International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Jan- 24)

http://dx.doi.org/10.12785/ijcds/150119

Convolutional Neural Network for Predicting Student Academic Performance in Intelligent Tutoring System

Fatema Alshaikh[1] **and Nabil Hewahi**[2]

[1,2]*Department of Computer Science,University of Bahrain, Sakheer, Bahrain*

Received 14 Sep. 2023, Revised 5 Jan. 2024, Accepted 7 Jan. 2024, Published 15 Jan. 2024

Abstract: One of the most significant research areas in education and Artificial Intelligence (AI) is the earlier prediction of students' academic achievement. Limited studies have been conducted using Deep Learning (DL) in the student domain of Intelligent Tutoring System (ITS). Traditional Machine Learning (ML) techniques have been employed in many earlier publications to predict student performance. This paper investigates the effectiveness of DL algorithms for predicting student academic performance. Three different DL architectures based on the structure of Convolutional Neural Networks (CNN) are presented. Two public datasets are used. Furthermore, two feature selection techniques are utilized in this experiment: Principal Component Analysis (PCA) and Decision Trees (DTs). Also, we applied a resampling technique for the first dataset to address the issue of an imbalanced dataset. According to the experimental findings, the proposed CNN model's success in predicting student performance at early stages reached an accuracy of 94.36% using the first dataset and 84.83% using the second dataset. Comparing the proposed approach with the previous studies, the proposed approach outperformed all previous studies when dataset 2 and part of dataset 1 were used. For the complete dataset 1, the proposed model performed very well.

Keywords: Machine Learning , Deep Learning , Convolutional Neural Networks, Intelligent Tutoring System, Principal Component Analysis, Decision Trees

1. INTRODUCTION

Education plays a crucial role in the growth of a country, and it is a key factor in achieving success in life. Academic institutions strive to provide their learners with a high-quality education to enhance learning [1]. Students' academic achievement is a vital part that determines the success of any educational institution. Educational organizations have started to use Artificial Intelligence technology to improve student learning. Currently, these organizations have significant difficulty providing quality learning for their learners while improving their success rate.

In contemporary times, Machine Learning holds a significant position in forecasting students' academic achievement, thereby facilitating them to attain higher grades. It is a helpful tool for taking early steps to improve student's performance and minimize failures at the end of the course. Intelligent Tutoring Systems (ITS), based on Artificial Intelligence (AI) and Machine Learning (ML), are an example of a tool that improves teaching abilities, assists students in learning, and may engage students in dialogues, reply to them, and provide feedback [2].

Student modeling is the central component of an ITS and a broad research field. Student profile modeling [3] is based on a learner profile description that combines the crucial features and provides the student's most coherent, complete, and operational picture. Background knowledge, learning preferences, behaviors, talents, objectives, and so on are all student characteristics. The student profile model may be built by analyzing data from many places (e.g., online learning platforms, social media, and school records). Indeed, some studies have utilized the student profile to suggest adaptive learning, advise them on academic choices [4]. Machine learning techniques are employed to derive meaningful insights from data for the purposes of informed decision-making in student profile modeling.

Researchers from many disciplines propose various techniques and approaches for predicting student performance. They also improve current approaches to increase prediction outcomes. Among all disciplines, machine learning plays a critical role in building models used for prediction purposes. Furthermore, many datasets and performance evaluation criteria are publicly available, allowing researchers to verify performance and provide improved

E-mail address: 20005999@stu.uob.edu.bh, nhewahi@uob.edu.bh

findings.

This paper proposes a novel approach depending on deep learning algorithms, including the Convolutional Neural Network (CNN), to predict student performance using two educational public datasets. We proposed three different models depending on deep learning techniques. Different evaluation metrics are utilized to assess student's performance within the year.

The remainder of the paper is structured in the following manner: Section 2 presents a review of the literature on predicting student performance using machine learning and deep learning techniques. Section 3 provides an overview of the model proposed. The datasets and evaluation metrics used to assess the performance of each model are described in sections 4 and 5, respectively. The experimental findings are presented in section 6. The discussion is reported in section 7. Finally, section 8 depicts the conclusion.

2. Related work

Many studies aim to improve student academic performance, analyze obstacles faced throughout the learning process, and provide ways to improve competencies. In other studies, finding failure/dropout criteria was challenging. They analyzed contexts to predict academic results and provide educational solutions. Another common purpose is adaptive learning, which enhances the quality of the educational environment and, ultimately, the learners' achievements.

The personalization that occurs in adaptive educational software and Intelligent Tutoring Systems is built on the foundation of student profiles. The results of specific exercises relating to a topic are used to evaluate a student's level of understanding. To do this, Sanjay et al. [5] have developed a variety of model extensions throughout the years that include certain cognitive traits in the student profile. An innovative approach to student profiling has been put forward in this study [5]. The suggested approach considers characteristics like the quality of questions and mistakes caused when practicing.

A good survey [3] presented the current state of the art in student profile modeling by employing machine learning strategies throughout the past four years. For various objectives, including failure, dropout, orientation, academic performance, etc., the study examined popular and effective machine learning approaches in traditional and online classrooms. The findings [3] indicated that most research investigations employ decision trees because they are the most effective and widely used. Moreover, the essential traits utilized to create a student profile are academic, personal identification, and internet behavior. An experiment based on ML algorithms was conducted to reinforce the survey findings. The decision tree performed best, which supports the survey results.

A recent study by Khan et al. [6] evaluated the efficiency of machine learning algorithms for measuring students' educational progress and alerting tutors to students' challenges, which might improve learning outcomes. This research hypothesis's result includes supportive strategies to control students' progress from the course's beginning effectively and a preventative approach to providing struggling students with great attention. Four machine learning algorithms were selected to predict the performance of students (k-Nearest Neighbours, Decision Tree (DT), Naive Bayes (NB), and Artificial Neural Networks (ANN)). The study proposed choosing a decision tree model as the best model for evaluation.

Recently, Ng et al. [7] have proposed a data mining approach for identifying essential factors that affect student performance based on data from two secondary schools in Portugal. Several machine learning algorithms are used for classification: Support Vector Machine (SVM), NB, and Multilayer Perceptron (MLP). The highest performance was obtained from SVM, scoring 91%. In addition, research by Liu and Koedinger [8] used data mining techniques to improve an Intelligent Tutoring System. The study results demonstrated that student learning outcomes are increased using the proposed system compared with the prior version.

A previous study by Imran et al. [9] has built a prediction model using the ensemble method, which combines several models to increase the accuracy of student performance prediction rather than using a single model. In addition, Lincke et al. [10] have recognized using machine learning algorithms to predict learning outcomes from student quiz answers and reading records from a Web-based learning system. According to the study's findings, the difficulty of the question and the number of times it has been answered incorrectly are valuable parameters for determining whether a student's response is correct. Moreover, Ghorbani et al. [11] evaluated several resampling methods for dealing with imbalanced data while predicting student performance. They used different ML techniques, such as SVM, Random Forest (RF), ANN, DT, and NB. The result indicates that the performance of classifiers was increased when using balanced datasets compared to imbalanced ones.

Many existing studies in the broader literature have examined the use of Deep Learning (DL) to increase the performance of student prediction. DL has recently allowed researchers to have the ability to extract high-level features from raw data automatically, affecting student performance prediction. A study in 2021 [12] proposed a model to predict student performance from historical records using a Deep Neural Network (DNN) called Bidirectional Long Short-Term Memory (BiLSTM). The prediction accuracy of the developed model was 90.16%. Another exciting study [13] investigated the role of student drawing in learning. The authors introduced a diagrammatic student model based on neural network architecture. Compared to competitive baseline techniques, it can predict student performance

more precisely.

Seminal contributions have been made by Siddique et al. [14]. The study aimed to identify the essential aspects influencing secondary school student's performance and to develop an effective classification model for academic performance prediction by combining single and ensemble-based classifiers. Three single classifiers were examined individually, comprising a Multilayer Perceptron (MLP), J48 (DT), and PART, as well as three ensemble techniques, including Bagging, MultiBoost, and Voting. In addition, nine new models were created by combining single and ensemble classifiers to improve previous algorithm performance. MultiBoost with MLP outscored the others in the study, scoring 98.7% accuracy. According to the study [11], the suggested model would effectively assess secondary school students' academic performance at an early stage to enhance academic achievement. Moreover, a study by Li et al. [15] proved that predicting student performance helps in course selection and developing suitable future study plans for learners. Furthermore, teachers and supervisors monitor students, give support, and implement training plans to achieve the best results. The proposed technique used a DNN to extract valuable data as a feature with appropriate weights. Neural networks use multiple hidden layers regulated by feed forwarding and backpropagation data from prior cases. The suggested model achieving the best prediction in MAE (0.593) and RMSE (0.785).

Many academics have utilized traditional machine learning algorithms to predict student academic achievement. Still, relatively limited studies have used convolutional neural networks' structure in the intelligent tutoring discipline. A survey by Poudyal et al. [16] found a novel model using a 2D CNN model by integrating two 2D CNN models to outperform standard baseline models (NB, DT, and logistic regression).

Many studies in the literature have examined the datasets utilized in this study. Two datasets were collected; the first was gathered and investigated by Paulo Cortez and Alice Silva of Portugal's University of Minho. It is available in two subjects from two Portuguese secondary schools: Mathematics and Portuguese language. The second dataset was provided by the Kalboard 360 learning management system. It is a three-class dataset where students are sorted into three grade levels: Low, Middle, and High, according to their total grade.

Several authors applied different machine learning algorithms to predict student performance using both datasets. For instance, a study in 2019 [9] used supervised machine learning algorithms and ensemble methods and compared the results using the first dataset. The study presented in [9] highlights the importance of data preparation and algorithm fine-tuning in addressing issues related to data quality. In this experiment, three supervised learning algorithms, namely J48, NNge, and MLP were used. The results indicate that J48 outperformed its competitors with a 95.78% accuracy rate. Furthermore, an ensemble approach is used to improve the precision of weak classifiers, and when compared to predictions from a single model, ensemble predictions are often more precise.

Another study [7] used the same dataset with three classification models: SVM, NB, and MLP. The performance indicators were F1-Score, recall, accuracy, and precision. SVM achieved the highest accuracy either for the binary or five class classifications. In addition to the previous studies, a work by Hamoud [17] used and evaluated the first dataset with three DT algorithms (J48, RepTree, and Hoeffding Trees). The results demonstrated that the J48 algorithm accurately classified and predicted students' intent to finish higher education and course progress.

On the other hand, several publications have been released documenting the use of machine learning and deep learning algorithms with the second dataset used in the current study. An improved model of student performance was created using a different set of behavioral characteristics [18]. This sort of feature is associated with learner interaction with an e-learning system. The study [18] examined data mining techniques such as ANN, NB, and DT classifiers to determine the effect of such features on student academic achievement. The findings showed a substantial link between learner behaviors and educational outcomes. Compared to the same dataset, results with various classification algorithms employing behavioral characteristics improved classification accuracy by up to 29%.

One year later, in 2016, the same researchers [19] used ensemble approaches, such as Bagging, Boosting, and RF, to increase the model's performance. The model's accuracy improved by up to 25.8% when using ensemble approaches. The model's accuracy was greater than 80% when tested on new learners. This result validates the suggested model's dependability.

A proposed framework by Saleem et al.[20] includes five machine-learning algorithms and four ensemble techniques: bagging, boosting, stacking, and voting. Using ensemble techniques, the model performance has improved significantly. Among alternative ensemble approaches, the stacking model excelled and achieved the highest F1 score (0.819) by integrating all five classifiers. The ML model integration enhanced the prediction accuracy and outperformed all other ensemble techniques. The suggested methodology can effectively evaluate student performance and assist instructors in making informed decisions. Another study [21] in 2019 developed a Deep Neural Network-based student performance prediction system. After training and evaluating the model with Kaggle datasets using several techniques in R Programming such as Decision Tree (C5.0), NB, RF, SVM, KNN, and DNN, and compared the accuracy of all other algorithms. The result indicated that DNN exceeded the six different algorithms with an accuracy of 84%.

A work by Akour et al. [22] explored the efficiency of using CNN to estimate students' performance and predict whether a student could complete their degree. The experimental findings showed that the suggested model outperformed the current methodologies regarding prediction accuracy. Table I summarizes some previous studies in this field.

3. The proposed Model

Based on the literature review, more research needs to be done on using DL algorithms, specifically CNN, to predict students' academic performance. Previous studies have focused on using traditional machine learning algorithms, for instance, NB, DT, SVM, and RF, to predict performance. Despite some existing research using KNN and DNN to build the models, their experiments' results needed more consistent accuracy. One research gap has been noted as none of the works under review examined building different models using CNN and fully connected layers. To overcome these research gaps, we proposed three different models based on CNN architecture using two educational datasets. Several authors are trying many algorithms and techniques to determine the best and most reliable results. We believe that our work conducted in this paper would significantly contribute to the field of deep learning.

A. Environment

The primary purpose of this research is to create prediction models utilizing DNN approaches to predict students' academic success in coming courses early in the semester based on their prior academic success. These models will be useful since learners will be notified of their expected outcomes earlier in the semester. As a result, students will be able to enhance their learning performance at the end of the term. Three DNN classifier models were built to predict student performance. We used an interactive computing data science platform called Jupyter Notebook. It is a flexible tool to create and configure scientific computing and machine learning workflows. Moreover, the models were created with Python 3 and TensorFlow 1.15.0.

Python is a high-level programming language used for scientific research and computation. Python is home to multiple open-source and general-purpose ML libraries used to train DL models. TensorFlow is an open-source machine learning library for DNN training and inference. It is used for numerical computation utilizing data flow graphs. Data flow graphs are often referred to as Static Computation graphs. A developer must first create the input layer and link each input layer to the hidden layer, followed by a similar process from the hidden layer to the output layer. The graphs comprise tensors and operations, which define all the neural networks and mathematical calculations. TensorFlow has a Graphical Processing Unit package where all matrix calculations may be performed [12].

B. The Architecture of the Proposed Model

The proposed model's structure is depicted in Figure 1. The model is composed of eight distinct stages. The initial stage involves the acquisition of the datasets gathered for the study. The use of robust, high-quality datasets is essential to ensure the accuracy and reliability of the model. More explanation about the datasets and their attributes is shown in sections 4-A and 4-B. The second stage is to preprocess the dataset and improve its quality which is discussed in section 4-C. The process of extracting suitable features from the data to improve the performance of the models arises in the third stage. Two types of feature selection were obtained (i.e., Decision Trees and Principal Component Analysis) to figure out which one might give better results, which is demonstrated clearly in section 4-D. The next stage is to create a training set consisting of (70%) of the data and a testing set consisting of (30%) of the data. The attributes and classes in the training dataset are separated and saved in a TensorFlow. The fifth stage is building the deep learning models using a Convolutional Neural Network (CNN) architecture. Three different models were developed to explore the best architecture that gives the highest accuracy; Model 1 used CNN as input, Model 2 used fully connected layers, followed by CNN, and Model 3 used CNN, followed by fully connected layers, followed by CNN. Section 3-D provides additional details about these layers. Section 6 presents experimental results obtained from predicting student performance. The next stage is to evaluate the models using some evaluation metrics defined in section 5. A comparison between the proposed models and other previous studies is shown in sections 6-B and 6-D.

C. Overview of CNN

The structure of CNN consists of three main layers [23] : convolution, pooling, and fully connected layers. The feature extraction is carried out by the first two layers (convolution and pooling). The final output is mapped onto the extracted features by the third layer (fully connected layer). More descriptions about these layers in the following:

1) Convolution layer : A convolution layer [23] is a central part of the CNN structure. The main purpose of the convolution layer is to extract features from data. It comprises a variety of functions, such as convolution and activation functions. The essential idea is that a kernel, a small array of numbers, generates a feature map by taking the element-wise product between the kernel and input tensor and summing up the results to generate the output value.
2) Pooling layer : A pooling layer handles a conventional A pooling layer down samples feature maps to reduce dimensionality, introduce translation invariance, and reduce learnable parameters. Max pooling is the most commonly used type of pooling, which selects patches and outputs the patch's maximum value.
3) Fully connected layers : Fully connected layers are also called dense layers, where each input neuron is associated with each output neuron in the prior layer. In most cases, the output of the last convolution or

TABLE I. Summary of the previous related work.

Ref	Year	Techniques	Best Technique	Dataset size	Evaluation Metrics
[18]	2015	ANN, NB, and DT	ANN	480	Accuracy, Precision, Recall, and F-Measure
[19]	2016	ensemble approaches, such as Bagging, Boosting, and RF	Ensemble	480	Accuracy, Precision, Recall, and F-Measure
[17]	2016	DT (J48, RepTree, and Hoeffding Tree)	J48	1044	Accuracy
[9]	2019	Ensemble (J48, Realada- boost)	Ensemble	1044	Accuracy
[21]	2019	Decision Tree (C5.0), NB, RF, SVM, KNN, DNN	DNN	500	accuracy, precision, recall, F-score, ROC curve, RMSE
[11]	2020	RF	RF	1044	Accuracy
[22]	2020	CNN	CNN	480	Accuracy, Precision, Recall, and F-Measure
[20]	2021	ML algorithms with ensemble: BAG, boosting, stacking, and VT	Stacking	480	F1 score
[3]	2021	DT, KNN, SVM, NN, NB	DT	395	Accuracy
[6]	2021	k-NN , decision tree , MLP	DT	151	Accuracy, precision, recall, F-Measure and MCC)
[7]	2021	NB, MLP, SVM	SVM	1044	Accuracy, precision, recall, and F-Measure
[12]	2021	CNN, Attention-based BiLSTM	BiLSTM	1044	Accuracy
[14]	2021	MLP, J48, and PART BAG, MB, VT	MB,MLP	1227	Accuracy, precision, recall, and F-score
[15]	2021	DNN	DNN	3,828,879	MAE , RMSE
[16]	2022	CNN, k-NN, NB, DT, and logistic regression	CNN,	32,593	Accuracy

pooling layer is converted into one flatten dimension array [23]. A subset of fully connected layers maps the features to the network's final outputs. The number of output nodes in the final dense layer typically equals the number of classes. A nonlinear function, for example, rectified linear unit (ReLU), is followed by each fully connected layer.

D. Model Creation

The planned architecture is divided into three different models. We used different epochs for all the built models: 10, 50, 100, and 200. The details regarding the proposed model can be found in the following subsections. The structure of each model includes details about the layers, the name and type of all layers in the model, their placement within the model, the output shapes of each layer, the number of parameters (weights) in each layer, and the overall number of trainable and non-trainable parameters of the model. Tables II, III, IV and Figures 2,3,4 summarise the layers of each model. To be noted that the (None) values in the tables and figures mean that the model can accept inputs of any dimension, a flexible batch size.

1) Model 1: CNN with one dimension as the input

Model 1 architecturecomprises five layers: an input layer, a pooling layer, a flatten, and 2 dense layers, as shown in Figure 2 and Table II. This layer converts the sample into a (58,32) shape vector. The second layer is a pooling layer called MaxPooling1D, using a length and stride of 2, which splits the size of the convolutional layer's feature maps in half. The pooling layer produces an output with a shape of (29, 32). The next layer is called Flatten. It enables

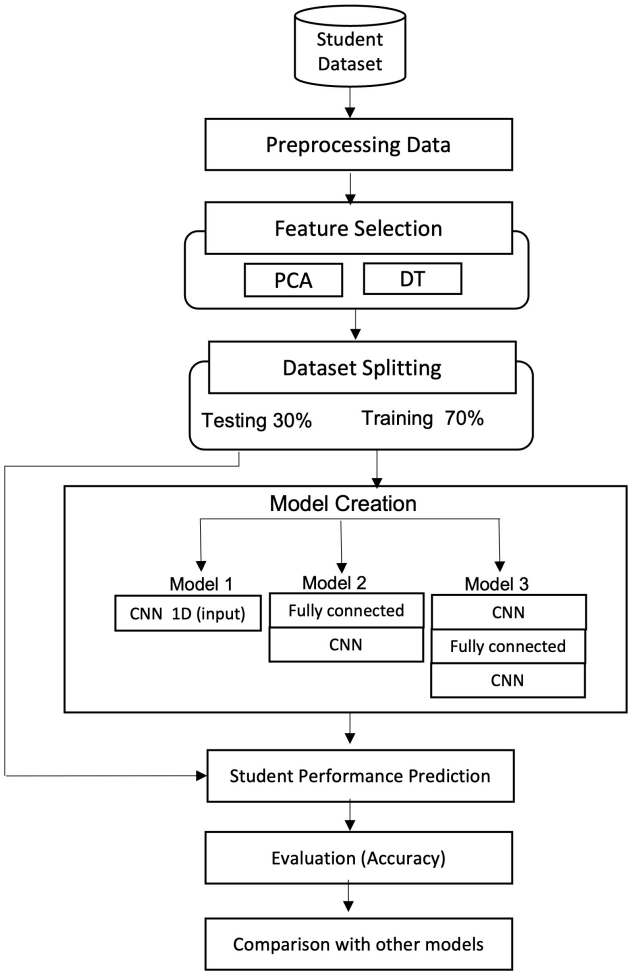

Figure 1. The structure of the proposed model

the processing of the output by traditional, fully connected layers. The following layer used a fully connected layer with 64 neurons and a rectifier activation function. The last layer is the output layer, which has three neurons representing the three categories' classes and a SoftMax activation function. The optimization technique Adaptive Moment Estimation (Adam) is applied to calculate the adaptive momentum value. The error with sparse categorical cross entropy is used as a loss function. We used a batch size of 32 and a different number of epochs 10,50,100, and 200.

2) Model 2: Fully connected, followed by CNN

The second model starts with the input layer with 32 neurons, followed by three fully connected layers, each with 64 neurons with ReLU as the activation layer. The next layer is called Conv1D, which is a convolutional layer with 16 feature maps of size 3×3. The following layer is a max pooling layer, which takes the max over 2×2 patches. The next layer is a second convolutional layer with 16 feature maps of size 3×3. After the convolutional layer, a max pooling layer was added with a pool size of 2×2.

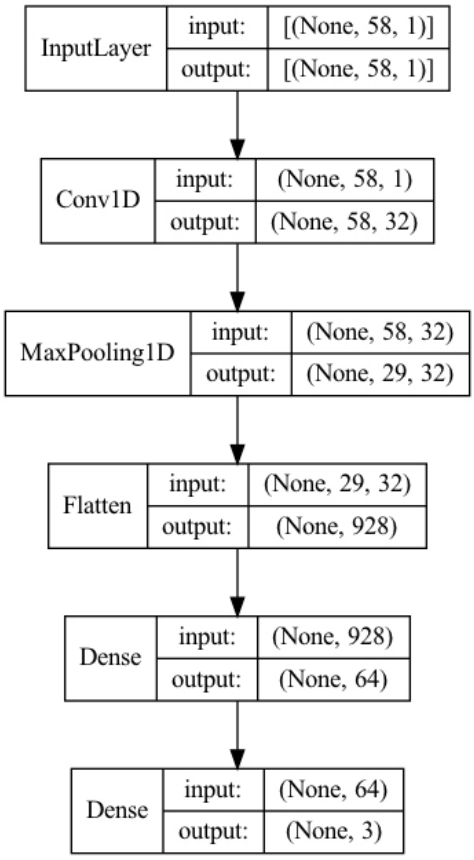

Figure 2. Model 1 structure

TABLE II. Model 1 layer's description

Layer (type)	Output Shape	Param#
conv1d (Conv1D)	(None, 58, 32)	128
maxpooling1d (MaxPooling1D)	(None, 29, 32)	0
flatten (Flatten)	(None, 928)	0
dense (Dense)	(None, 64)	59456
dense1 (Dense)	(None, 3)	195
Total params: 59,779 Trainable params: 59,779 Non-trainable params: 0		

The last hidden layer is a the flatten layer. The output layer has 3 neurons corresponding to the 3 output classes. We used the ReLU function for the hidden layers and the SoftMax function for the output layer. Figure 3 illustrates the structure of Model 2, and Table III summarises the model layer and number of parameters in each layer.

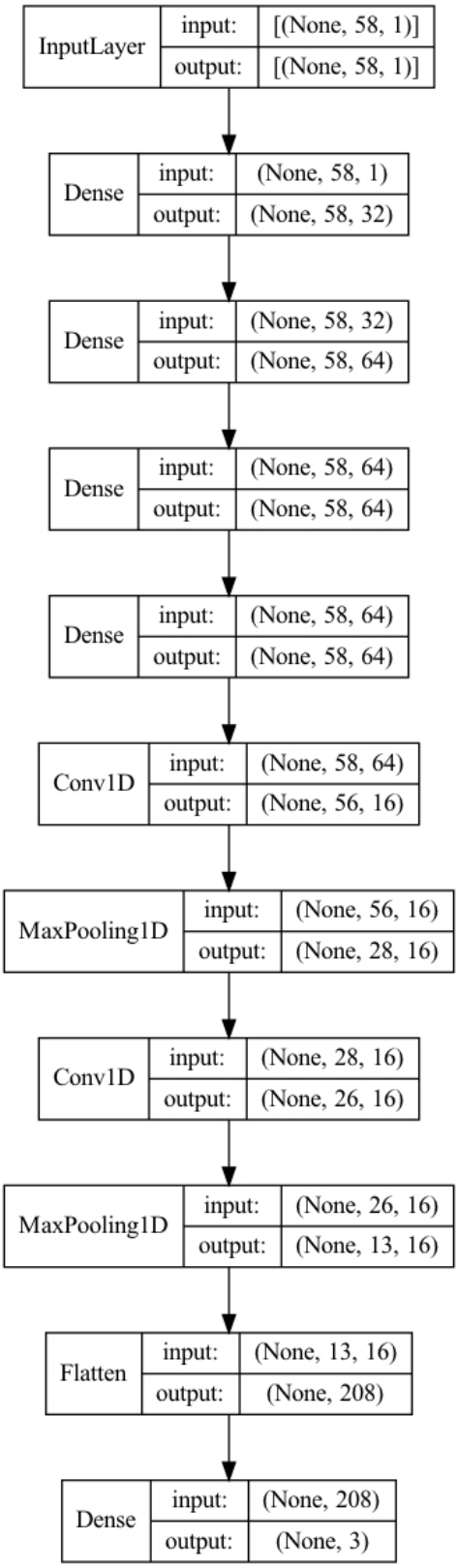

Figure 3. Model 2 structure

TABLE III. Model 2 layer's description

Layer (type)	Output Shape	Param#
dense1 (Dense)	(None, 58, 32)	64
dense2 (Dense)	(None, 58, 64)	2112
dense3 (Dense)	(None, 58, 64)	4160
dense4 (Dense)	(None, 58, 64)	4160
conv1d1 (Conv1D)	(None, 56, 16)	3088
maxpooling1d1 (MaxPooling)	(None, 28, 16)	0
conv1d2 (Conv1D)	(None, 26, 16)	784
maxpooling1d2 (MaxPooling)	(None, 13, 16)	0
flatten1 (Flatten)	(None, 208)	0
dense5 (Dense)	(None, 3)	627
Total params: 14,995 Trainable params: 14,995 Non-trainable params: 0		

3) Model 3: CNN, followed by Fully connected, followed by CNN

Model 3 consists of 14 layers: an input layer, two convolutional layers, three fully connected layers, two more convolutional layers, a flatten layer, and the last, the output layer. The Model starts with the input layer that contains 32 neurons. The output shape from the input layer is (58,32) with 64 parameters. The following layer is a one-dimension convolutional layer with 16 filters of size 3×3 and ReLU as an activation function. Then, a pooling layer of pool size of 2×2. Next, a second convolutional layer with 16 filters of size 3×3 followed by a max pooling layer size of 2×2. The output shape from the convolutional layers is (15,16). After that, three fully connected layers with different numbers of neurons 32, 16, and 16, respectively, with ReLU activation function, are added to the model.

The next layer is a one-dimension convolutional layer with 16 neurons of size 3×3 followed by a max pooling layer. In addition, another one-dimension convolutional layer followed by a pooling layer are added. The final layer is the output layer which, has 3 neurons and SoftMax as the activation function. For all the hidden layers in the model, we used the ReLU function. Figure 4 and Table IV summarize the structure of Model 3.

4. Dataset Description

A. Dataset 1 Description

This dataset approaches student performance in two Portuguese secondary schools. Two datasets are available in two subjects: Mathematics and Portuguese language. It was retrieved from the website of the UCI machine learning repository [24]. The data was gathered via school reports and surveys, including information about a student's grade, demographics, and social background. The dataset

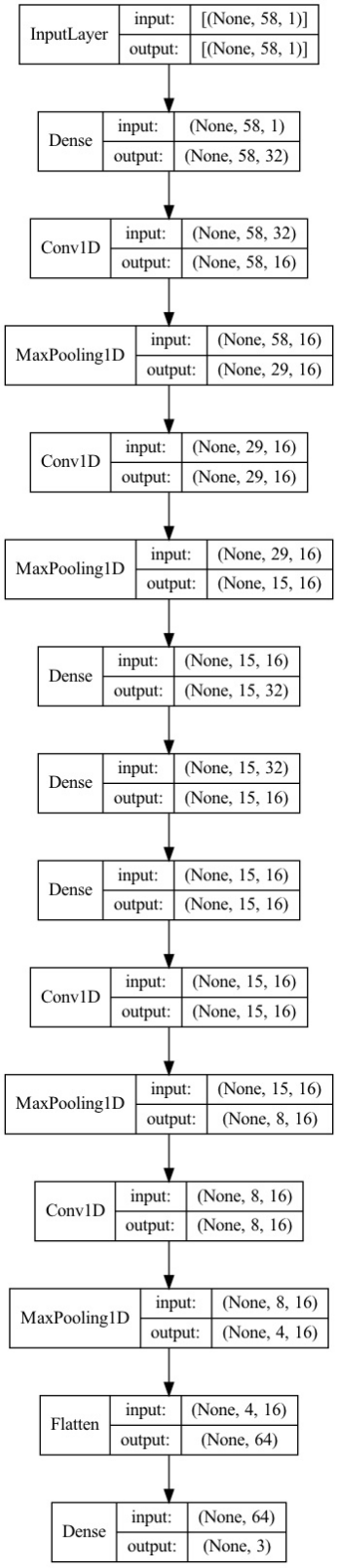

Figure 4. Model 3 structure

TABLE IV. Model 3 layer's description

Layer (type)	Output Shape	Param#
dense1 (Dense)	(None, 58, 32)	64
conv1d1 (Conv1D)	(None, 58, 16)	1552
maxpooling1d1 (MaxPooling)	(None, 29, 16)	0
conv1d2 (Conv1D)	(None, 29, 16)	784
maxpooling1d2 (MaxPooling)	(None, 15, 16)	0
Dense2 (Dense)	(None, 15, 32)	544
dense3 (Dense)	(None, 15, 16)	528
dense4 (Dense)	(None, 15, 16)	272
conv1d3 (Conv1D)	(None, 15, 16)	784
maxpooling1d3 (MaxPooling	(None, 8, 16)	0
conv1d4(Conv1D)	(None, 8, 16)	784
maxpooling1d4 (MaxPooling	(None, 4, 16)	0
flatten (Flatten)	(None, 64)	0
Dense5 (Dense)	(None, 3)	195
Total params: 5,507 Trainable params: 5,507 Non-trainable params: 0		

was represented using three-level classification tasks [25]. It is a three-class dataset (Good, Fair, and Bad) based on the final grade. The dataset consists of 1044 observations with 33 attributes. More details about the dataset used, and its attributes are presented in Table V and Table VI.

TABLE V. Dataset 1 description

Dataset Features	Multivariate
Attribute Types	Integer
Related Tasks	Classification
No. of Instances	1044
No. of Attributes	33
Missing Values?	N/A
Domain	Social
Date Donated	27-11-2014

B. Dataset 2 Description

This educational dataset was obtained from the Kalboard 360 learning management system (LMS) [26]. The dataset contains 480 instances and 16 attributes. The attributes are divided into three primary categories: (a) Demographic characteristics such as nationality and gender. (b) Educational background characteristics include educational stage, grades, and section. (c) Behavioral characteristics include raising hands in class, opening resources, responding to

parent surveys, and school satisfaction. It is a three-class dataset in which learners are assigned to one of three numeric intervals based on their grades (Low-Level, Middle-Level, High-Level). The description and the attributes information of dataset 2 are shown in Table VII and Table VIII.

C. Data preprocessing

Data preprocessing is necessary after data collection to enhance dataset quality. Data preparation includes data attribute selection, cleansing, transformation, and reduction. It contributes to the process of knowledge discovery. Data transformation [27] is described as the technical process of transforming data from one format, standard, or structure to another without affecting the content of the data. It can be evaluated to help decision-making processes and improve data quality. Both datasets were transformed into a numeric format. The attributes of dataset 1 student's school, sex, address type, family size, parents' status, school support, family support, activities, nursery school, higher education, home internet, and romantic relationship were converted to binary '0' and '1'. Other nominal attributes include the mother and father's jobs, the reason for choosing this school, and the student's guardians' transition into a numerical data type. Additionally, integer numbers were created from the output feature. Good is equal to two, Fair to one, and Poor to zero. The second dataset's characteristics, such as the gender of the student, the responsible parents, school year semester, parent responding to questioners, and parent satisfaction, are also transformed into binary data, '0' and '1'. Country of origin, birthplace, level of education, grade of the student, student classroom, and course subject are additional nominal data type features that have been converted to numerical data types.

We noticed a highly imbalanced dataset when we performed the discretization procedure to dataset 1, and the distribution of the students' class labels was inconsistent. The class label contains much more instances for the class "Fair" but fewer samples for classes "Good" and" Poor". Figure 5 illustrates the issue with a significantly imbalanced dataset. The class labels (Good, Fair, Poor) had corresponding distributions of 20%, 65%, and 15%, respectively. This serious difficulty manifests in categorization issues and decreases the model's performance. One of the most crucial aspects of enhancing the models' performance is addressing the problem of an unbalanced dataset. Due to this issue, the dominant class dominates the minority class. As a result, the classifier's performance is unreliable, and they frequently fall into the majority class [11]. Therefore, we must find a solution to this issue as it might provide incorrect outcomes [28].

To address the issue of an imbalanced dataset, we applied resampling techniques. Resampling falls into three categories: oversampling, under-sampling [29], and hybrid sampling [30]. As the datasets used in this experiment are small, oversampling techniques are ideal for handling imbalanced data. It creates additional instances to enhance the number of minority classes in the dataset [31]. The Synthetic Minority Oversampling Technique (SMOTE) method was employed in this experimentation [32].

To investigate the effectiveness of employing balanced data on model performance, we also evaluated how well our models performed on unbalanced data versus other balanced datasets. On the other hand, the class label for the second dataset was balanced (Low, Middle, High) and had corresponding distributions of 44%, 26%, and 30%, respectively.

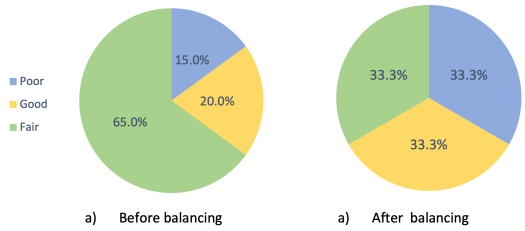

Figure 5. Visualization graph for balancing dataset 1

D. Feature selection

Feature selection [33] is vital for increasing classifier accuracy, saving data-collecting effort, improving model interpretability, and shortening prediction time. Feature importance scores are crucial in a predictive modeling task because they give an understanding of the data and the models [34]. The basis for dimensionality reduction and feature selection may enhance a predictive model's effectiveness and performance on the problem.

This experiment applied two feature selection techniques to both datasets, Principal Component Analysis (PCA) and Decision Trees (DTs). These two approaches are used independently from each other to determine which one of the approaches incorporates improving the accuracy of the classification. The principal component analysis is an approach to minimize the dimensionality of such data, improving interpretability while preventing information loss [35]. It helps identify the most critical attributes in a dataset and simplifies data visualization in 2D and 3D. PCA helped to identify a sequence of linear variable combinations [36].

The second feature selection technique is the decision tree. DT is a commonly used technique in machine learning and data mining [37]. ID3, C4.5, and CART are the traditional decision tree construction methods [38]. C4.5 improves ID3 by avoiding biased attributes. CART can handle features with more values [39][40]. The feature selection process is the procedure for creating a decision tree. Each feature is computed by particular standards [38]. The decision tree algorithm has several key advantages, including adequate classification accuracy and high reliability. We employed the decision tree classifier (CART method) in selecting features, and it was implemented in the Scikit-Learn library [34]. The model offers a feature importance

TABLE VI. Attributes information of dataset 1

No	Attribute	Attribute Description	Type	Range of Values
1	school	School name	binary	GP or MS
2	gender	Students gender	binary	F or M
3	age	Students age	numeric	15, 16,17, 18, 19, 20, 21, 22
4	address	Home address type	binary	Urban (U) or Rural (R)
5	famsize	Size of family	binary	LE3 if < 3 or GT3 > 3
6	Pstatus	Cohabitation status of parents	binary	(T) live with parents or (A) not living together
7	M_educ	Education of mother	numeric	No school (0) Elementary schooling up to the fourth year (1) Grades 5 through 9 (2) Secondary schooling (3) University level education (4)
8	F_educ	Education of father	numeric	No school (0) Elementary schooling up to the fourth year (1) Grades 5 through 9 (2) Secondary schooling (3) University level education (4)
9	M_job	Mother's work	nominal	schoolteacher, medical, public service, in the house, other
10	F_job	Father's work	nominal	schoolteacher, medical, public service, in the house, other
11	reason	Chosen school due to	nominal	near to home, school status, course preferences, or other
12	guardian	Guardian of the learner	nominal	mother, father, other
13	traveltime	Time(home-school)	numeric	<15m (m for minute) 15m to 30m 30m to 1h (h for hours), >1 h
14	studytime	Weekly study time	numeric	<2 h (h for hours), 2 - 5 h, 5 - 10 h, >10 h
15	failures	Previous course failures	numeric	n if 1<=n<3, else 4
16	schoolsup	Further educational assistance	binary	yes or no
17	famsup	Family educational assistance	binary	yes or no
18	paid	Additional paid courses within the course	binary	yes or no
19	activities	Activities outside of the classroom	binary	yes or no
20	nursery	Went to nursery school	binary	yes or no
21	higher	Intends to pursue further education	binary	yes or no
22	internet	Home Net access	binary	yes or no
23	romantic	Involving a relationship	binary	yes or no
24	famrel	family relationship quality	numeric	1 (extremely poor) to 5 (extremely high)
25	freetime	After-school free hours	numeric	1 (extremely poor) to 5 (extremely high)
26	goout	Interacting out with friends	numeric	1 (extremely poor) to 5 (extremely high)
27	Dalc	Alcohol usage at workday	numeric	1 (extremely poor) to 5 (extremely high)
28	Walc	Alcohol drinking throughout the weekend	numeric	1 (extremely poor) to 5 (extremely high)
29	health	Present state of health	numeric	1 (extremely poor) to 5 (extremely high)
30	absences	Amount of absences from school	numeric	0 to 93
31	G1	First semester score	numeric	0 to 20
32	G2	Second semester score	numeric	0 to 20
33	G3	Final score	numeric	0 to 20, output target

TABLE VII. Dataset 2 description

Dataset Features	Multivariate
Attribute Types	Integer/Categorical
Related Tasks	Classification
No. of Instances	480
No. of Attributes	16
Missing Values?	N/A
Domain	Education and Data Mining
Date Donated	8-11-2016

property that can be used to get the relative importance scores for each input feature once it has been fitted.

5. Model Evaluation

Evaluation metrics were used to assess the trained classifier's model performance. In this study, we evaluated the performance of the models using four evaluation metrics: accuracy, precision, recall, and F1-score, as shown in Table IX. Accuracy is commonly used to assess classifier generalization capacity. The trained classifier's accuracy is determined using total correctness, which refers to the total number of occurrences correctly predicted by the trained classifier. In general, the accuracy metric computes the ratio of accurate predictions to the total number of instances investigated. Precision is the second evaluation metric used, which is specified as the proportion of correctly predicted positive patterns in a class to all predicted positive patterns. The third metric used is Recall, which determines the proportion of positive patterns that are identified correctly. F-Measure is also used as an evaluation metric, which shows the harmonic mean of recall and precision levels [41].

6. Experimental Results

This section describes the outcomes of the three proposed models using dataset 1 and dataset 2. The sample data were divided into training data (70%) and testing data (30%) for both datasets.

A. Dataset 1 Results

The experiment was repeated in different cases for the first dataset, as shown in Table X. Cases 1, 2, and 3 with all data, 1044 instances for both subjects (Math and Portugal), and cases 4,5,6,7,8 and 9 with only one subject (Portugal) with 649 instances and 33 attributes.

1) Case 1: All instances (1044) without any feature selection

Table XI shows the results of the experiment using the entire dataset for each model. Comparing the results of the three models, Model 2 achieved the highest accuracy of 89.81% when the number of epochs reached 100. Precision, Recall, and F score were high with results of 0.8922,0.9479, and 0.9192, respectively. Furthermore, the highest accuracy of Model 3 when the number of epochs is 50 was 88.85% which is higher than that of Model 1 when the number of epochs is 100, 87.68%.

2) Case 2: All instances (1044) with DT as a feature selection

The experiment was repeated after applying a decision tree (DT) for feature selection. Table XII illustrates the accuracy of each model with DT. We can observe from the results that the highest accuracy was 89.81% when we used a CNN with one dimension as the model's input. Comparing the outcome of the proposed models with feature selection and without using any features, we can see that almost all accuracies were obtained from models with feature selection more than the ones obtained without using any feature. However, the highest accuracy from Model 2 was 89.81% ,which is equal to the highest accuracy when using Model 1 when CNN is used as the model's input.

3) Case 3: All instances (1044) with PCA as a feature extraction

The second feature extraction used in this examination is PCA. We tried to extract 40 features from dataset 1. Table XIII displays the results of the experiment using PCA as feature extraction. All models acquired polite accuracy results, and Model 2 reached the highest of 88.22% with 200 epochs. Furthermore, Table XIV compares the accuracies of dataset 1 when applied feature selection and without any features. We can observe from the results that the accuracies obtained from the models with DT as feature selection was better than PCA. The highest accuracy was 89.81% with DT and 88.22% with PCA. Overall, the accuracies when applied DT with Model 1 are higher than the accuracies obtained from the same model without any feature selection.

4) Case 4: Part of the dataset (649 instances) without any feature selection

We employed the experiment using part of the imbalanced dataset, selecting only one subject with 649 instances. From Table XV, we can observe that the highest accuracy of 94.36% when the model was built using fully connected layers, followed by CNN with 100 epochs. Furthermore, the results of the accuracies obtained from all models were high, between 89%- 94%.

5) Case 5: Part of the dataset (649 instances) with DT as a feature selection

The investigation was replicated using only 649 instances from all datasets after selecting features using decision trees. Table XV indicates the results of the accuracy of the three models with DT as feature selection and without them. The highest score obtained reached 87.69% when using CNN as the input layer to Model 1 after selecting features from the dataset.

6) Case 6: Part of the dataset (649 instances) balanced without DT

In addition, we tried to oversample the dataset, as shown in Table XV. The highest accuracy for the balanced dataset

TABLE VIII. Attributes information of dataset 2

No	Attribute Name	Description	Type	Range of Values
1	Gender	Student sex	nominal	M or F
2	Nationality	Student's nationality	nominal	LB, EG, KSA, USA, JOR, VEN, IRAN, TUN, SYR, LYB, KUW, MOR, IRQ, PAL
3	Placeofbirth	Place of Birth	nominal	LB, EG, KSA, USA, JOR, VEN, IRAN, TUN, SYR, LYB, KUW, MOR, IRQ, PAL
4	StageID	Educational level	nominal	L (lower), M (middle), H (high)
5	GradeID	Student grades	nominal	From G1 to G12
6	SectionID	Student class	nominal	A,B,C
7	Topic	Course subject	nominal	Eng, Span, Fren, Ar, IT, Math, Chem, Bio, Sci, His, Quran, Geo
8	Semester	School year semester	nominal	1st or 2nd
9	Relation	Responsibility of the student	nominal	Mother or Father
10	Raisedhands	how often does a student raise their hand in class	numeric	0 -100
11	VisITedResources	how frequently a student accesses course material	numeric	0 -100
12	AnnouncView	how often the student looks at the most recent announcements	numeric	0 -100
13	Discgroups	how frequently a student takes part in discussion groups	numeric	0 -100
14	PAnsweringSurvey	if the parent responded to the questionnaires supplied by the school	nominal	Yes or NO
15	PschoolSatisfi	how satisfied parents are with the school	nominal	Yes or NO
16	StudentAbsenDays	Students absents	nominal	more than 7, less than 7
17	Class	Student level	nominal	L, M, H

TABLE IX. Evaluation metrics used in the proposed model

Metrics	Formula
Accuracy (Acc)	(TP + TN) / TS
Precision (PR)	TP / (TP + FP)
Recall (RE)	TP / (TP + FN)
F1-score (FS)	2 x ((PR * RE)/(PR + RE))

TP: predicted Yes, and actual output was Yes.
TN: predicted NO, and actual output was NO
FP: predicted YES, and actual output was NO.
FN: predicted NO, and actual output was YES.
TS: Total number of samples

without any feature selection reached 89.74% when we applied Model 2 with 100 epochs. Moreover, the accuracies from Model 3 are near the highest score, which reached 88.21% and 88.72% with epochs 50 and 100, respectively.

7) Case 7: Part of the dataset (649 instances) balanced with DT

In case 7, we used 649 instances from the balanced dataset with DT as a feature selection as exhibited in Table XV. From the outcomes, we can observe that all the accuracies obtained decreased compared to the same dataset without feature selection. The highest accuracy was 86.67% using Model 2 and Model 3 with 100 epochs.

8) Case 8: Part of the dataset (649 instances) imbalanced with PCA

For the feature extraction part, we applied PCA to the part of the imbalanced dataset. The result indicates that the highest accuracy obtained was 85.13% from Model 1, when using CNN as an input layer. From Table XV, Model 1 achieves the highest accuracy; either we applied DT or PCA as feature selection.

TABLE X. Different cases for the dataset 1

Cases	Instances	With DT	With PCA	imbalanced	Accuracy%	Best proposed Model
Case 1	1044			✓	89.81	Model 2
Case 2	1044	✓		✓	89.81	Model 1
Case 3	1044		✓	✓	88.22	Model 2
Case 4	649			✓	94.36	Model 2
Case 5	649	✓			87.69	Model 1
Case 6	649				89.74	Model 2
Case 7	649	✓			86.67	Model 2,3
Case 8	649		✓	✓	85.13	Model 1
Case 9	649		✓		88.21	Model 1,3

9) Case 9: Part of the dataset (649 instances) balanced with PCA

In the last case in our experiment, we tried to extract features from 649 instances after resampling the dataset. Table XV showed that the highest accuracy was 88.21% when using CNN as an input layer with 50 epochs or using CNN followed by fully connected followed by CNN with 100 epochs.

Table XVI summarizes the results of the accuracies of dataset 1 with feature selections and without any feature selection. Comparing the outcomes when we used all the datasets and a part of them, the accuracy from 649 instances was higher than the accuracy obtained when we used 1044 instances from the dataset. Moreover, the performance of models when we used an imbalanced dataset was greater than the performance of balanced dataset. The highest accuracy for the imbalanced dataset was 94.36% and 89.74% for the balanced. In addition, the accuracies with feature selection were lower than that with DT as feature selection. However, using PCA as a feature selection for a balanced dataset gives higher accuracies than using DT as feature selection.

B. Comparison with previous studies using dataset 1

This study produced results that corroborate the findings of many of the earlier studies in this field. Table XVII compares our model's performance to previous works. All these studies, which are shown in Table XVII were used dataset 1. In 2017 [42], researchers used Decision Stump to predict the academic performance of students with 649 instances of the dataset. The accuracy obtained was 90.81% ,which is good. Furthermore, a study in 2019 [9] used ensemble techniques (J48 and Realada- boost) to predict student achievement with an accuracy of 95.78% using all instances of the dataset.

A recent study by Yousafzai [12] used a deep learning algorithm, such as a convolutional neural network, to get an accuracy of 85.56%. Several previous researchers performed

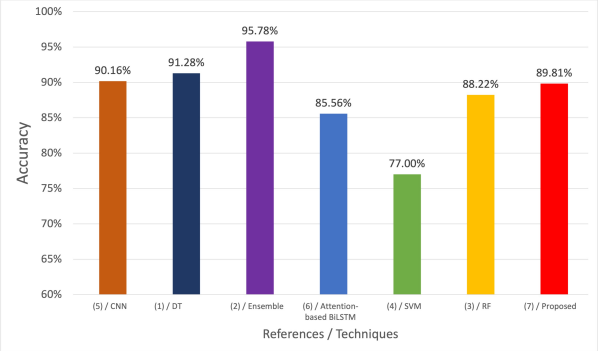

Figure 6. Accuracy comparison between the proposed model and previous studies for all instances of dataset 1

a similar series of experiments using different machine learning algorithms to obtain good accuracy results. A work by Ghorbani et al. [11] used random forest, and another study [7] predicted academic performance using a support vector machine with 88.22% and 77% accuracy, respectively.

Our model performed well compared with all the models mentioned in this paper, with an accuracy rate of 89.81% if we used the entire dataset with 1044 instances, 89.74% if we used 649 instances and applying the second model (fully connected followed by CNN) after oversampling, and 94.36% if the dataset was imbalanced. Our model got good findings compared to the baseline model and models from other researchers, as shown in Figure 6 and Figure 7. This difference depends on how the other studies dealt with the datasets, how they pre-processed the data, what algorithms were used, and how they evaluated their models and got their results. In addition, we compared the results of the proposed model accuracy with previous studies applying the CNN model; it performed better than them. As a result, according to our study topic, the proposed CNN models succeed in predicting student academic performance with educational datasets.

TABLE XI. Performance evaluation of dataset 1 (1044) without any feature selection

Models	No. Epochs	Accuracy %	Accuracy	Precision	Recall	F score
Model 1 CNN as the input	10	85.67	0.8567	0.9061	0.8542	0.8794
	50	87.58	0.8758	0.8923	0.9062	0.8992
	100	87.68	0.8768	0.8732	0.9323	0.9018
	200	86.94	0.8694	0.8802	0.9037	0.8918
Model 2 fully connected - CNN	10	86.62	0.8662	0.8989	0.8802	0.8895
	50	88.85	0.8885	0.8792	0.9479	0.9123
	100	89.81	0.8981	0.8922	0.9479	0.9192
	200	85.03	0.8503	0.8156	0.8750	0.8398
Model 3 CNN - Fully connected - CNN	10	86.94	0.8694	0.8646	0.9171	0.8901
	50	88.85	0.8885	0.9062	0.9110	0.9086
	100	86.62	0.8662	0.8854	0.8947	0.8901
	200	85.99	0.8599	0.8442	0.8554	0.8427

TABLE XII. Comparison of accuracy of dataset 1 (1044) with DT as feature selection and without any feature selection

Models	No. Epochs	Accuracy%	Accuracy% with DT
Model 1 CNN as the input	10	85.67	84.39
	50	87.58	89.81
	100	87.68	89.49
	200	86.94	88.22
Model 2 fully connected - CNN	10	86.62	81.53
	50	88.85	86.94
	100	89.81	86.94
	200	85.03	85.99
Model 3 CNN - Fully connected - CNN	10	86.94	85.67
	50	88.85	87.58
	100	86.62	88.85
	200	85.99	85.67

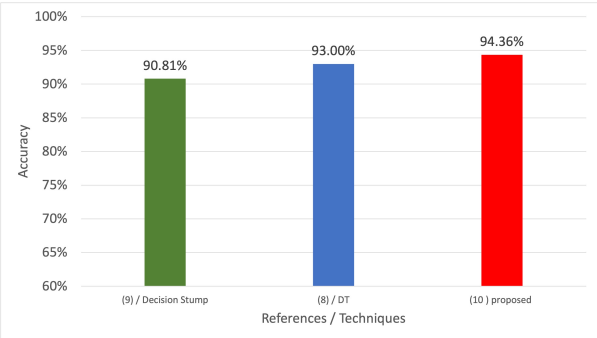

Figure 7. Comparison of proposed and previous models' accuracy for a portion of dataset 1

C. Dataset 2 Results

For the second dataset as presents in Table XVIII, we have three cases as the follows:

1) Case 1: Without feature selection

The highest accuracy obtained was 83.45% when we used CNN as the model's input, reaching 200 epochs. As shown in Table XIX, the accuracies of Model 3 are higher than that of Model 2. For instance, the accuracy received from Model 3 was 80.13%, with 100 epochs, while the highest accuracy achieved from Model 2 was 72.52% with 200 epochs.

TABLE XIII. Accuracy of dataset 1 (1044) with PCA as feature selection

Models	No. Epochs	Accuracy% with PCA
Model 1 CNN as the input	10	85.03
	50	85.35
	100	87.26
	200	84.39
Model 2 Fully connected - CNN	10	82.09
	50	86.31
	100	86.62
	200	88.22
Model 3 CNN - Fully connected - CNN	10	81.85
	50	85.03
	100	83.44
	200	80.89

TABLE XIV. Comparison of accuracy of dataset 1 (1044) with DT, PCA as feature selection and without feature selection

Models	No. Epochs	Accuracy%	Accuracy% with DT	Accuracy% with PCA
Model 1 CNN as the input	10	85.67	84.39	85.03
	50	87.58	89.81	85.35
	100	87.68	89.49	87.26
	200	86.94	88.22	84.39
Model 2 Fully connected - CNN	10	86.62	81.53	82.09
	50	88.85	86.94	86.31
	100	89.81	86.94	86.62
	200	85.03	85.99	88.22
Model 3 CNN - Fully connected - CNN	10	86.94	85.67	81.85
	50	88.85	87.58	85.03
	100	86.62	88.85	83.44
	200	85.99	85.67	80.89

2) Case 2: With DT as a feature selection

We repeated the experiment with DT as a feature selection. The highest score was 84.83% when using CNN as the model's input. Furthermore, we can observe that the accuracies achieved from Model 1 with different epochs are higher than those from Model 2 and Model 3. However, Model 3 accuracies are higher than those of Model 2. The highest accuracy was 80.21% for Model 3 and 70.83% for Model 2.

3) Case 3: With PCA as feature extraction

In the third case we applied feature extraction using PCA. The highest accuracy obtained was 82.29% when CNN was used as the model input. For the other models, the performance of Model 3 was higher than that of Model 2. In addition, when the number of epochs is 100, the accuracy of the Model 3 is 81.25%, and 77.08% for the Model 2. If we compare the accuracy of the three models as shown in Table XIX, the highest accuracy results when we use CNN as the input for Model 1. It reaches 84.83% when using DT as a feature selection with 50 epochs and 83.45% without feature selection.

D. Comparison with previous studies using dataset 2

Comparative analysis was done with the previous studies that used dataset 2, as shown in Table XX. In 2015 a paper

TABLE XV. Comparison of accuracy of dataset 1 (645) with feature selection and without feature selection

		Imbalanced			Balanced		
Model	No. Epochs	Accuracy%	Accuracy% with DT	Accuracy% with PCA	Accuracy%	Accuracy% with DT	Accuracy% with PCA
Model 1 CNN as the input	10	90.77	84.62	66.67	85.13	84.62	70.26
	50	89.74	87.69	81.03	88.72	84.62	88.21
	100	91.28	87.18	80.00	83.08	85.64	87.18
	200	92.31	87.69	85.13	86.67	86.15	87.18
Model 2 fully connected - CNN	10	89.74	86.15	66.15	82.56	83.08	81.54
	50	90.77	86.15	70.26	87.18	86.67	84.10
	100	94.36	86.59	73.33	89.74	85.13	85.13
	200	90.26	85.64	73.85	86.15	82.05	85.15
Model 3 CNN - Fully connected - CNN	10	93.85	85.13	69.74	86.61	82.56	83.59
	50	91.79	86.15	70.26	88.21	81.03	86.67
	100	92.31	86.67	71.79	88.72	86.67	88.21
	200	90.77	87.18	72.82	85.13	85.64	86.15

TABLE XVI. Comparison of accuracy of dataset 1 with a feature selection and without any feature selection

		imbalanced (1044 instances)			imbalanced (649 instances)			Balanced (649 instances)		
Models	No. Epochs	Accuracy%	Accuracy% with DT	Accuracy% with PCA	Accuracy%	Accuracy% with DT	Accuracy% with PCA	Accuracy %	Accuracy% with DT	Accuracy% withPCA
Model 1 CNN as the input	10	85.67	84.39	85.03	90.77	84.62	66.67	85.13	84.62	70.26
	50	87.58	89.81	85.35	89.74	87.69	81.03	88.72	84.62	88.21
	100	87.68	89.49	87.26	91.28	87.18	80.00	83.08	85.64	87.18
	200	86.94	88.22	84.39	92.31	87.69	85.13	86.67	86.15	87.18
Model 2 fully connected - CNN	10	86.62	81.53	82.09	89.74	86.15	66.15	82.56	83.08	81.54
	50	88.85	86.94	86.31	90.77	86.15	70.26	87.18	86.67	84.10
	100	89.81	86.94	86.62	94.36	86.59	73.33	89.74	85.13	85.13
	200	85.03	85.99	88.22	90.26	85.64	73.85	86.15	82.05	85.15
Model 3 CNN - Fully connected - CNN	10	86.94	85.67	81.85	93.85	85.13	69.74	86.61	82.56	83.59
	50	88.85	87.58	85.03	91.79	86.15	70.26	88.21	81.03	86.67
	100	86.62	88.85	83.44	92.31	86.67	71.79	88.72	86.67	88.21
	200	85.99	85.67	80.89	90.77	87.18	72.82	85.13	85.64	86.15

by Amrieh et al. [18] utilized the same dataset and three distinct machine learning algorithms in their study: Decision Tree, Naive Bayes, and Artificial Neural Networks. Their results indicated that the accuracies were 61.30% for DT, 73.8% for ANN, and 72.5% for NB. One year later, the same authors [19] used ensemble techniques to predict the performance and got an accuracy of 79.1%. Another work by Pujianto et al. [43] analyzed the effectiveness of two classifiers, C4.5 and KNN, using the SMOTE preprocessing approach. The C4.5 decision tree technique produced improved prediction performance in experiments with accuracy, recall, and precision scores of 71.09%, 71.63%, and 71.54%, respectively. Moreover, in 2021, a publication [44] constructed a prediction model using varies machine learning techniques on the entire dataset. In addition, they implemented several ensemble meta-based models that were integrated with ML algorithms for classifying data. For instance: Bagging, AdaBoostM1, and RandomSubSpace. The results showed that the Multilayer Perceptron Machine Learning approach had up to 80.33% accuracy performance using the ensemble meta-based technique (AdaBoostM1). Compared to all the models mentioned in previous studies, our model performed better, with an accuracy rate of 84.83% as presented in Figure 8. As a result, Model 1 using CNN with one dimension as the input and DT as feature extraction can be used to predict student performance

TABLE XVII. Comparison of accuracy of previous studies using dataset 1 and the proposed model

No	Ref	Year	Techniques	No. of instances	Accuracy
1	[17]	2016	DT	1044	91.28%
2	[9]	2019	Ensemble (J48, Realada- boost)	1044	95.78 %
3	[7]	2021	SVM	1044	77.00%
4	[11]	2020	RF	1044	88.22%
5	[12]	2021	CNN	1044	85.56 %
6	[12]	2021	Attention-based BiLSTM	1044	90.16 %
7	proposed		fully connected – CNN / CNN as the input	1044	89.81%
8	[25]	2008	DT	649	93.00%
9	[42]	2017	Decision Stump	649	90.81 %
10	proposed		fully connected – CNN	649	94.36%

TABLE XVIII. Different cases for dataset 2

Cases	With DT	With PCA	Accuracy%	Best proposed Model
Case 1			83.45	Model 1
Case 2	✓		84.83	Model 1
Case 3		✓	82.29	Model 1

successfully.

7. Discussion

From the experiment, the best algorithms for the first dataset compared to other classifiers is Model 2, using fully connected layers followed by CNN. The accuracy obtained using all instances was 89.81% for imbalanced without feature selection, 86.94% with DT, and 88.22% with PCA as feature selection. Moreover, using part of dataset 1 with imbalanced reached accuracy of 94.36% without feature selection and 86.59% with DT as feature selection and 85.13% with PCA as feature selection. However, after resampling the dataset, the accuracy was 89.74% without feature selection and 86.67% with DT as feature selection, and 88.21% with PCA. In addition, the other evaluation metrics results for Model 2 were higher than other models, the obtained precision, recall, and F1 score were 0.892, 0.948, and 0.919, respectively. The final results from dataset 1 indicate that the best model compared with other proposed models is Model 2, which uses fully connected layers followed by CNN. The second was Model 1 and last was Model 3. Despite that Model 2 does not get the best accuracy in all cases, but it still achieves a higher accuracy than any of the accuracies obtained in the previous studies. This makes Model 2 is a strong architecture for student performance prediction.

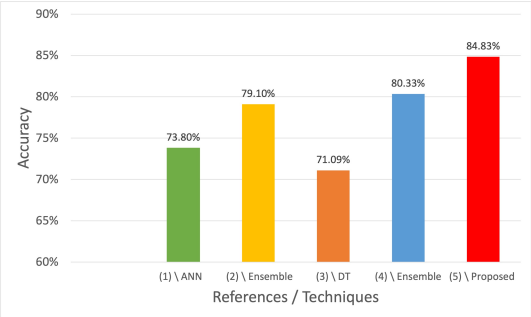

Figure 8. Comparison of accuracy of the proposed model and previous studies for dataset 2

For the second dataset used in this experiment, the best model among the other models is Model 1, using CNN as model input. The accuracy results of Model 1 are 83.45% without any feature selection, 84.83% with DT, and 82.29% with PCA as feature selection. The results obtained indicate that Model 1 has better performance than other previous studies, with an accuracy of 84.83%.

The findings are consistent with other studies showing that machine learning and deep learning techniques are ideal for predicting student performance with more than 84% accuracy. Furthermore, different datasets and machine learning techniques result in various performances. When comparing the performance of the selected datasets, dataset 1 performed better than dataset 2. When comparing the performance of the chosen models, Model 2 performed better than Model 1 in dataset 1 but was lower in dataset 2.

8. Conclusion

Identifying students' future academic progress and improving their grades are major benefits of applying machine learning and deep learning. It can be used for student profile modeling, aiming to produce knowledge from data automatically. This study proposes three models using convolutional neural networks and fully connected layers to predict student success in the year. Model 1 used CNN as the model input, Model 2 used fully connected layers followed by CNN, and Model 3 used CNN followed by fully connected layers followed by CNN. In addition, two datasets available publicly were used. The first dataset was imbalanced, and the second dataset was balanced. We applied two feature selection techniques: decision tree and principal component analysis. The model was evaluated with feature selection and without any feature selection. Moreover, different evaluation metrics were employed to measure the performance of each model. Among the models obtained for the first dataset, Model 2 has the best results when using part of the dataset, with an accuracy of 94.36% without any feature selection when the data was imbalanced and 89.74% when the data was balanced. However, for the second dataset, Model 1 obtained the best results either with or without feature selection. It was 84.83% accuracy with DT, 82.29% with PCA, and 83.45% without feature selection. The proposed model's results were compared with previous models, and based on the outcomes, it can be concluded that the suggested models are highly effective in predicting student performance. It can also be valuable in assisting teachers and improving their teaching skills. Based on the promising findings presented in this study, further research in the field of intelligent tutoring

TABLE XIX. Comparison of accuracy of dataset 2 with a feature selection and without any feature selection

Models	No. Epochs	Accuracy%	Accuracy% with DT	Accuracy% with PCA
Model 1 CNN as the input	10	76.16	81.13	81.25
	50	78.25	84.83	80.21
	100	80.52	83.33	81.25
	200	83.45	82.29	82.29
Model 2 fully connected - CNN	10	69.16	65.14	70.83
	50	67.14	67.71	72.92
	100	71.25	70.83	77.08
	200	72.52	69.79	71.88
Model 3 CNN - Fully connected - CNN	10	76.13	75.00	72.92
	50	78.45	76.04	76.04
	100	80.13	78.12	81.25
	200	77.12	80.21	79.17

TABLE XX. Using dataset 2, compare the accuracy of earlier research and the proposed model.

No	Ref	Year	Techniques	No. of instances	Accuracy
1	[18]	2015	ANN, NB, DT (best ANN)	480	73.8 %
2	[19]	2016	Ensemble methods	480	79.1 %
3	[43]	2021	DT and KNN	480	71.09%
4	[44]	2021	Ensemble Meta-Based Tree	480	80.33%
5	proposed		CNN with one dimension as the input (DT as feature extraction)	480	84.83%

systems are strongly encouraged, including utilizing more machine learning models with deep learning, investigating new approaches to conduct experiments, and selecting the appropriate features influencing student performance.

References

[1] K. I. M. Ramaphosa, T. Zuva, and R. Kwuimi, "Educational data mining to improve learner performance in gauteng primary schools," in *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*. IEEE, 2018, pp. 1–6.

[2] P. Asthana and B. Hazela, "Applications of machine learning in improving learning environment," *Multimedia Big Data Computing for IoT Applications: Concepts, Paradigms and Solutions*, pp. 417–433, 2020.

[3] T. Hamim, F. Benabbou, and N. Sael, "Survey of machine learning techniques for student profile modeling," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 4, pp. 136–151, 2021.

[4] A. Abyaa, M. K. Idrissi, and S. Bennani, "Towards an adult learner model in an online learning environment," in *2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET)*. IEEE, 2016, pp. 1–5.

[5] S. Singh and S. Vikram, "A novel approach to student profiling in intelligent tutoring systems," *Wesleyan Journal of Research*, vol. 13, no. 27, 2020.

[6] I. Khan, A. R. Ahmad, N. Jabeur, and M. N. Mahdi, "An artificial intelligence approach to monitor student performance and devise preventive measures," *Smart Learning Environments*, vol. 8, no. 1, pp. 1–18, 2021.

[7] H. Ng, A. A. bin Mohd Azha, T. T. V. Yap, and V. T. Goh, "A machine learning approach to predictive modelling of student performance," *F1000Research*, vol. 10, 2021.

[8] R. Liu and K. R. Koedinger, "Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains." *Journal of Educational Data Mining*, vol. 9, no. 1, pp. 25–41, 2017.

[9] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, "Student academic performance prediction using supervised learning techniques." *International Journal of Emerging Technologies in Learning*, vol. 14, no. 14, 2019.

[10] A. Lincke, M. Jansen, M. Milrad, and E. Berge, "The performance of some machine learning approaches and a rich context model in student answer prediction," *Research and Practice in Technology Enhanced Learning*, vol. 16, no. 1, pp. 1–16, 2021.

[11] R. Ghorbani and R. Ghousi, "Comparing different resampling methods in predicting students' performance using machine learning techniques," *IEEE Access*, vol. 8, pp. 67 899–67 911, 2020.

[12] B. K. Yousafzai, S. A. Khan, T. Rahman, I. Khan, I. Ullah, A. Ur Rehman, M. Baz, H. Hamam, and O. Cheikhrouhou, "Student-performulator: student academic performance using hybrid deep neural network," *Sustainability*, vol. 13, no. 17, p. 9775, 2021.

[13] A. Smith, W. Min, B. W. Mott, and J. C. Lester, "Diagrammatic student models: Modeling student drawing performance with deep learning," in *User Modeling, Adaptation and Personalization: 23rd International Conference, UMAP 2015, Dublin, Ireland, June 29–*

July 3, 2015. Proceedings 23. Springer, 2015, pp. 216–227.

[14] A. Siddique, A. Jan, F. Majeed, A. I. Qahmash, N. N. Quadri, and M. O. A. Wahab, "Predicting academic performance using an efficient model based on fusion of classifiers," *Applied Sciences*, vol. 11, no. 24, p. 11845, 2021.

[15] S. Li and T. Liu, "Performance prediction for higher education students using deep learning," *Complexity*, vol. 2021, pp. 1–10, 2021.

[16] S. Poudyal, M. J. Mohammadi-Aragh, and J. E. Ball, "Prediction of student academic performance using a hybrid 2d cnn model," *Electronics*, vol. 11, no. 7, p. 1005, 2022.

[17] A. Hamoud, "Selection of best decision tree algorithm for prediction and classification of students' action," *American International Journal of Research in Science, Technology, Engineering & Mathematics*, vol. 16, no. 1, pp. 26–32, 2016.

[18] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using x-api for improving student's performance," in *2015 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*. IEEE, 2015, pp. 1–5.

[19] E. A. Amrieh, T. Hamtini, and Aljarah, "Mining educational data to predict student's academic performance using ensemble methods," *International journal of database theory and application*, vol. 9, no. 8, pp. 119–136, 2016.

[20] F. Saleem, Z. Ullah, B. Fakieh, and F. Kateb, "Intelligent decision support system for predicting student's e-learning performance using ensemble machine learning," *Mathematics*, vol. 9, no. 17, p. 2078, 2021.

[21] V. Vijayalakshmi and K. Venkatachalapathy, "Comparison of predicting student's performance using machine learning algorithms," *International Journal of Intelligent Systems and Applications*, vol. 11, no. 12, p. 34, 2019.

[22] M. Akour, H. Alsghaier, and O. Al Qasem, "The effectiveness of using deep learning algorithms in predicting students achievements," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 387–393, 2020.

[23] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, pp. 611–629, 2018.

[24] D. Dua and C. Graff, "Uci machine learning repository-student performance," 2017.

[25] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance." EUROSIS-ETI, 2008.

[26] J. Khalifeh, "Kalboard: Amman governorate," 2012. [Online]. Available: https://www.kalboard.com

[27] BasuMallick, "What is data transformation?" 2006. [Online]. Available: https://www.spiceworks.com/tech/big-data/articles/what-is-data-transformation/

[28] S. Kotsiantis, D. Kanellopoulos, P. Pintelas *et al.*, "Handling imbalanced datasets: A review," *GESTS international transactions on computer science and engineering*, vol. 30, no. 1, pp. 25–36, 2006.

[29] M. Bach, A. Werner, and M. Palt, "The proposal of undersampling method for learning from imbalanced datasets," *Procedia Computer Science*, vol. 159, pp. 125–134, 2019.

[30] S. Gazzah, A. Hechkel, and N. E. B. Amara, "A hybrid sampling method for imbalanced data," in *2015 IEEE 12th International Multi-Conference on Systems, Signals & Devices (SSD15)*. IEEE, 2015, pp. 1–6.

[31] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," *Decision Analytics*, vol. 2, pp. 1–25, 2015.

[32] H. Hassan, N. B. Ahmad, and S. Anuar, "Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining," in *Journal of Physics: Conference Series*, vol. 1529, no. 5. IOP Publishing, 2020, p. 052041.

[33] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*. Springer, 2008, vol. 207.

[34] J. Brownlee, "How to calculate feature importance with python," *Machine Learning Mastery. https://machinelearningmastery.com/calculate-feature-importance-with-python*, 2020.

[35] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[36] Biswal, "Principal component analysis in machine learning," 2009. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/principal-component-analysis

[37] H. Sun and X. Hu, "Attribute selection for decision tree learning with class constraint," *Chemometrics and Intelligent Laboratory Systems*, vol. 163, pp. 16–23, 2017.

[38] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, pp. 81–106, 1986.

[39] S. Roy, S. Mondal, A. Ekbal, and M. S. Desarkar, "Crdt: correlation ratio based decision tree model for healthcare data mining," in *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2016, pp. 36–43.

[40] H. Zhou, J. Zhang, Y. Zhou, X. Guo, and Y. Ma, "A feature selection algorithm of decision tree based on feature weight," *Expert Systems with Applications*, vol. 164, p. 113842, 2021.

[41] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[42] S. Akhai, R. Karia, A. Mahadik, A. Shah, and M. Gidwani, "Automated performance evaluation system," 2017.

[43] U. Pujianto, W. A. Prasetyo, and A. R. Taufani, "Students academic performance prediction with k-nearest neighbor and c4. 5 on smote-balanced data," in *2020 3rd international seminar on research of information technology and intelligent systems (ISRITI)*. IEEE, 2020, pp. 348–353.

[44] M. Kumar, G. Mehta, N. Nayar, and A. Sharma, "Emt: Ensemble meta-based tree model for predicting student performance in

academics," in *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1. IOP Publishing, 2021, p. 012062.

Fatema Alshaikh is a Ph.D. candidate in the Computing and Information Science program at the University of Bahrain. In 2012, she completed her Master's degree in computer science from the Open University Malaysia, Bahrain, and in 2005, she graduated with a B.Sc. in Computer Science from the University of Bahrain. Her research interests are mainly focused on artificial intelligence, machine learning, and intelligent systems.

Nabil M. Hewahi is a professor of computer science since 2006. He is currently working at the University of Bahrain, Bahrain. He obtained his B.Sc degree from Al-Fateh University, Libya in 1986, M.Tech degree from the Indian Institute of Technology (IIT), Bombay, India in 1991 and PhD degree from Jawaharlal Nehru University, new Delhi, India in 1994. All in computer Science. The main research focus and interest is in the fields of AI and ML. Prof. Hewahi has published around 100 papers in international journals and conferences.