The Journal of Supercomputing
https://doi.org/10.1007/s11227-024-06087-9

Transforming educational insights: strategic integration of federated learning for enhanced prediction of student learning outcomes

Umer Farooq[1,2] · Shahid Naseem[2] · Tariq Mahmood[3,4] · Jianqiang Li[1,5] · Amjad Rehman[3] · Tanzila Saba[3] · Luqman Mustafa[2]

Accepted: 18 March 2024

Abstract

Numerous educational institutions utilize data mining techniques to manage student records, particularly those related to academic achievements, which are essential in improving learning experiences and overall outcomes. Educational data mining (EDM) is a thriving research field that employs data mining and machine learning methods to extract valuable insights from educational databases, primarily focused on predicting students' academic performance. This study proposes a novel federated learning (FL) standard that ensures the confidentiality of the dataset and allows for the prediction of student grades, categorized into four levels: low, good, average, and drop. Optimized features are incorporated into the training process to enhance model precision. This study evaluates the optimized dataset using five machine learning (ML) algorithms, namely support vector machine (SVM), decision tree, Naïve Bayes, K-nearest neighbors, and the proposed federated learning model. The models' performance is assessed regarding accuracy, precision, recall, and F1-score, followed by a comprehensive comparative analysis. The results reveal that FL and SVM outperform the alternative models, demonstrating superior predictive performance for student grade classification. This study showcases the potential of federated learning in effectively utilizing educational data from various institutes while maintaining data privacy, contributing to educational data mining and machine learning advancements for student performance prediction.

Keywords Prediction · Federal learning · Machine learning · Learning outcome · EDM · SVM · Inclusive innovation

Extended author information available on the last page of the article

Published online: 10 April 2024

1 Introduction

Schooling plays a pivotal role in a nation's socioeconomic progress and advancement. Forecasting the academic performance of students has emerged as a crucial undertaking. There is a pressing requirement for implementing a digital infrastructure that can facilitate the cultivation of a generation equipped with the necessary skills to advance a nation across various domains, including but not limited to scientific, economic, social, and military spheres. One of the primary rationales is that educational institutions rely on the government as the principal entity responsible for exerting substantial and arduous endeavors to facilitate the progression of education in a consistent and accelerated manner by expanding educational accessibility. The initial approach to acquiring comprehension of future events involves the act of making predictions about forthcoming occurrences. The second method, data mining, involves extracting previously unknown information. To enhance its predictive capabilities, data mining can be optimized by utilizing extensive databases and diverse data sources pertaining to various sectors, including commerce, social media, healthcare, and education [1, 2]. An optimal educational setting involves exchanging information and implementing extensive deep-learning experiments. The significance of education in the context of sustainable development (ESD) is showing a notable increase in recent years, as evidenced by the analysis of educational data obtained from diverse e-learning platforms and the advancement of conventional educational systems. To illustrate the full range of possibilities EDM offers, various datasets sourced from diverse origins are interconnected. The institution's data are carefully examined to identify factors that may contribute to advancing the educational process. This is in contrast to conventional database analysis, which can generate responses to queries like "Who is the student that received a failing grade?" By employing Educational Data Mining (EDM), researchers can obtain insights into more comprehensive inquiries, such as predicting a student's final grade or determining their exam outcome as a pass or fail [3].

Educational results delineate the potential or capacity of students as a direct outcome of modifications in their behavior stemming from their educational encounters. The prediction of students' performance results from transforming their behavior after participating in academic courses, encompassing information acquisition, cognitive abilities, and physical aptitude [4]. Particularly, building materials forms of students' achievement include the comprehension of subject matter, the processing of information, and the utilization of intellectual or physical abilities to make informed decisions. Within the sequence of programs to be completed, it is feasible to systematically monitor and assess students' cognitive, effective, and psychomotor proficiency, drawing upon a comprehensive comprehension of their academic achievements. A student's academic efficacy is contingent upon the caliber of the instruction and learning process. To enhance this process, it is possible to utilize learning outcomes to assess and improve [5].

1.1 Federated learning

As in federated learning, data are distributed across numerous devices, often leading to unbalanced datasets. This heterogeneity can result in challenges in model training and performance and make it difficult to develop models that perform consistently across all devices. Previously, to overcome such issues techniques like advanced data sampling and model personalization were used to ensure more uniform model training and performance across diverse datasets. Analyzing student performance and engagement data locally allows them to adapt learning content to individual needs, thus improving educational outcomes while safeguarding student privacy and promoting privacy-conscious educational environments. Federated machine learning achieves higher performance as it can assist in identifying learning gaps and provides targeted interventions needed to improve and better acquisition of educational programs. Federated machine learning algorithms can also facilitate the optimization of student data processing by identifying inefficiencies and predicting future performance trends, which results in more streamlined the student's results.

FL is an ML-based approach that can be employed in made-available edge computing devices or servers to preserve personal data samples without data exchange, thereby facilitating the training of an algorithm. FL is an algorithmic approach wherein multiple clients collaborate with a central aggregator to address machine learning problems, as presented in Fig. 1. The system receives training data from FL, ensuring the users' privacy is upheld. FL effectively handles and solves the previously mentioned concerns using a centralized aggregator server. This server provides access to a universally shared deep learning (DL) structure on a global scale.

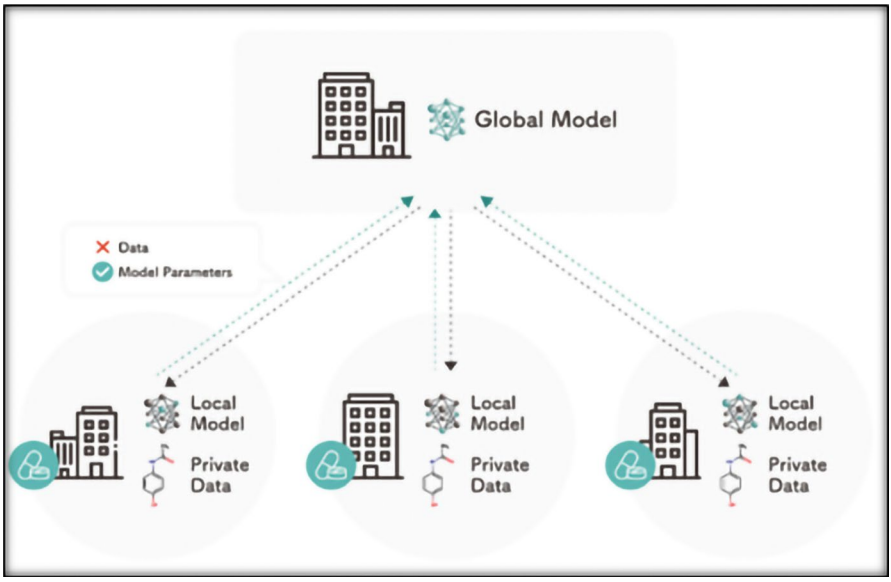

Fig. 1 Represents a computational strategy of FL where many clients work with a core aggregator to tackle issues related to machine learning

This article examines the different uses of FL in diverse fields, including medical services, intelligent health reasons evaluation networks, mobile devices, and assistance concepts.

Furthermore, we offer an extensive examination of the prevailing challenges faced by FL, encompassing aspects such as safeguarding user privacy, communication expenses, system heterogeneity, and the lack of reliability in model uploads. Currently, federated learning is confronted with several challenges. Subsequently, the various potential avenues for future research will be examined. The federated neural network algorithm trains diverse machine models by leveraging multiple local datasets instead of exchanging data with external datasets. As a result of this, enterprises have the capability to construct a cohesive global model without the necessity of storing training information in a centralized location. The achievement of accurate prediction performance is a direct outcome of federated learning. This collaborative approach enables scientists from different institutions and geographical locations to collectively educate multi-centric artificial intelligence (AI) algorithms using diverse datasets. Preservation of the information belonging to an organization's administration and human being users' privacy is ensured by solely transferring algorithms while keeping the data intact [6].

1.2 Cross-silo federal learning

The cross-silo FL system facilitates the integration of multiple silos, each connected to an essential server. Numerous companies can communicate through a unified approach while preserving their raw data in separate silos. The technology efficiently manages extensive datasets across multiple organizations while upholding data privacy principles. Figure 2 depicts a cross-silo system that connects a nearby hospital, a local testing facility, and a clinical research facility [7]. Each organization serves the community, providing them with distinct yet analogous environments to conduct their business.

1.3 Cross-device federated learning

Cross-device FL operates on devices that are interconnected within a network. Various instruments used by users, including adaptable phones and laptops, function as sources of information that train the model locally, as shown in Fig. 3. Subsequently, the server amalgamates these components to construct a comprehensive global model. The utilization of this technique by Google to train subsequent prediction models is a noteworthy illustration. To predict the following word, an initial step involves the creation of a localized linguistic framework, which is established by utilizing the user's input within the given device. The process consists of using federated sequential models to store and recall previous user input data, enabling the generation of future predictions. Consequently, the network experiences enhanced efficiency, improving user experience. In the realm of textual composition, it is a widely held sentiment that individuals generally appreciate the presence of accurate predictive suggestions for subsequent words during writing [8].

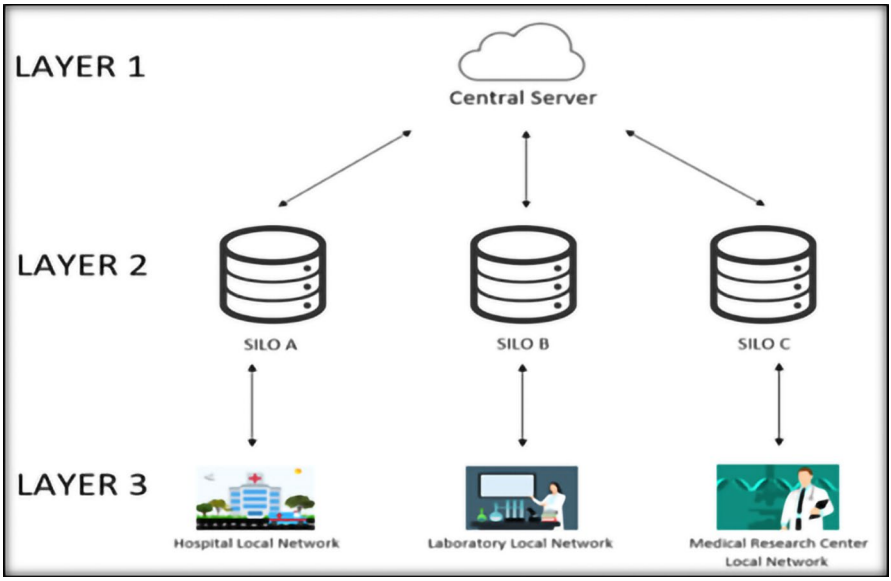

Fig. 2 Demonstrates a cross-silo FL system that connects a nearby hospital, a local diagnostic facility, and a clinical research complex

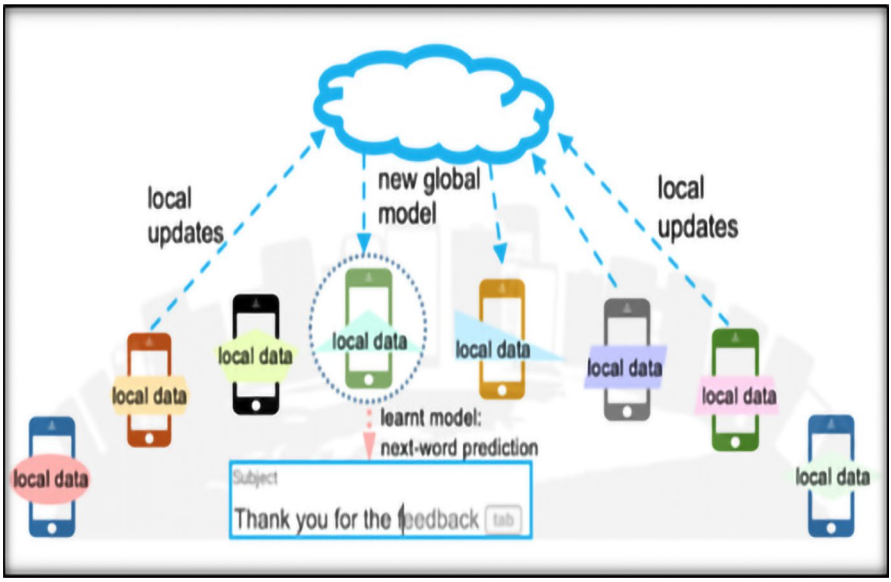

Fig. 3 The provides a visual representation of a collaborative system where users use adaptable phones and laptops as information sources for local model training. This decentralized approach ensures privacy and security while leveraging collective intelligence to enhance the system

2 Literature review

Kamar et al. [9] have extensively investigated academic records mining, leading to a significant body of research in this area. Numerous studies within this domain have focused on developing predictive models to evaluate the capabilities of university students and identify potential risks within educational institutions. Some approaches analyze the influence of different factors on student performance, while others aim to build optimal classification models for predicting uncertain future performance. In the following section, we delve into the interrelated research endeavors, shedding light on various aspects explored in this field.

The proposed study assesses student performance exclusively based on specific factors that yield suitable outcomes. In addition, the proposed methodology incorporates the characteristics of the J48 Classifier and Naive Bayes classification technique to ascertain the effects and classify students' overall academic achievement in a specific manner [10]. Multiple classification algorithms were applied to the student's data in a study conducted by [11]. These algorithms included fully connected layers, Naive Bayes (NB), sequential minimal optimization (SMO), J48, and REPTree. Altabrawee and Dr. Ami developed a novel prediction model in their research. The Ami model effectively assists students who lack motivation and exhibit weak academic performance to enhance their study habits and attain satisfactory grades. Altabrawee et al. propose a novel prediction model for forecasting student performance. The proposed study involved the development of a classifier utilizing four distinct ML-based techniques to predict students' overall academic performance in a specific subject. Artificial Neural Networks (ANN), NB, decision trees (DT), and Linear Regression (LR) constitute a set of machine learning techniques [12]. This study investigates the impact of utilizing the Internet as an educational platform and the influence of college and university students' social media usage on their academic performance.

Education data mining (EDM) approaches have been instrumental in facilitating the rapid growth of the education industry. These approaches have contributed to the development of new and improved services for students. Karthikeyan et al. [13] investigate various factors associated with Educational Data Mining (EDM), including students' performance, attrition rates, cognitive abilities, teacher and administrator performance, course delivery, and material composition. Researchers employ diverse data analytics tools to derive meaningful insights from educational data. The complexity of the developed predictive models renders them impracticable for making accurate future predictions, posing a significant challenge in deciphering their underlying mechanisms. Additionally, it offers enhanced predictive capabilities and a heightened emphasis on the comprehensive evaluation of learners' performance. Dhilipan et al. [14] conducted, addresses the intricate challenge that needs to be resolved to predict students' future success in online courses accurately. The ability to anticipate student performance during the initial phases of a course is of utmost importance to enable timely educational interventions. However, prior research has not comprehensively analyzed this issue from the deep learning perspective. This research

introduces a novel approach known as Grit Net, which leverages bidirectional long short-term memory (LSTM) to address the challenge of predicting pupil achievement. According to our data, the Grit Net consistently outperforms the conventional logistic regression technique. This advantage is particularly evident during the early weeks when accurate predictions are the most challenging.

Li et al. [15] utilization of federated learning (FL) in this study facilitated a collaboration in data science that resulted in the development of a predictive model for clinical outcomes in COVID-19 patients. This work has set a significant precedent for applying FL in the medical domain. FL is an approach employed in training AI-based models, wherein data from multiple sources is utilized while ensuring data privacy. This approach effectively addresses the challenges associated with data sharing. FL is an algorithmic approach wherein multiple clients collaborate with a central aggregator to address machine-learning tasks. The system receives training data from FL, ensuring the preservation of user confidentiality. FL addresses and resolves the previously mentioned concerns using a centralized aggregator server. This server facilitates access to a universally shared DL model on a global scale. This article examines fuzzy logic (FuL) applications in diverse fields, including healthcare, intelligent medical diagnostic systems, wireless communications, and service concepts [16].

Furthermore, we offer an extensive examination of the challenges faced by FL, encompassing aspects such as safeguarding user privacy, communication expenses, system heterogeneity, and the lack of reliability in model uploads [17]. Currently, federated learning is encountering a range of challenges, a few of which will be discussed herein. Subsequently numerous potential avenues for future research will be examined. The federated learning algorithm trains diverse machine models by utilizing multiple local datasets instead of sharing data with external datasets. As a result of this, enterprises have the capability to establish a cohesive global model without the necessity of storing training data in a centralized location. The achievement of strong predictive capabilities is a direct outcome of federated learning, a collaborative approach that enables scientists from diverse institutions across different geographical locations to train multi-centric artificial intelligence models using disparate datasets collectively. An organization's data governance and individual users' privacy are maintained by solely transferring algorithms and not the data itself. FL has a higher potential and is utilized to predict learners' academic performance. Hu et al. [18] provide an initial framework for understanding and contrasting FL with traditional distributed learning. The study focuses on elucidating the various aspects of FL, including its development process, definition, architecture, and categorization. Subsequently, the document delineates several prevalent FL concerns requiring attention. This paper provides a concise overview and comparison of various federated machine learning algorithms, specifically those rooted in conventional FL algorithms. The primary emphasis is placed on deep learning and classification techniques. This study also addresses future advancements of federated learning by utilizing deep learning techniques [19]. The proposed FL framework has demonstrated superior performance to existing machine learning models for predicting student performance in terms of efficiency, reliability, and precision. To facilitate the creation

of multiple machine learning-based localized data models, we conducted an assessment of three distinct local datasets.

Realinho et al. [20] utilized a combination of social community assessment and data mining techniques to predict the prospective success or failure of high school students at the outset of their education. The aim of their observation was modified to increase the accuracy of academic information inputted while considering the social community's opinion. Tyler et al. [21] proposed a theoretical framework that utilizes data collected during admission to predict students' academic performance and evaluate their initial term achievements. This novel methodology enables data from multiple colleges and universities to be categorized into two distinct categories. The initial cohort of children has been identified as exhibiting subpar academic achievement and being susceptible to adverse outcomes, while the subsequent cohort has been categorized as demonstrating commendable academic performance. Kaur et al. [22] developed an innovative model that centers on various student attributes to predict academic performance. The instructional performance model is influenced by multiple factors, making it crucial to establish a predictive data mining technique for analyzing students' performance. This technique is necessary to differentiate between advanced beginners and gradual beginners. Bhardwaj et al. [23] proposed a predictive student achievement model incorporating multiple classifiers. The suggested model, AODE, combines the aggregation of one-dependence estimators and K-nearest neighbors. A rule consisting of various possibilities is used to incorporate numerous classifiers for predicting academic performance among engineering students. The proposed methodology was employed and subjected to a t-test for comparison across three distinct datasets about student performance. The presented approach is also likened to KSTAR, OneR, Xero, NB, and NB tree algorithms and individual algorithms to establish a contrast.

Pandey et al. [24], a novel methodology for earthquakes, represent a naturally occurring phenomenon that can significantly harm human lives and infrastructure. Researchers have explored various AI techniques for earthquake forecasting. However, achieving high precision remains challenging due to vast multivariate data, delay in communication transmission latency, limited computational capacity, and data privacy concerns. FL is a machine learning-based approach that addresses these challenges by enabling on-site data collection and processing while ensuring data privacy without transmitting it to a central server. The federated approach aggregates local data models to construct a global data model, inherently ensuring data security and privacy and accounting for data heterogeneity. This study introduces an innovative conceptual framework for earthquake prediction, leveraging machine learning principles in seismology. The proposed FL framework outperforms traditional earthquake forecasting models by generating multiple ML-based local data methods from three distinct local datasets. These models are combined using the Fed Quake algorithm to form a global data model on a central FL server. A meta-classifier is trained using this global data model to enhance earthquake prediction accuracy. Glannini et al. [25] used AI and DL algorithms to analyze the complex problems of the students and also the instructors in an institute, such as "lack of communication," lack of discussion about subjects," and the harassment of the male and female instructors. Chai et al. [26] utilized AI and DL to enhance student's

learning abilities in both private and public sectors, not only for classroom teaching but also for management tasks in institutes. Chu et al. [27] propose a personalized federated learning framework to enhance learning outcomes for underrepresented students. The framework uses meta-learning, adapted local models from a global model, and personalized steps. The focus is on predicting learning outcomes from student activities, with subgroups formed based on community. The primary goal is to maximize predictive quality for all student subgroups, particularly underrepresented ones. The framework's well-organized student embedding correlates with improved student modeling. Syreen et al. [28] discuss the growing importance of federated learning (FL) in academia and industry, highlighting its potential to improve data-driven applications. Zhang et al. [29] explore federated learning as a new solution for safeguarding big data and artificial intelligence privacy, addressing issues of sharing mechanisms and task-specific framework migration. Wen et al. [30] the authors discuss FL's importance and rapid growth as a crucial solution for protecting user's privacy. A new text classification method involves preprocessing, feature extraction, feature selection, feature weighting, and classification. Preprocessing includes tokenization, stop word removal, lower-case conversion, and stemming techniques to enhance classification performance. The processed tokens are used to create a list of features for textual data. Parlak et al. [31] study on text classification uses a novel filter selection method called the Extensive Feature Selector. They use three local feature selection methods and three globalization techniques on four benchmark datasets: 20 Newsgroup, Enron1 and Polarity [32]. The datasets have four characteristics: multi-class unbalanced, multi-balanced, binary-class unbalanced, and binary-class balanced. The classification stage uses SVM and DT, two well-known classifiers.

The suggested framework is rigorously tested by evaluating multidimensional seismic data within a radial area of 100 km centered at coordinates 34.708° N and 72.5478° E in the Western Himalayas. The finding demonstrates the effectiveness and potential of the suggested federated learning-based approach for earthquake prediction in seismic studies. Upon conducting a comparative analysis between the outcomes of the suggested framework and the regional seismic data collected through instrumental recording over 35 years, it was determined that the prediction accuracy amounted to 88.87%. The findings derived from the proposed framework have the potential to serve as an important element in the creation of seismic early warning systems.

Only educational institutions always seeking new ways to improve quality can produce students with the greatest levels of accomplishment while still maintaining the lowest rates of failure. Educational federated learning has recently been more prevalent in educational settings. This is because federated learning enables teachers to analyze and forecast the performance of their students, as well as take the necessary actions in advance. Due to issues with prediction accuracy, erroneous attribute analysis, and insufficient datasets, educational systems are facing difficulties and roadblocks while trying to fully benefit from educational federated learning. To make the prediction process more accurate, it is essential to conduct an exhaustive analysis of the relevant research and then choose the most suitable prediction technique. The principal emphasis of this work is on federated learning techniques,

classification algorithms, and the effect that they have on datasets, in addition to the result of the prediction attribute. The purpose of this work is to give clear and brief comparisons between each of these aspects of federated learning. The research also identifies the most reliable features that may be used to accurately forecast students' academic performance.

3 Materials and methods

First, a federated educational framework is created in this research to forecast students' educational achievement into low, good, average, and drop utilizing various machine learning techniques and approaches, as shown in Fig. 4. Second, to

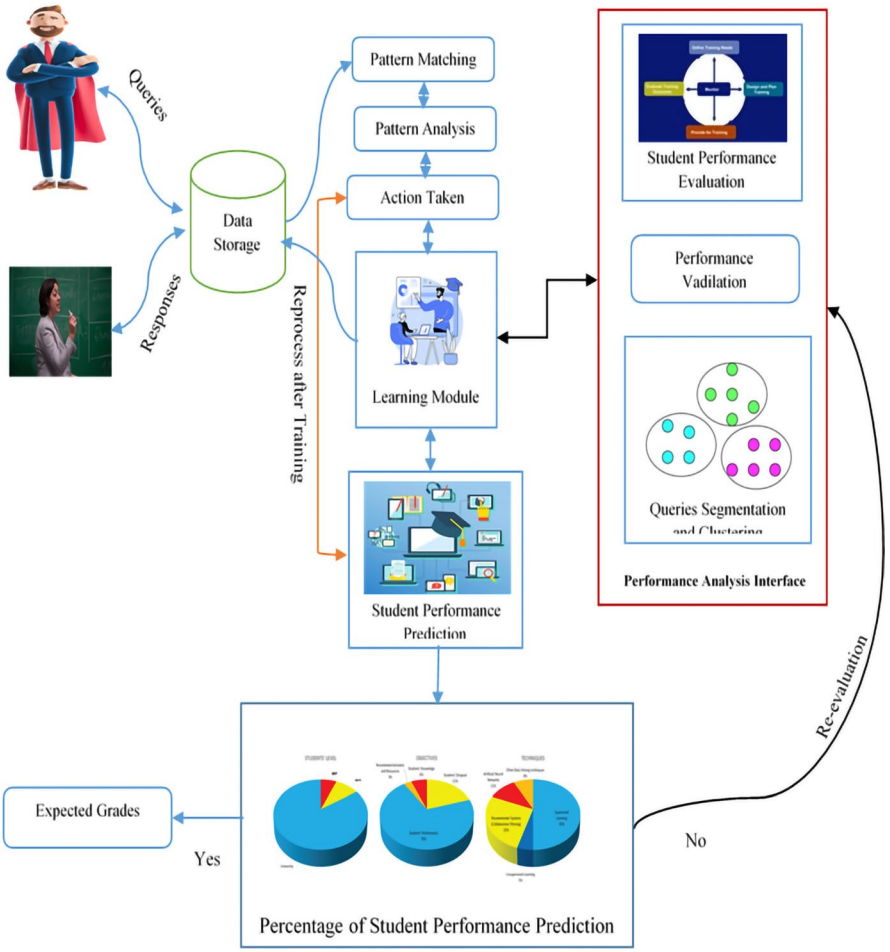

Fig. 4 Proposed model to predict students' educational achievement as low, good, average, or drop. The model incorporates diverse machine-learning techniques and approaches to achieve accurate forecasts

make the models more precise, we employ optimized features. Finally, five different machine learning models have been evaluated on our optimized dataset for training and testing. We employ a federated learning model, KNN, NB, SVM, and decision trees. At last, the F1-score of models, accuracy, precision, recall, and other metrics are computed based on the outcomes of each model [33, 34].

In Fig. 4, a massive amount of data are required to train the models for federated learning, so in the first layer, data are combined from several networks containing information on students. Layer 2 models are developed using the TensorFlow framework. As part of the first layer, an encryption system is applied, and after that, the obtained data are processed before being utilized for the training process of the proposed models. The proposed ML-based algorithms are trained using the processed data sent to the second layer. Each model's training data and resulting predictions are saved in a central cloud database. The third and final layer is an aggregation of instructors for the students, hosted on a cloud server and informed by a global model/metamodeling. After each iteration, the global model compiles the local models' insights and revises its prediction accordingly. The global prediction is provided after applying the meta-classifier to the results from the regional models. For pupils to become effective students is one of the most significant objectives of education. Everyone interested in determining effective teaching methods must consider the ability to predict outcomes for student learning through observation data and learners' achievement using federated learning. Information on students will be gathered, including demographics, academic performance, and more. The datasets will include feature information about the dataset's distribution alongside the samples used for training and assessment sample sizes. The student-to-sample ratio is also taken into account.

3.1 Dataset

Three separate Kaggle datasets were combined to form the dataset utilized in this study. Dataset 1 (StudentAsseccment.CSV, StudentInfo.CSV) includes information about individual students and their academic achievements. Dataset 2 (StudentPriction.CSV) includes a plethora of attributes, such as student family details, student education details, and current semester grades, while dataset 3 (StudentData.CSV) provides just eight features about students' performance in each topic as explained in Table 1. The dataset includes 33 variables, including the student's educational history, personal details, and family history. Columns 1 to 10 in the dataset contain the demographic questions, while columns 11 to 16 contain responses. Some questions focus on the student's family, while others probe his or her educational preferences. The dataset contains 22 columns that are displayed in Table 1. In Table 1, we have taken the following datasets of 145 students such as "Student ID", "Age", "Gender", "Highest Study Type", "Scholarship", "Work Activity", "Partner", "Salary", "Transport", "Living", "Mother Education", "Father Education", "Number of Siblings", "Kids", "Mother Job", "Father Job", "Study Hours", "Read Frequently", "Read Frequently Science", "Attending Department" and "The impact of the study on him/her".

Table 1 Student information dataset

Sr.no	Std detail	Discipline	Class category	Object
0	Std ID	145	Non-empty	int-64
1	Age	145	Non-empty	int-64
2	Gender	145	Non-empty	int-64
3	High Study	145	Non-empty	int-64
4	Scholarship	145	Non-empty	int-64
5	Work	145	Non-empty	int-64
6	Activity	145	Non-empty	int-64
7	Partner	145	Non-empty	int-64
8	Salary	145	Non-empty	int-64
9	Transport	145	Non-empty	int-64
10	Living	145	Non-empty	int-64
11	Mother-Edu	145	Non-empty	int-64
12	Father-Edu	145	Non-empty	int-64
13	No. of Sibling	145	Non-empty	int-64
14	Kids	145	Non-empty	int-64
15	Mother-Job	145	Non-empty	int-64
16	Father-Job	145	Non-empty	int-64
17	Study-Hours	145	Non-empty	int-64
18	Read-Freq	145	Non-empty	int-64
19	Read-Freq-SCI	145	Non-empty	int-64
20	Attend-Dept	145	Non-empty	int-64
21	Impact	145	Non-empty	int-64

3.2 Dataset preprocessing

The proposed study commenced by conducting preliminary data preprocessing to facilitate experimentation. This initial phase encompassed three primary procedures: data discretization, imputation of missing values, and addressing imbalanced data [35]. Initially, data integrity analysis was performed on the student dataset to identify variables with notably low data integrity. The dataset consisted of two types of variables: discrete and continuous. Discrete variables were encoded using binary values (0/1), while the continuous variables were subjected to discretization. Specific guidelines were employed during the discretization process. A binary encoding (0/1) was utilized for discrete features, while continuous features were discretized into suitable intervals to represent various feature levels. This thorough data preprocessing effectively eliminated noise from the dataset and improved classifiers' performance in subsequent analyses and experiments. Algorithm 1 shows the procedure for predicting the grades of 10 students in a class based on four parameters such as quiz, assignment, mid-term, and final-term marks. Initially, the student's details in maintained in an Excel sheet, and later on these data are into.CSV file. After that, the data are distributed into two sections. 30% data are used for testing purposes, and 70% data are used for training purposes. After training and testing the data, the students' details are mentioned

along the X-axis, and the parameters for defining the students' grades are mentioned along the Y-axis.

Algorithm 1 Student prediction data preprocessing with feature loop

Data: Student Prediction Dataset
Result: Standardized training and testing sets
Input : CSV file path for student prediction dataset
Output: Standardized training and testing sets
Extract the input features by dropping the 'GRADE' column.
Extract the target variable as the 'GRADE' column
Split the dataset into train and test datasets using train_test_split.
Set the test_size parameter to 0.3 (30% for testing, 70% for training).
Obtain X_train and X_test as the input features for training and testing, respectively.
Obtain y_train and y_test as the target variable for train and test, respectively.
Get the lengths of X_train and X_test
Initialize a StandardScaler object as sc.
for *each feature in X_train and X_test* **do**
    Apply sc.fit_transform to the current feature in X_train Apply sc.transform to the current feature in X_test
end
Note: The loop assumes individual feature scaling. If features are scaled together, apply sc.fit_transform on the entire X_train and sc.transform on the entire X_test without a loop.

3.3 Importance of dataset and expert-driven feature selection scheme

Adequate sample size is essential in creating a precise and reliable classification model. Superfluous features and entries can detrimentally affect classification accuracy, increasing computational demands [36]. Consequently, it is vital to identify the most critical features during the experimental design phase. To achieve this objective, several tests were performed, beginning with twenty features and expanding further until no notable progress was identified. The suggested classifier's superior classification performance was combined with the suggested MFV-based expert-driven feature identification technique during these investigations. The data from these experiments illustrate that classification accuracy decreases consistently when using more than an eighty-feature subset, signifying no considerable improvement in the predictive accuracy of the federated learning algorithm. Therefore, it was extrapolated that features from twenty are critically important in creating an effective classification model for the FL algorithm. Due to privacy and ethical restrictions limiting the availability of other public datasets related to the federated learning algorithm

for significance testing, two benchmark techniques were chosen for comparison with the suggested MFV-based expert-driven feature selection framework. The initial benchmark study employed an Information Gain (IG) technique for selecting features to develop the initial FL algorithm predictive model. The second benchmark study utilized a combined data imbalance technique to boost the federated learning algorithm's classification performance and advocated employing an SVM classifier.

4 Machine learning-based algorithm

This research paper extensively explores a range of machine learning (ML) models within data mining, employing widely accepted methodologies [37, 38]. The study selects and applies distinct ML models, including SVM, DT, NB, KNN, and an FL Model. The main objective revolves around utilizing these models for predictive analysis of student grades, effectively categorizing them into four discrete levels: low, good, average, and drop. Sarker et al. [39] proposed that machine learning involves machines autonomously acquiring data knowledge, improving performance through prior experiences, and generating predictions. ML algorithms efficiently process large datasets for their tasks after undergoing training. These techniques address various business challenges, encompassing organization, interactions, planning, classification, and regression. ML can be categorized into four types based on learning methodologies and techniques.

Alzubaidi et al. [40] highlight that machine learning constitutes a fundamental subfield within artificial intelligence (AI) and is the cornerstone of deep learning. Inspired by the neural networks in the human cerebral cortex, DL operates without explicit programming, instead relying on many nonlinear processing units to extract and transform features. Each layer of the network builds upon the outputs from the preceding layer, enabling the model to address the dimensionality challenge by autonomously identifying and prioritizing relevant features with minimal programming intervention. The versatility of deep learning makes it widely applicable, particularly in scenarios involving multiple inputs and outputs. As an integral branch of AI, deep learning aims to develop algorithms that emulate the human brain's cognitive processes, aligning with the overarching goal of artificial intelligence in replicating human behavior.

4.1 Support vector machine (SVM)

Our study employs the support vector machine (SVM) models presented by Cortes et al. [41], widely utilized for distribution estimation, regression, and classification tasks. SVMs use nonlinear techniques to transform independent variables into higher-dimensional feature spaces, ensuring better separation [42, 43]. They find extensive applications in remote sensing domains. When dealing with multi-class

classification scenarios where data are not linearly separable, SVMs rely on kernel functions like linear, polynomial, radial basis, and sigmoid, each requiring specific parameters. The selection and parameterization of these kernels significantly impact the accuracy of SVMs [41, 44]. To combat overfitting and regulate the trade-off between margins and training errors, tuning parameters, particularly the penalty factor C, are utilized [45, 46]. For our research, we chose the Radial Basis Function (RBF) kernel for classification and regression tasks. Two crucial parameters need specification: C and gamma. The C parameter balances maximizing the margin and minimizing classification/regression errors. Setting C to 22 emphasizes maximizing the margin and potentially reducing training error. However, it is essential to acknowledge that the optimal C value may differ depending on the dataset and specific problem. Algorithm 2 shows the number of variables denoted by p* for determining the grade level of the students in the form of "low", "drop", average," and "good". In line 14, we have applied an SVM algorithm to predict the student's grade level by applying different weights on each variable. Line 15 shows that if we have more than 2 sub-variables of a main variable, like quizzes 1, 2, and 3. We have selected the optimal sub-variables for analyzing the grade level. Lines 20-21 show if the variable values are less than the defined values, the grade level will be named "drop". At the last, the algorithm will show a grade level based on the subject value.

Algorithm 2 SVM model variable relevance ranking

Data: Set containing $p*$ variables and a binary result
Result: Relevance-based grading of a list of variables (low, drop, average, good)

13 Determine the optimal parameters or values for the SVM model;
14 Train the SVM standard $P^* \rightarrow P$
15 **if** $(p \geq 2)$ **then**
16 SVM $\rightarrow$ SVM$_p$ with optimal tuning settings for the p variables and data findings.
17 $W_p \rightarrow$ calculate the weight vector of SVM$_p$ $(w_{p1}, \ldots, w_{pp})$.
18 Low.grade.level $\rightarrow$ Variable having the smallest value in Grade.level vector.
19 Grade.level $\rightarrow (w_{p1}^2, \ldots, w_{pp}^2)$.
20 Low.grade.level $\rightarrow$ Variable having the smallest value in Grade.level vector.
21 Grade$_p \rightarrow$ min(Grade.level).
22 Remove drop.grade.level.
23 $p \rightarrow p - 1$
24 **end**
25 Grade_good $\rightarrow$ variable in data $\in$ (grade level: good, average, low, drop).
26 Return (grade level: good, average, low, drop)

4.2 Decision trees (DT)

Our study delves into nonparametric supervised learning, where decision trees (DT) serve as a widely adopted approach [47, 48]. These trees offer versatility for both classification and regression tasks, with prominent algorithms such as ID3 [49], C4.5 [47], and CART [50]. To achieve enhanced performance and superior results, tuning the parameters of DT algorithms becomes crucial. Our research stipulates a maximum depth of 15 for the decision tree, effectively capturing more complex data patterns without overfitting. Additionally, by setting the random state to 29, we ensure consistency in the random aspects of the algorithm across multiple runs. Algorithm 3 shows the working of the DTC model for classifying the students' grade level based on the different variables such as quizzes, assignments, mid-term and final term marks. In line 28, we have selected specific parameters for predicting the grade level, such as low, drop, average, and good. Line 29 shows that the maximum number of parameters might be taken as 6, the minimum number of parameters might be taken as 2, and the parameter level position must always be started with the '0' position. Lastly, the DTC algorithm is used to train different variables and test the grade level.

Algorithm 3 Decision tree classifier (DTC) algorithm

Data: Training features X_{train}, training labels y_{train}, test features X_{test}
Result: Predicted labels for test features y_{pred}
Input : None
Output: Predicted labels for test features y_{pred}
Import the DecisionTreeClassifier from scikit-learn.
Create a DecisionTreeClassifier object with specified parameters:
$DTC \leftarrow \text{DecisionTreeClassifier}(\text{max\_depth} = 6, \text{min\_samples\_split} = 2, \text{random\_state} = 0)$
Train the Decision Tree Classifier on the training data:
$DTC.\text{fit}(X_{\text{train}}, y_{\text{train}})$
Utilize the learned classifier to generate predictions on the test data.
$y_{\text{pred}} \leftarrow DTC.\text{predict}(X_{\text{test}})$

4.3 Naïve Bayes

In our research investigation, we explore the application of the Naïve Bayes algorithm for predicting student grades and effectively categorizing them into four levels: low, good, average, and drop [51, 52]. Naïve Bayes is a well-established probabilistic classification method known for its simplicity, efficiency, and effectiveness,

especially in managing high-dimensional datasets. The dataset used in this study comprises various features, including historical academic performance, attendance records, coursework scores, and other relevant factors that may influence student grades. The target variable represents the categorical grade level, with the four discrete categories mentioned earlier.

During the training phase, the Naive Bayes classifier acquires knowledge about the probabilities linked to each feature about each class label by using the labeled training data. The provided data are next used to compute the posterior probability of each class label based on the observable characteristics of a new and unobserved student record. The anticipated grade level for a particular student is determined by assigning the class label with the most excellent posterior probability. To evaluate the effectiveness of the NB model, we use commonly used assessment measures like accuracy, precision, recall, and F1-score. Furthermore, cross-validation methods are employed to guarantee the robustness of the model and mitigate the risk of overfitting. Our experimental results demonstrate the effectiveness of the NB approach in accurately predicting student grades and proficiently classifying them into four distinct levels. This research outcome provides valuable insights into the application of Naïve Bayes for educational data analysis. It establishes the groundwork for potential future research and practical implementation in student performance prediction and targeted intervention strategies.

4.4 K-nearest neighbors (KNN)

In our study, we explore the KNN algorithm, known as a non-generalizing learning method, often referred to as "lazy learning" or "instance-based learning" [53]. Unlike traditional approaches that build internal models, KNN retains all occurrences of the training set in an n-dimensional space. It then categorizes new data points by analyzing existing data and employing similarity metrics, such as the Euclidean distance function [54]. During the classification process, the KNN of each data point contributes their votes, and a simple majority determines the final classification. The accuracy of KNN heavily depends on data quality and exhibits some resistance to noisy training data. However, selecting the optimal number of neighbors to consider, in this case, K=1, poses a significant challenge in KNN. Nevertheless, KNN finds practical applications in both classification and regression tasks. The KK-nearest neighbors (KNN) classifier algorithm is illustrated in Algorithm 4. The algorithm is used for analyzing the grade level depending on data quality, exhibits some resistance to noisy training data, and analyzes the accuracy of the algorithm through cross-validation. Line 35 shows the optimal number of neighbors to be considered. Line 36 shows that six neighbors have been selected for training and testing purposes. Line 39 shows the procedure for analyzing the accuracy of the KNN algorithm after validating the data.

Algorithm 4 K-nearest neighbors (KNN) algorithm

Data: Training features X_{train}, training labels y_{train}, test features X_{test}, test labels y_{test}
Result: Predicted labels for test features y_{pred} and evaluation metrics
Input : None
Output: Predicted labels for test features y_{pred}, train and test accuracy, and KNN cross-validation score

Import the KNeighborsClassifier from sci-kit-learn.
Create a KNeighborsClassifier object with the number of neighbors set to 6:
$KNN \leftarrow \text{KNeighborsClassifier}(\text{n\_neighbors} = 6)$
Train the KNN Classifier on the training data:
$KNN.\text{fit}(X_{\text{train}}, y_{\text{train}})$
Make predictions on the test data using the trained classifier:
$y_{\text{pred}} \leftarrow KNN.\text{predict}(X_{\text{test}})$
Calculate and print the KNN train accuracy:
$\text{train\_accuracy} \leftarrow KNN.\text{score}(X_{\text{train}}, y_{\text{train}})$
print("KNN train accuracy: %0.3f" % train_accuracy)
Calculate and print the KNN test accuracy: $\text{test\_accuracy} \leftarrow KNN.\text{score}(X_{\text{test}}, y_{\text{test}})$
print("KNN test accuracy: %0.3f" % test_accuracy)
Import $\text{cross}_v al_s core from scikit - learn$
Perform 10-fold cross-validation on KNN with the entire dataset:
$\text{DTCCV} \leftarrow \text{cross\_val\_score}(KNN, \text{inputs}, \text{target}, \text{cv} = 10)$
print("KNNCV %0.2f" % (DTCCV.mean()))

4.5 State-of-the-art federated learning model for enhancing student learning outcomes

The federated learning classifier is a proposed model for securing the privacy of the dataset to predict student learning performance by using the educational data of different institutes. Firstly, this study develops a federated learning model to predict the grades of the students into low, good, average, and drop. Secondly, we use optimized features to improve the accuracy of the models. Thirdly, our optimized dataset is evaluated on five machine-learning models to train and test the models. We use a support vector machine, decision tree, Naïve Bayes, K-nearest neighbors, and a federated learning model. Finally, the results of all the models are calculated in terms of accuracy, precision, Recall, and F1-score of Models, and a complete comparative analysis is done. Federated learning and KNN perform best as compared to other models. The proposed research aims to ensure students' success as learners by employing federated learning to predict learning outcomes based on data and achievements. Real data from a multifunctional university is used for testing, making the model applicable to other institutions. The collected data includes learners' information, such as classes, grades, and relevant details, with insights

into data distribution and trainee-to-sample ratio. Data will be gathered from various university campuses and stored in a database, including different data types like text. Pattern recognition techniques will then swiftly and accurately identify familiar patterns in the data. Relevant data will be processed, while irrelevant data will be iteratively refined until relevancy is achieved. The output is subjected to secure and efficient FL, which trains algorithms across decentralized peripheral devices or servers containing local data samples, thereby significantly enhancing performance [55]. An integrated interface uses three key factors to assess the model's performance: Performance Evaluation, Performance Validation, and Segregation and Clustering. Performance prediction factors are also integrated into federated learning, enabling precise outcome predictions. Accurate predictions lead to assigning final grades, while inaccuracies trigger further iterations in the prediction cycle [56].

The research addresses challenges related to imbalanced data through concrete techniques, providing effective solutions. Moreover, concrete deep-learning classifiers are proposed and applied based on performance evaluation matrices, further enhancing the learning process. Pattern recognition, an essential data analysis method, utilizes ML algorithms to automatically identify patterns and regularities in various data types, ensuring swift and accurate recognition of familiar patterns. In this research, we utilize a comprehensive database as a repository of institutional data, encompassing teacher and student information. The database contains a vast array of information, including assessments, assignments, attendance records, exams, results, and alterations in behavior. To optimize the analysis of this data, we utilize pattern recognition, a data analysis technique that utilizes machine learning algorithms to autonomously detect patterns and consistencies within a wide range of datasets, encompassing text, images, audio, and other specified attributes. Pattern recognition systems demonstrate the capability to swiftly and accurately recognize familiar patterns. We extract meaningful insights from the database through pattern recognition analysis, aligning with our research objectives. Specifically, we analyze assignments, quizzes, and result patterns to identify regularities. The FL-SVM Model for Predicting Student Performance is presented in Algorithm 5.

In this study, the data analyzer plays a crucial role in processing and validating data obtained through pattern recognition. Data analysis is an essential practice that involves cleaning, transforming, and modeling data to extract meaningful insights for decision-making. Various interdisciplinary domains such as artificial intelligence, Pattern Recognition, and Neural Networks heavily rely on data analysis. The collected data from the data analyzer serves multiple purposes, including the federated learning process and decision-making for performance factors. A central model is created on the server and distributed to each participant, where batches of data are used for training and updating the models as new data enters the system. Performance evaluation of learners is conducted using fresh data points collected at admission to ensure the validity of each treatment. The final models are compared using testing datasets to assess the effectiveness of each processing step. The averaged model is then returned to the institutions and evaluated against local validation sets. The exchange of projections and actual values with the central server enables the calculation of performance metrics, which facilitates the classification of student performance using the federated learning model. The collected data, combined

with federated learning, allows the calculation of probability percentages to improve future results. This process holds significant promise in enhancing educational outcomes through data-driven decision-making and performance assessment.

Algorithm 5 FL-SVM model for predicting student performance

Data: u: number of students in a section
for $u = 1$ **to** U **do**
  $y_u \in (0, 1)$: Prediction whether student u passes the course or not.
  $a_u(t)$, where $t \in (1, \ldots, L_u)$: Time-dependent activity for student u.
  $h_u(t) = FL(a_u(t), h_u(t-1))$: Hidden state generates learned state for student u.
  $h_u^{=\sum_t \alpha(t) h_u(t)}$: Outcome of FL system for student u.
  $\alpha(t) = \frac{e^{e(t)}}{\sum_t e^{e(t)}}$: Weights applied on input parameters.
  $e(t) = p(t)^T \tanh(W_\alpha h_u(t))$, where W_α: Learning parameter.
  $y_u = \text{softmax}(W_l h_u^{+b_l})$: Predicted probability of pass/fail for student u.
end
$W_l \in \mathbb{R}^{k \times 2}$: Weight matrix for linear transformation of h_u, where $k = 48$ is the hidden dimension $b_l \in \mathbb{R}^2$: Bias vector
$L_{BCE} = \sum_u \mathcal{E}_\Omega[y_u \log(y_u) + (1 - y_u) \log(1 - y_u)]$: Used for evaluating the prediction quality

4.6 Evaluation parameter

In the study, Eq. 1 is used to assess the accuracy of each ML model. Accuracy, defined as the proximity of a measurement to the true value being sought, is a crucial metric in evaluating the models [57].

$$\text{Accuracy} = \frac{\text{TrPos} + \text{TrNeg}}{\text{TrPos} + \text{FaPos} + \text{TrNeg} + \text{FaPos}} \tag{1}$$

In this context, Eq. 2 is used to calculate Precision, a measure of the agreement between multiple measurements of the same quantity. Precision is determined by the number of true positive (TrPos) and false positive (FaPos) findings, highlighting the ability to identify positive instances accurately. Higher precision indicates reduced outcome variation, making them highly reproducible and reliable.

$$\text{Precision} = \frac{\text{TrPos}}{\text{TrPos} + \text{FaPos}} \tag{2}$$

In this context, Eq. 3 defines recall as a measure of a model's capability to recognize all relevant instances within a data stream. The recall is calculated by summing the correct identifications (true positives) and subtracting the false alarms (false

negatives). A higher recall value indicates that the model can effectively capture a larger proportion of relevant instances, making it more adept at recognizing true positives.

$$\text{Recall} = \frac{\text{TrPos}}{\text{TrPos} + \text{FaNeg}} \tag{3}$$

The F1-score, represented by Eq. 4, is computed using the harmonic mean of recall and precision. It is a symmetrical metric that considers both the model's ability to identify pertinent instances (recall) and the accuracy of its identifications. The harmonic mean combines these metrics to evaluate the model's efficacy, particularly when precision and recall are equally important.

$$F1\text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{precision} + \text{Recall}} \tag{4}$$

5 Results and discussion

The proposed system evaluates student academic achievement using different ML-based algorithms applied to a dataset. The assessment criteria encompass model accuracy, precision, recall, and F1-score [58, 59]. The results, depicted in Fig. 5, showcase the aggregated testing outcomes of the models, with 10 clients executing 90 epochs. The final global model has achieved an exceptional accuracy rate of 99% along with a loss rate of 1.5% [60]. Furthermore, Table 2 illustrates the effectiveness of the federated learning approach, achieving a 99% training accuracy after one hundred rounds of communication and a 99% testing accuracy. The federated learning confusion matrix demonstrates accurate results, with all classes displaying perfect precision, recall, and F1-score. The application of FL for analyzing the accuracy of the proposed system by training the four parameters, quiz, assignment, mid-term, and final-term scores, to predict the grade level based on these parameters. Line 61 shows the procedure of comparison of the accuracy of different classifiers such as Gaussian Naïve Bayes, DTC, and RFC, and after comparison, develop a fusion matrix for predicting its accuracy and plot its graph to analyze the accuracy ratio. For this purpose, initially, calculate the accuracy of DTC after training the pre-defined parameters, then cross-validate the accuracy by developing a confusion matrix and, similarly, predict the accuracy of the other classifiers.

5.1 Accuracy comparison between proposed ML-based models

Our study utilized the KNN, DT, SVM, and NB algorithms for training and testing. We split the dataset with 75% for training using KNN with 6 neighbors, and the remaining 25% was used for 10-fold cross-validation testing. The performance results for the presented models are shown in Fig. 5a–d and Table 2.

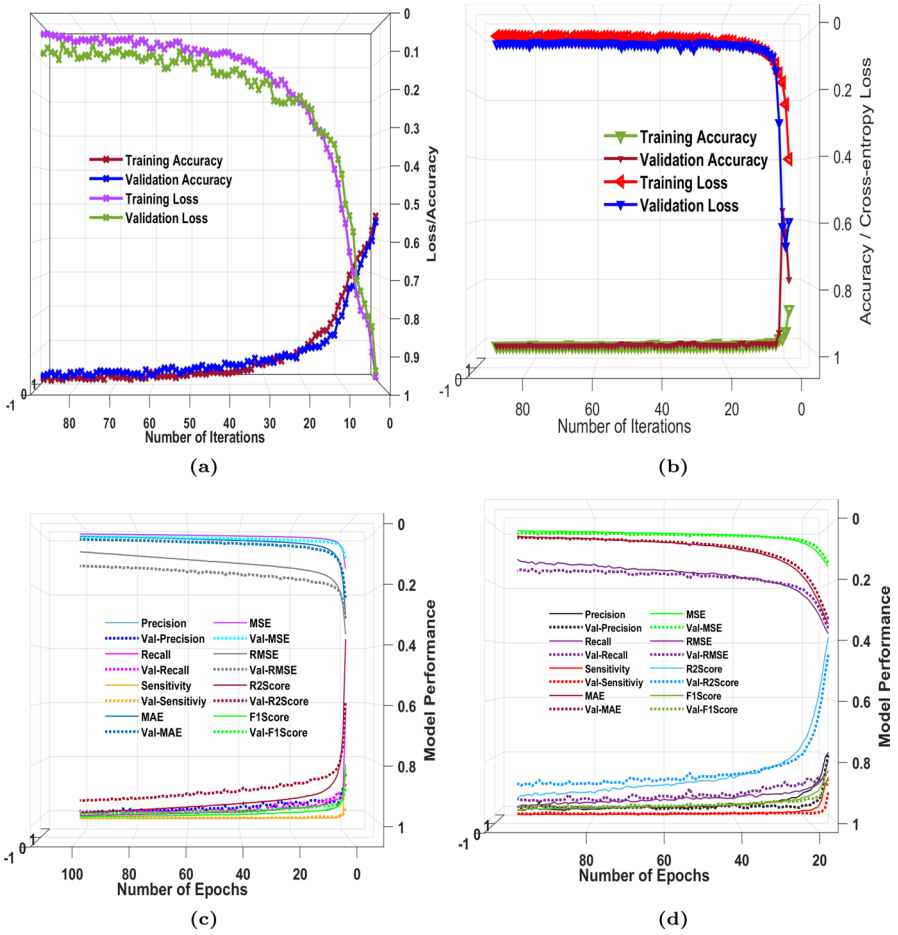

Fig. 5 The performance of proposed models **a** confusion matrix of FL, **b** SVM, **c** KNN, **d** DT, **e** NB, **f** AUC comparison of proposed algorithms

Table 2 Accuracy precision comparison of the proposed models

Models	Testing accuracy	Training accuracy	Cross validation
KNN	0.924	0.891	0.952
DT	0.941	0.952	0.950
SVM	0.981	0.987	0.989
NB	0.929	0.938	0.928
Proposed FL	0.990	0.995	0.992

The KNN model exhibited a training accuracy of 0.891, surpassed by the testing accuracy of 0.924 and the cross-validation accuracy of 0.952. This indicates that the model learned well from the training data and performed even better on unseen instances during testing and cross-validation. The model's predictions for each class were carefully evaluated, revealing true examples, false examples, and detailed explanations for each class, providing valuable insights into its classification capabilities. The DT model demonstrated excellent performance, achieving a higher training accuracy of 0.957 compared to testing (0.941) and cross-validation (0.950) accuracy. This indicates that the model effectively learned from the training data without overfitting and generalizes well to new instances. The model's interpretability and robustness further contribute to its suitability for decision-making tasks across various domains. The SVM model exhibited a training accuracy of 0.987, higher than both testing 0.981 and cross-validation 0.989 accuracies. The Fig. 5b depicted the training and validation accuracy and loss against the 90 epoch. The SVM model performed well on the testing database. The NB model demonstrated a training accuracy of 0.938, outperforming the testing accuracy of 0.929 and the cross-validation accuracy of 0.928. This indicates that the model learned well from the training data and displayed strong generalization abilities on unseen data during testing and cross-validation. The model's high training accuracy suggests its capability to capture essential patterns and features from the training dataset, leading to effective and accurate predictions. Moreover, as illustrated in Fig. 5a, the federated learning approach demonstrated remarkable performance, achieving a training accuracy of 0.995 after ninety rounds of communication and a testing accuracy of 0.990. Notably, the federated learning model outperformed other models in terms of accuracy. While both the SVM and FL models showed strong performance on the training and testing datasets, the federated learning model stood out as the clear winner, attaining impressive accuracies of 0.99 for both training and testing. The study compared the performance of three models—SVM, NB, and FL—on both training and testing datasets. These results unequivocally highlight the effectiveness of the proposed federated learning model in producing highly accurate predictions, surpassing the performance of other models in terms of training and testing accuracy. The federated learning approach exhibits robustness and efficiency in learning from distributed data sources, making it a powerful and promising technique for various real-world applications. The federated learning (FL) model emerged as the clear winner with impressive accuracies of 0.99 for training and testing. The SVM and FL models performed strongly on both datasets, but FL outperformed them. In contrast, the NB classifier achieved the lowest accuracy, with 7% for training and 6% for testing. These results reveal the effectiveness of the FL model in producing accurate predictions, highlighting its potential for handling distributed data and real-world applications.

5.2 Precision comparison of the proposed models

The evaluation of model performance revealed varying precision rates for different classes. Table 3 and Fig. 5c, d illustrate the precision trends against the 90

Table 3 Precision comparison of the proposed models

Models	Low (%)	Good (%)	Drop (%)	Average (%)
KNN	0.97	0.96	0.95	0.96
DT	0.97	0.97	0.96	0.97
SVM	0.98	0.98	0.98	0.99
NB	0.94	0.95	0.95	0.94
FL	0.99	0.99	0.99	0.99

epochs for all proposed models. The KNN model demonstrated high precision rates for different classes, achieving 0.97 for low, 0.96 for good, 0.95 for drop, and 0.96 for average. These precision rates highlight the model's accuracy in correctly predicting instances belonging to each class. The KNN model effectively classified data points into specific categories, making it a valuable tool for various classification tasks. The DT model demonstrated impressive precision rates for different classes, achieving 0.97 for low, 0.97 for good, 0.96 for drop, and an impressive 0.97 for average. These precision rates highlight the model's accuracy in correctly predicting instances belonging to each class. It is noteworthy that class low had the least accurate predictions, while class average exhibited the best performance. The DT model's ability to achieve consistently high precision across various classes makes it a reliable and effective tool for classification tasks. The SVM model displayed remarkable accuracy rates for different classes, achieving 0.98 for the low class, 0.98 for good, 0.98 for drop, and an impressive 0.99 for average. These accuracy rates indicate the model's proficiency in correctly classifying instances belonging to each category. As observed previously, class low had the least accurate outcomes, while class average showcased the highest level of accuracy, as depicted in Fig. 5d. The SVM model's consistently high accuracy across various classes makes it a robust and reliable classifier for diverse classification tasks.

In contrast to the other models, the NB model demonstrated slightly lower accuracy rates for different classes. Specifically, it achieved accuracy rates of 0.94 for the low class, 0.95 for good, 0.95 for drop, and a perfect 0.94 for average. Class drop had the least accurate results, indicating that the model faced challenges in correctly classifying instances belonging to this category. On the other hand, the class average exhibited the highest level of accuracy, showcasing the NB model's ability to predict instances in this class effectively. Despite the lower accuracy rates for certain classes, the NB model's overall performance was commendable, especially in the average class. The precision analysis of the models reveals that Model FL achieved high precision values of 0.99 for both the low and good classes, indicating its strong ability to predict instances in these categories accurately, as depicted in Fig. 5c. In contrast, Model NB obtained slightly lower precision values of 0.94 for the low class and 0.99 for the good class. These precision metrics provide valuable insights into the models' performance differences and their strengths in predicting specific class

outcomes. Model FL demonstrates superior precision in identifying instances for the low and good classes, while Model NB has relatively lower precision for the low class. These findings aid in understanding the models' capabilities and can inform future decisions for improving their predictive performance.

5.3 Recall comparison of the proposed models

Table 4 presents a comparison of recall rates for the proposed models. Model FL achieves a remarkable recall rate of 0.99 for the class drop, outperforming model NB with a recall rate of 0.94 for the same class. Model FL also demonstrates a perfect recall rate of 0.99 for the class average, while model NB achieves a lower recall rate of 0.94 for the same class. In the KNN model, the highest recall rate is 0.97 for the good class, followed by 0.96 for low and drop classes, and 0.95 for the average class. Similarly, the DT model's recall rate is highest for the drop class at 0.97 and lowest for the average class at 0.95, with 0.96 for low and good classes. The SVM model achieves a recall rate of 0.98 for the low class, followed by 0.97 for the average class, and 0.98 for both the good and drop classes. On the other hand, the NB model exhibits a recall rate of 0.93 for the drop class, while both the good and average classes have a recall rate of 0.94. The low class has the lowest recall rate for the NB model at 0.03. Model FL demonstrates exceptional recall performance for the low class with a perfect recall rate of 0.99, while the KNN model falls slightly behind with a recall rate of 0.97 for the same class. Fig. 5c, d visually depicts the recall rates for the good class, with model FL achieving a recall rate of 0.99, outperforming the KNN model, which achieves a recall rate of 0.97.

5.4 F1-score comparison of the proposed models

Table 5 shows that NB with a lower F1-score of 0.94 for the same class. Furthermore, model FL achieves a perfect F1-score of 0.99 for the class average, while the NB model obtains a lower F1-score of 0.93 for the same class. For the DT model, the F1-score ranges from 0.97 for the good class to 0.97 for the average class, with 0.96 and 0.97 for the drop and low classes, respectively. Similarly, the SVM model demonstrates F1-scores ranging from 0.99 for the good class to 0.97 for the average class, with 0.99 and 0.98 for the drop and low classes, respectively. The SVM model achieves the highest F1-score of 0.99 for the good class and the lowest F1-score of 0.97 for the average class. F1-scores of 0.97 and 0.98 are obtained for the average and low classes, respectively. Model FL excels in F1-score for the low class with a value of 0.99, while model NB lags with an F1-score of 0.93 for the same class. In this table, we have compared the proposed classifier federated learning with the rest of the art classifiers, including Naïve Bayes, SVM, decision tree, and KNN. In these models, we have calculated four levels of F1-score, including drop, low, average, and good. After comparison, it is

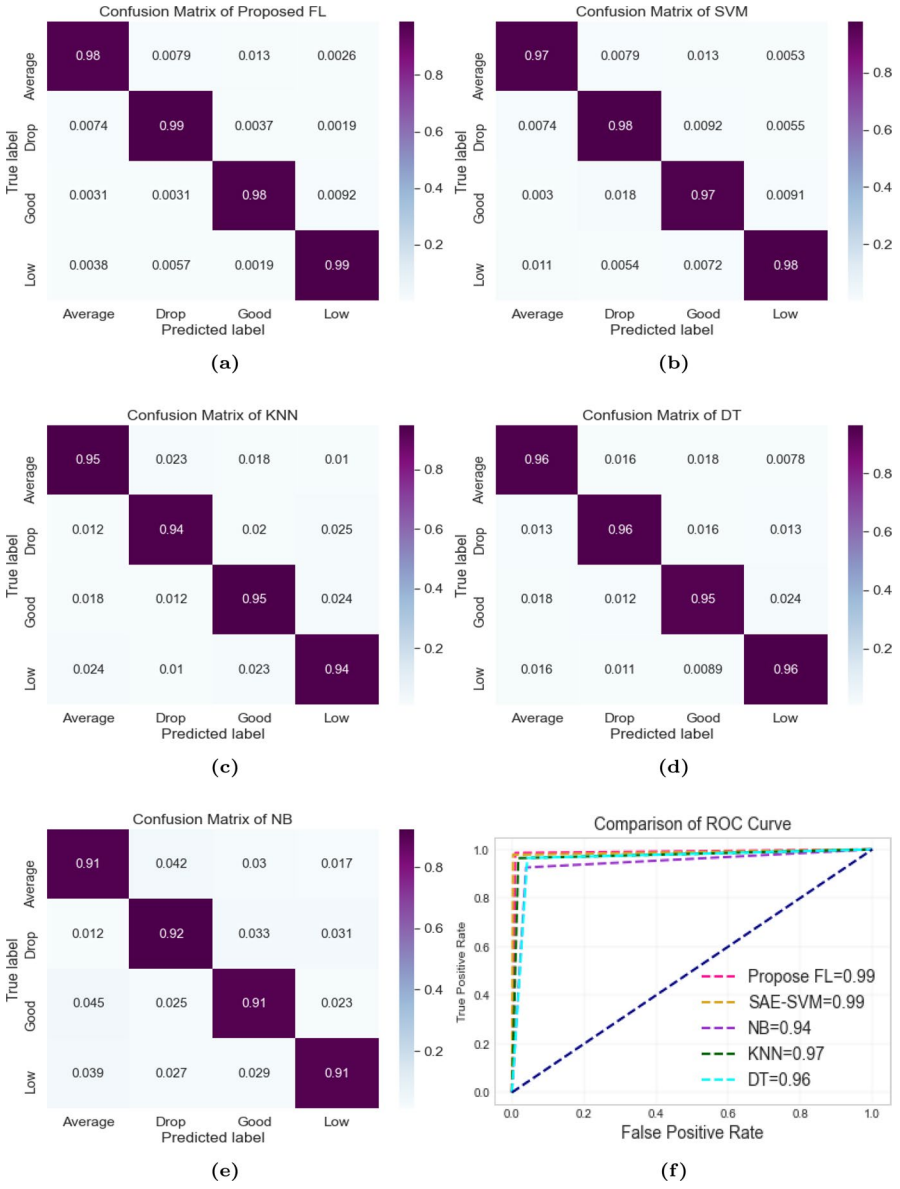

Fig. 6 The performance of proposed models **a** confusion matrix of FL, **b** SVM, **c** KNN, **d** DT, **e** NB, **f** AUC comparison of proposed algorithms

found that the F1-score of the proposed FL classifier is better than the rest of the art classifiers.

Figure 6a–f visually illustrates that model FL attains a perfect F1-score of 0.99 for the good class, while model NB achieves a lower F1-score of 0.94 for

Table 4 Recall comparison of the proposed models

Models	Low (%)	Good (%)	Drop (%)	Average (%)
SVM	0.98	0.98	0.97	0.97
DT	0.95	0.96	0.97	0.95
KNN	0.97	0.97	0.96	0.95
NB	0.93	0.94	0.94	0.94
FL	0.99	0.99	0.99	0.99

Table 5 F1-score comparison of the proposed models

Models	Low (%)	Good (%)	Drop (%)	Average (%)
SVM	0.98	0.99	0.98	0.97
DT	0.97	0.97	0.96	0.97
KNN	0.95	0.96	0.95	0.97
NB	0.93	0.94	0.94	0.93
FL	0.99	0.99	0.99	0.99

the same class. In this table, we have compared the proposed classifier federated learning with the rest of the art classifiers, including Naïve Bayes, SVM, decision tree, and KNN. In these models, we have calculated four levels of F1-score, including drop, low, average, and good. After comparison, it is found that the F1-score of the proposed FL classifier is better than the rest of the art classifiers.

5.5 Confusion matrix comparison of the proposed models

Confusion matrices play a confusion matrices play an essential part in evaluating the efficacy of each model. They offer valuable insights into the accuracy, precision, recall, and other performance metrics for different classes, comprehensively evaluating each model's capabilities in predicting class outcomes accurately. The presented confusion matrix represents a multi-class classification problem with four distinct classes, as visualized in Fig. 6c–e. From this matrix, we derive the performance metrics for the proposed FL algorithm, which demonstrate outstanding values across all classes. The precision, recall, and F1-score for classes 0, 1, 2, and 3 are approximately 0.992, 0.991, 0.996, and 0.991, respectively. These high scores indicate the model's accuracy in making predictions with minimal false positives or negatives. Furthermore, the model's overall accuracy, which signifies its performance across all classes, is approximately 0.994 for each class. This suggests that the FL model performs well in correctly identifying instances belonging to all classes within the dataset. The exhaustive evaluation provided by the confusion matrices enables us to comprehend the strengths and weaknesses of each model in terms of accurately classifying the data.

Fig. 6f illustrates the AUC scores, a robust measure of the classification performance of five proposed models in our study. The FL model stands out with exceptional performance, achieving an impressive AUC score of approximately 0.99. This

signifies a high TPR and a low FPR, emphasizing the FL model's proficiency in accurately classifying the data. Likewise, the SVM model exhibits superior performance among the four algorithms, attaining a notable AUC score of approximately 0.99. This highlights its ability to accurately classify the data with a high TPR and a low FPR. The RF model closely follows with an AUC score of 0.96, indicating strong classification performance, although it incurs slightly more prediction errors than the SVM model. The KNN model achieves an AUC score of 0.97, denoting fair classification performance, but it demonstrates relatively more classification errors than the SVM and RF models. Conversely, the NB model exhibits the weakest performance, obtaining an AUC score of 0.94, implying lower classification accuracy than the other models. The ROC curves further reinforce these findings, where an ideal classification performance would result in a curve extending vertically along the y-axis and then horizontally along the top of the plot. The curves of the FL and SVM models closely align with this ideal, reaffirming their strong performance, while the KNN and NB models show more deviation, indicating relatively weaker performance. Consequently, the FL model outperforms the other models in classifying the data in our study, with the SVM model following closely. The KNN and NB models, while achieving fair performance, demonstrate comparatively lower accuracy in classification tasks.

5.6 Comparing the proposed model with other methods

Table 6 presents a comprehensive comparison of performance metrics achieved in various studies for the task of student grade prediction. The table includes traditional approaches and our proposed federated learning (FL) model. Each study's algorithms, dataset characteristics, and corresponding accuracy rates are highlighted, providing insights into the efficacy of the models in predicting student grades. Our study, employing FL with SVM, DT, KNN, and NB algorithms, demonstrates remarkable accuracy at 99%, outperforming traditional studies with accuracies of 88%, 60%, 80%, and 74%. This highlights the superiority of our approach in accurately classifying student grades. The tailored dataset used in our study was specifically optimized for the prediction task, potentially contributing to the higher accuracy and improved precision achieved. Additionally, we used AUC scores to assess the models' classification performance, offering a more detailed analysis of their TPR and FPR. This approach allowed for a deeper understanding of each model's strengths and weaknesses, aiding in identifying the most proficient algorithm for the classification task. In contrast, traditional studies predominantly relied on accuracy metrics, providing a more limited perspective on model performance. The unique aspects of our study, such as the utilization of federated learning, a comprehensive algorithm comparison, a tailored dataset, and the use of AUC scores, contributed to its outstanding performance in accurately predicting student grades. These distinctive features set our research apart from traditional studies and establish its significance in the field of student performance prediction.

Table 6 Comparison of performance metrics for student grade prediction using various machine learning algorithms

Study	Techniques	Algorithms	Dataset	Accuracy (%)
Anupam et al [61]	Machine learning algorithms	Linear regression K-neighbors regressor Decision tree Random forest XGBRegressor CatBoosting regressor	Missing values Duplicate values Check data types	88
Ismail et al. [62]	WEKA data mining	Decision tree Naïve Bayes MLP classification	first-year bachelor's students in a computer science course	60
Pandey et al. [63]	DT algorithm	KDD, WEKA J48, NBtreee, Reptree Simple Cart	Prediction of students academic performance	80
Veselina et al. [64]	SVM, DT, NB, RF	BayesNet MLP SMO Decision tress	Demographic (age, gender) Personality (self-efficacy, self-regulation) Academic (high school performance); Behavioral log data;Institutional (quality teaching approach)h	74
Proposed	Federated learning	SVM, DT, KNN, NB	Student grades (low, good, average, and drop).	99

6 Conclusion

Students' academic performance should be examined in order to boost both individual student outcomes and institution-wide outcomes. Moreover, educational data mining (EDM), in which data mining and machine learning techniques are employed for extracting data from educational warehouses, has been a significant and developing research subject focused on the prediction of students' learning achievement. A machine learning (ML) technique known as federated learning (FL) enables on-site data collection and processing without sacrificing data privacy or requiring data to be sent to a central server. We proposed a federated learning model to predict student learning performance by using the educational data of different institutes. Firstly, this work focused on a federated learning model to predict the grades of the students into low, good, average, and drop. Secondly, different optimized features are used that improve the accuracy of the models. Thirdly optimized dataset is evaluated on machine learning models. We used a support vector machine, decision tree, Naïve Bayes, K-nearest neighbors, and a federated learning model to train and test the dataset. Finally, the results of all the models (accuracy, precision, recall, and F1-score) are compared. Federated learning and KNN performed well on the training and testing dataset as compared to other models. Federated learning gives accuracy testing accuracy of 96.5 and performs better as compared to other models. This work can be expanded by using more features to train and test the machine learning and deep learning models. The security of the federated learning is improved by using more communication channels. The self-collected dataset should be used by using questionnaires and online form submissions. More deep-learning techniques should be used to improve the accuracy of models.

Acknowledgements The authors would like to thank China's National Key R & D Program for providing the experimental facilities to perform these experiments. The author would like to thank Artificial Intelligence and Data Analytics Lab (AIDA) CCIS Prince Sultan University for their support.

Author Contributions Farooq, Naseem, Mahmood, LI, Saba, Rehman, and Mustafa conceived the study and experimented. Farooq, Naseem, Mahmood, Saba, and Li reviewed, drafted, and revised the study. Naseem, Mahmood, Rehman and Mustafa contributed to the design and analyzed data. Farooq, Mahmood, Saba, Li, and Mustafa conducted the proofreading of the study. All authors have read and agreed to the published version of the manuscript.

Funding This study is supported by the National Key R &D Program of China with project no. 2020YFB2104402.

Availability of data and materials Not applicable.

Declarations

Ethical approval Not applicable.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Yassein NA, Helali RGM, Mohomad SB et al (2017) Predicting student academic performance in ksa using data mining techniques. J Inf Technol Softw Eng 7(5):1–5
2. Mahmood T, Li J, Pei Y, Akhtar F, Imran A, Rehman KU (2020) A brief survey on breast cancer diagnostic with deep learning schemes using multi-image modalities. IEEE Access 8:165779–165809
3. Siddique A, Jan A, Majeed F, Qahmash AI, Quadri NN, Wahab MOA (2021) Predicting academic performance using an efficient model based on fusion of classifiers. Appl Sci 11(24):11845
4. Pujianto U, Prasetyo WA, Taufani AR (2020) 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), pp. 348–353. IEEE
5. Alwarthan SA, Aslam N, Khan IU (2022) Predicting student academic performance at higher education using data mining: a systematic review. Appl Comput Intell Soft Comput 2022:8924028
6. Namoun A, Alshanqiti A (2020) Predicting student performance using data mining and learning analytics techniques: a systematic literature review. Appl Sci 11(1):237
7. Al-Mahmoud H, Al-Razgan M (2015) Arabic text mining a systematic review of the published literature 2002-2014. In: 2015 International Conference on Cloud Computing (ICCC), pp. 1–7. IEEE
8. Chen D, Gao D, Xie Y, Pan X, Li Z, Li Y, Ding B, Zhou J (2023) Fs-real: Towards real-world cross-device federated learning. arXiv preprint arXiv:2303.13363
9. Kumar M, Mehta G, Nayar N, Sharma A (2021) Emt: Ensemble meta-based tree model for predicting student performance in academics. In: IOP Conference Series: Materials Science and Engineering, vol. 1022, p. 012062. IOP Publishing
10. Makhtar M, Nawang H, WAN SHAMSUDDIN SN (2017) Analysis on students performance using naïve bayes classifier. J Theoret Appl Inf Technol **95**(16)
11. Altabrawee H, Ali OAJ, Ajmi SQ (2019) Predicting students' performance using machine learning techniques. J Univ BABYLON Pure Appl Sci 27(1):194–205
12. Apolinar-Gotardo M (2019) Using decision tree algorithm to predict student performance. Indian J Sci Technol 12:5
13. Karthikeyan VG, Thangaraj P, Karthik S (2020) Towards developing hybrid educational data mining model (hedm) for efficient and accurate student performance evaluation. Soft Comput 24(24):18477–18487
14. Dhilipan J, Vijayalakshmi N, Suriya, S., Christopher A (2021) Prediction of students performance using machine learning. In: IOP Conference Series: Materials Science and Engineering, vol. 1055, p. 012122. IOP Publishing
15. Li S, Liu T (2021) Performance prediction for higher education students using deep learning. Complexity 2021:1–10
16. Khan MS, Mansour M, Khadar S, Mallick Z (2020) Evaluating healthcare performance using fuzzy logic. S Afr J Ind Eng 31(1):133–143
17. Zhang J, Zhu H, Wang F, Zhao J, Xu Q, Li H et al (2022) Security and privacy threats to federated learning: Issues, methods, and challenges. Secur Commun Netw
18. Hu Z, Shaloudegi K, Zhang G, Yu Y (2022) Federated learning meets multi-objective optimization. IEEE Trans Netw Sci Eng 9(4):2039–2051
19. Chen H, Wang H, Jin D, Li Y (2023) Advancements in federated learning: Models, methods, and privacy. arXiv preprint arXiv:2302.11466
20. Realinho V, Machado J, Baptista L, Martins MV (2022) Predicting student dropout and academic success. Data 7(11):146
21. Tyler JH, Taylor ES, Kane TJ, Wooten AL (2010) Using student performance data to identify effective classroom practices. Am Econ Rev 100(2):256–260
22. Kaur K, Kaur K (2015) Analyzing the effect of difficulty level of a course on students performance prediction using data mining. In: 2015 1st International Conference on Next Generation Computing Technologies (NGCT), pp. 756–761. IEEE
23. Bhardwaj BK, Pal S (2012) Data mining: A prediction for performance improvement using classification. arXiv preprint arXiv:1201.3418
24. Pandey M, Taruna S (2016) Towards the integration of multiple classifier pertaining to the student's performance prediction. Perspect Sci 8:364–366

25. Ch'ng LK (2024) Standing on the shoulders of generative ai. In: Transforming Education With Generative AI: Prompt Engineering and Synthetic Content Creation, pp. 1–21. IGI Global
26. Chai CS, Chiu TK, Wang X, Jiang F, Lin X-F (2022) Modeling Chinese secondary school students' behavioral intentions to learn artificial intelligence with the theory of planned behavior and self-determination theory. Sustainability 15(1):605
27. Chu Y-W, Hosseinalipour S, Tenorio E, Cruz L, Douglas K, Lan A, Brinton C (2022) Mitigating biases in student performance prediction via attention-based personalized federated learning. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 3033–3042
28. Banabilah S, Aloqaily M, Alsayed E, Malik N, Jararweh Y (2022) Federated learning review: Fundamentals, enabling technologies, and future applications. Inf Process Manag 59(6):103061
29. Zhang C, Xie Y, Bai H, Yu B, Li W, Gao Y (2021) A survey on federated learning. Knowl-Based Syst 216:106775
30. Wen J, Zhang Z, Lan Y, Cui Z, Cai J, Zhang W (2023) A survey on federated learning: challenges and applications. Int J Mach Learn Cybern 14(2):513–535
31. Parlak B, Uysal AK (2023) A novel filter feature selection method for text classification: extensive feature selector. J Inf Sci 49(1):59–78
32. Parlak B, Uysal AK (2021) The effects of globalisation techniques on feature selection for text classification. J Inf Sci 47(6):727–739
33. Janan F, Ghosh SK (2021) Prediction of student's performance using support vector machine classifier. In: Proc. Int. Conf. Ind. Eng. Oper. Manag, pp. 7078–7088
34. Mahmood T, Li J, Pei Y, Akhtar F, Imran A, Yaqub M (2021) An automatic detection and localization of mammographic microcalcifications roi with multi-scale features using the radiomics analysis approach. Cancers 13(23):5916
35. Mahmood T, Li J, Pei Y, Akhtar F, Rehman MU, Wasti SH (2022) Breast lesions classifications of mammographic images using a deep convolutional neural network-based approach. PLoS ONE 17(1):0263126
36. Mahmood T, Li J, Pei Y, Akhtar F (2021) An automated in-depth feature learning algorithm for breast abnormality prognosis and robust characterization from mammography images using deep transfer learning. Biology 10(9):859
37. Rehman KU, Li J, Pei Y, Yasin A, Ali S, Mahmood T (2021) Computer vision-based microcalcification detection in digital mammograms using fully connected depthwise separable convolutional neural network. Sensors 21(14):4854
38. Mahmood T, Li J, Pei Y, Akhtar F, Jia Y, Khand ZH (2021) Breast mass detection and classification using deep convolutional neural networks for radiologist diagnosis assistance. In: 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), pp. 1918–1923. IEEE
39. Sarker IH (2022) Ai-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. SN Comput Sci 3(2):158
40. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L (2021) Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. J Big Data 8:1–74
41. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297
42. Moguerza JM, Muñoz A (2006) Support vector machines with applications. Stat Sci 21(3):322–336. https://doi.org/10.1214/088342306000000493
43. Ali S, Li J, Pei Y, Khurram R, Rehman KU, Mahmood T (2022) A comprehensive survey on brain tumor diagnosis using deep learning and emerging hybrid techniques with multi-modal mr image. Arch Comput Methods Eng 29(7):4871–4896
44. Yaqub M, Jinchao F, Arshid K, Ahmed S, Zhang W, Nawaz MZ, Mahmood T (2022) Deep learning-based image reconstruction for different medical imaging modalities. Comput Math Methods Med 2022:8750648
45. Marjanović M, Kovačević M, Bajat B, Voženílek V (2011) Landslide susceptibility assessment using svm machine learning algorithm. Eng Geol 123(3):225–234
46. Iqbal S, Qureshi AN, Li J, Choudhry IA, Mahmood T (2023) Dynamic learning for imbalance data in learning chest x-ray and ct images. Heliyon
47. Quinlan JR (2014) C4. 5: Programs for Machine Learning. Elsevier, Amsterdam
48. Iqbal S, Qureshi NA, Li J, Mahmood T (2023) On the analyses of medical images using traditional machine learning techniques and convolutional neural networks. Arch Comput Methods Eng 30(5):3173–3233

49. Quinlan JR (1986) Induction of decision trees. Mach Learn 1:81–106
50. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. CRC press, Florida, Boca Raton
51. Divyabharathi Y, Someswari P (2018) A Framework for Student Academic Performance Using Naïve Bayes Classification. JAET
52. Iqbal S, Qureshi AN, Ullah A, Li J, Mahmood T (2022) Improving the robustness and quality of biomedical cnn models through adaptive hyperparameter tuning. Appl Sci 12(22):11870
53. Jabbar A, Naseem S, Mahmood T, Saba T, Alamri FS, Rehman A (2023) Brain tumor detection and multi-grade segmentation through hybrid caps-vggnet model. IEEE Access 11(1):72518–72536
54. Rehman A, Sadad T, Saba T, Hussain A, Tariq U (2021) Real-time diagnosis system of covid-19 using x-ray images and deep learning. It Professional 23(4):57–62
55. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA (2017) Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics, pp. 1273–1282. PMLR
56. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
57. Saba T, Khan SU, Islam N, Abbas N, Rehman A, Javaid N, Anjum A (2019) Cloud-based decision support system for the detection and classification of malignant cells in breast cancer using breast cytology images. Microsc Res Tech 82(6):775–785
58. Sandra L, Lumbangaol F, Matsuo T (2021) Machine learning algorithm to predict student's performance: a systematic literature review. TEM J 10(4):1919–1927
59. Naseem S, Mahmood T, Saba T, Alamri FS, Bahaj SA, Ateeq H, Farooq U (2023) Deepfert: An intelligent fertility rate prediction approach for men based on deep learning neural networks. IEEE Access
60. Chen H-C, Prasetyo E, Tseng S-S, Putra KT, Kusumawardani SS, Weng C-E (2022) Week-wise student performance early prediction in virtual learning environment using a deep explainable artificial intelligence. Appl Sci 12(4):1885
61. Khan A, Ghosh SK (2021) Student performance analysis and prediction in classroom learning: a review of educational data mining studies. Educ Inf Technol 26:205–240
62. Ismail NH, Ahmad F, Aziz AA (2013) Implementing weka as a data mining tool to analyze students' academic performances using naïve bayes classifier. In: UniSZA Postgraduate Research Conference
63. Pandey M, Sharma VK (2013) A decision tree algorithm pertaining to the student performance analysis and prediction. Int J Comput Appl 61(13):1–5
64. Nedeva V, Pehlivanova T (2021) Students' performance analyses using machine learning algorithms in weka. In: IOP Conference Series: Materials Science and Engineering, vol. 1031, pp 012061. IOP Publishing

Authors and Affiliations

Umer Farooq[1,2] · Shahid Naseem[2] · Tariq Mahmood[3,4] · Jianqiang Li[1,5] · Amjad Rehman[3] · Tanzila Saba[3] · Luqman Mustafa[2]

✉ Tariq Mahmood
tmsherazi@ue.edu.pk

Umer Farooq
umerfarooq@ue.edu.pk

Shahid Naseem
shahid.naseem@ue.edu.pk

Jianqiang Li
lijianqiang@bjut.edu.cn

Amjad Rehman
arkhan@psu.edu.sa

Tanzila Saba
tsaba@psu.edu.sa

Luqman Mustafa
mluqman564@gmail.com

[1] Faculty of Information Technology, Beijing University of Technology, Beijing 100024, China

[2] Faculty of Information Sciences, Division of Science and Technology, University of Education, Lahore 54000, Pakistan

[3] Artificial Intelligence and Data Analytics (AIDA) Lab, CCIS Prince Sultan University, Riyadh 11586, Kingdom of Saudi Arabia

[4] Faculty of Information Sciences, University of Educatio, Vehari Campus, Vehari 61161, Pakistan

[5] Beijing Engineering Research Center for IoT Software and Systems, Beijing 100124, China