTAI. pp. 200–203, 1992, https://doi.org/10.1109/TAI.1992.246402.
27. Ramaswami M, Bhaskaran R. A study on feature selection techniques in educational data mining. 2009;1(1): 7–11. [Online]. Available: http://arxiv.org/abs/0912.3924.
28. Jović A, Brkić K, Bogunović N. A review of feature selection methods with applications. In: 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 Proceedings, 2015, pp. 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458.
29. Luo S. Data mining of many-attribute data: investigating the interaction between feature selection strategy and statistical features of datasets.
30. Breiman L. Randon Forests. Machinelearning202.Pbworks.Com, pp. 1–35, 1999, [Online]. Available: http://machinelearning202.pbworks.com/w/file/fetch/60606349/breiman_randomforests.pdf.
31. Belayneh A, Adamowski J, Khalil B, Quilty J. Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. Atmos Res. 2016;172–173:37–47. https://doi.org/10.1016/j.atmosres.2015.12.017.
32. Schapire RE. The strength of weak learnability. Mach Learn. 1990;227:197–227.
33. Aguirre-Gutiérrez J, Seijmonsbergen AC, Duivenvoorden JF. Optimizing land cover classification accuracy for change detection, a combined pixel-based and object-based approach in a mountainous area in Mexico. Appl Geogr. 2012;34:29–37. https://doi.org/10.1016/j.apgeog.2011.10.010.
34. Ko BC, Kim HH, Nam JY. Classification of Potential Water Bodies Using Landsat 8 OLI and a Combination of Two Boosted Random Forest Classifiers. 2015; 13763–13777. https://doi.org/10.3390/s150613763.
35. Sikder MF, Uddin MJ, Halder S. Predicting students yearly performance using neural network: a case study of BSMRSTU. In: 2016 5th International Conference on Informatics, Electronics and Vision, ICIEV 2016, Nov. 2016, pp. 524–529. https://doi.org/10.1109/ICIEV.2016.7760058.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.