IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 September 2024, accepted 21 October 2024, date of publication 8 November 2024, date of current version 18 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3488743

RESEARCH ARTICLE

Evaluation of Feature Transformation and Machine Learning Models on Early Detection of Diabetes Mellitus

AHMED ALI LINKON[1], INSHAD RAHMAN NOMAN[2], MD RASHEDUL ISLAM[3], JOY CHAKRA BORTTY[1], KANCHON KUMAR BISHNU[2], ARAF ISLAM[1], RAKIBUL HASAN[3], (Member, IEEE), AND MASUK ABDULLAH[4]

[1]Department of Computer Science, Westcliff University, Irvine, CA 92614, USA
[2]Department of Computer Science, California State University, Los Angeles, CA 90032, USA
[3]Department of Business Administration, Westcliff University, Irvine, CA 92614, USA
[4]Department of Vehicles Engineering, Faculty of Engineering, University of Debrecen, 4028 Debrecen, Hungary

Corresponding author: Rakibul Hasan (r.hasan.179@westcliff.edu)

ABSTRACT The increasing prevalence of diabetes necessitates the development of effective early detection methods to mitigate its health impacts. This paper investigates the impact of feature transformation and machine learning (ML) models on the early detection of diabetes using a binary tabular classification dataset. We explore three feature transformation techniques, no transformation, normalization, and min-max scaling, to assess their influence on the performance of various ML models. To comprehensively evaluate the effectiveness of these preprocessing techniques, we experimented with twelve different ML models, including both traditional algorithms and ensemble methods. A publicly available dataset has been used for this research, containing 768 samples and 8 features. To ensure their effectiveness, the models are assessed using several evaluation metrics, including accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Among the ML models, Light Gradient Boosting Machine (LGBM) achieved the highest accuracy of 82.91% when min-max scaling was applied to the data. Our results demonstrate the varying effectiveness of different combinations of feature transformation techniques and ML models in enhancing diabetes detection performance. Furthermore, it has been observed that the ensemble models generally achieved better performance than traditional ML models. These findings provide valuable insights for optimizing preprocessing and model selection strategies in the development of robust early diabetes detection systems.

INDEX TERMS Machine learning, diabetes detection, feature transformation, data preprocessing.

I. INTRODUCTION

Diabetes mellitus, more commonly known as diabetes, is an emerging worldwide health crisis that has extensive consequences for both individuals and healthcare systems [1], [2]. Diabetes, which is distinguished by its chronic hyperglycemia, is subdivided into various varieties [3], the most prevalent of which are Type 1, Type 2, and gestational diabetes. The condition's intricate nature and initially mute progression emphasize the critical importance of timely identification and precise diagnosis [4], [5], [6]. According to statistical findings from 2017, 451 million individuals globally have diabetes; by 2045, that number is expected to rise to 693 million [7]. Other statistical research illustrates the severity of diabetes; it states that half a billion people globally have the disease, and that figure will rise to 25% invasive, cost-effective, and more proactive approaches in the medical field to identify the risk of diabetes at an earlier stage [8]. Although efficacious, conventional diagnostic techniques, including fasting glucose tests and the HbA1c blood test, depend on biochemical analysis and frequently fall short in discerning the subtle advancement of the ailment

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

during its early phases. Diabetes mellitus, more commonly known as diabetes, is an emerging worldwide health crisis that has extensive consequences for both individuals and healthcare systems [1], [2]. Diabetes, which is distinguished by its chronic hyperglycemia, is subdivided into various varieties [3], the most prevalent of which are Type 1, Type 2, and gestational diabetes. The condition's intricate nature and initially mute progression emphasize the critical importance of timely identification and precise diagnosis [4], [5], [6]. According to statistical findings from 2017, 451 million individuals globally have diabetes; by 2045, that number is expected to rise to 693 million [7]. Other statistical research illustrates the severity of diabetes; it states that half a billion people globally have the disease, and that figure will rise to 25% invasive, cost-effective, and more proactive approaches in the medical field to identify the risk of diabetes at an earlier stage [8]. Although efficacious, conventional diagnostic techniques, including fasting glucose tests and the HbA1c blood test, depend on biochemical analysis and frequently fall short in discerning the subtle advancement of the ailment during its early phases.

Amidst the dynamic nature of the current environment, the introduction of ML and artificial intelligence (AI) signifies a transformative period in the healthcare sector [9], presenting novel approaches to identifying and controlling intricate ailments such as diabetes. These technologies have demonstrated tremendous promise in analyzing enormous datasets to discover patterns and predictive indicators that may not be readily discernible to the naked eye. Diabetes continues to be a prominent contributor to mortality on a global scale, underscoring the criticality of sophisticated diagnostic instruments capable of accurately and efficiently predicting heart disease. The increasing sense of urgency has generated a burgeoning fascination with the application of ML [10], [11], [12], [13] and Deep Learning (DL) [14], [15], [16] methods to improve the predictive precision and effectiveness of diabetes detection. ML and DL, which fall under the umbrella of artificial intelligence (AI) [17], [18], [19], [20], provide robust functionalities for examining intricate datasets. It facilitates the detection of correlations and patterns that may elude human analysts at first glance. ML algorithms can analyze extensive quantities of patient data, including clinical symptoms, medical histories, laboratory results, and imaging studies, to forecast the probability of diabetes with an unprecedented degree of precision that was hitherto unattainable using only conventional statistical approaches. ML and DL have been utilized in various fields, including hand-written digit recognition [21], [22], diagnosis of Autism Spectrum Disorder [23], [24], [25], depression detection [26], [27], [28], object detection [29], [30], and suspicious activity detection [31], [32], [33], etc. Furthermore, the capacity of ML models to consistently acquire knowledge from fresh data offers a prospect for creating dynamic diagnostic instruments that can progress in tandem with emergent patterns in the manifestation of diabetes and the results of treatments [34]. Accurate predictions in this domain necessitate the integration of complex data. Therefore, the incorporation of ML methods into the diagnostic procedure signifies a substantial progression in the domain of diabetes, holding the potential to revolutionize the existing strategy for identifying and treating diabetes [35]. By optimizing diagnostic instruments' predictive accuracy and efficiency, ML can enable timely interventions, customize treatment approaches according to unique patient profiles, and ultimately enhance the prognosis of those susceptible to diabetes [36]. Through the utilization of ML models' predictive functionalities, healthcare providers can detect individuals at risk of developing a particular condition prematurely and execute intervention strategies tailored to target the distinct factors contributing to each patient's condition. As a whole, the introduction of ML and artificial intelligence (AI) into the field of diabetes signifies a turning point in the battle against diabetes [37]. Through the utilization of these technologies to analyze intricate medical data, the medical community has achieved an unprecedented level of capability in promptly identifying individuals who are at risk of developing diabetes and in formulating individualized treatment approaches that have the potential to transform the trajectory of this perilous condition substantially. A thorough study of ML algorithms and various preprocessing techniques must be explored for this. In this study, we have addressed the following issues.

The main contribution of this study includes:

1) This study rigorously assesses the impact of different feature transformation techniques—no transformation, normalization, and min-max scaling—on the performance of machine learning models in early diabetes detection. This evaluation provides critical insights into how preprocessing methods affect model accuracy and other performance metrics.
2) The research compares twelve machine learning models, including traditional algorithms and ensemble methods, to determine their effectiveness in detecting diabetes. This analysis identifies the most suitable models for this task, contributing valuable knowledge to medical diagnostics using machine learning.
3) The study reveals that ensemble models outperform traditional machine-learning models in early diabetes detection. This finding supports the growing consensus on the effectiveness of ensemble methods and provides practical recommendations for improving diagnostic accuracy in healthcare applications.
4) Lastly, this study establishes a foundation for subsequent investigations concerning ML and the detection of diabetes. By identifying deficiencies in the existing body of literature and proposing potential avenues for additional research, our objective is to stimulate ongoing innovation and inquiry into the utilization of AI in healthcare.

The remaining sections of this paper are structured as follows: Section two details the datasets, experimental procedure, and ML and DL techniques employed in this research. The outcomes of multiple experiments are presented in section

three. Finally, section four highlights potential challenges and future avenues for applying ML and DL in diabetes detection.

II. RELATED WORKS

Many works have been conducted to diagnose diabetes using ML-based models. Sisodia et al. [38] utilized three distinct ML classifiers, including Decision Tree (DT), Support Vector Machines (SVM), and Naive Bayes (NB), to most accurately predict the likelihood of diabetes. They provided evidence that NB exhibits superior performance, as evidenced by its AUC of 0.819. Pradhan et al. [39] proposed genetic programming for the forecasting of diabetes, where their design performed better than other techniques they implemented. Mohan and Jain [40] analyzed and predicted diabetes using the SVM algorithm in conjunction with four distinct types of algorithms: linear, polynomial, RBF, and sigmoid. The authors achieved varying accuracies across distinct kernels, with values of 51% in 2030 and 2045 [41]. It is crucial to put into practice prompt intervention and effective management strategies in order to prevent or delay the development of complications linked to diabetes, such as cardiovascular diseases, renal failure, vision impairment, and neuropathy, which have a significant negative impact on quality of life and increase the risk of mortality. Traditional diagnostic methods, while effective, often fail to capitalize on the nuanced patterns hidden in the vast amounts of data generated by modern healthcare systems. Traditionally, diabetes detection has relied on clinical screenings and biochemical tests, which often require specific conditions and expert interpretation. However, these methods can be invasive, costly, and typically detect the disease only after it has manifested to a considerable degree. Hence, a compelling requirement is needed for non-spanning ranges of AUC values between 0.69 to 0.82. Goyal et al. [42] developed a digital home health monitoring system to detect diabetes. In addition, the researchers incorporated the Pima Indian dataset into their study. They implemented conditional decision-making to forecast blood pressure status and SVM, KNN, and decision trees to predict diabetes. SVM performed the best among these models, achieving an accuracy of 75%, surpassing the performance of alternative classifier algorithms. Hasan et al. [43] utilized various ensemble method-based ML algorithms to predict diabetes. The authors utilized AUC as a metric for assessing accuracy. In conclusion, the proposed ensemble classifier achieved an AUC value of 0.95. Jackins et al. [44] proposed an ML-based multi-disease prediction system that can be used to detect diabetes. The authors claim that the Naive Bayes algorithm outperformed the random forest method by a margin of 0.43 percent in terms of accuracy.

Kumari et al. [45] attempted to predict diabetes using an ensemble approach based on fuzzy voting classifiers. The fuzzy voting classifier, as proposed, achieved the maximum overall accuracy of 0.791 and F1 score of 0.716. Pranto et al. [46] developed an automated diabetes prediction method to detect diabetes among female patients in Bangladesh utilizing two datasets and multiple ML methods. DT and k-Nearest Neighbors (KNN) achieved significant accuracy in the experiment. With the Pima Indian dataset, Ramesh et al. [47] created an automated and remote diabetes predicting system. The authors used SMOTE, feature selection, and feature scaling, among other data preparation methods. A maximum accuracy of 83.2% was reached using SVM with Radial Basis Function (RBF) kernel. An Android application uses the suggested ML framework. Deberneh and Kim [48] presented an ML-based type 2 diabetes early prediction system. We employ various ML techniques to predict this illness using the patient data of the previous year. The RF and SVM classifiers achieved the greatest F1 score of 74%. Ahmed et al. [49] built a website to forecast diabetes automatically. This paper used several well-known ML techniques along with two open-source datasets. With an accuracy of 0.968, DT and Random Forest (RF) classifiers performed the best in this investigation. Using ML algorithms and sophisticated feature selection, Olisah et al. [50] accomplished diabetes mellitus prediction. The authors used the LMCH Iraqi and Pima Indian databases, two open-source datasets. Preprocessing based on polynomial regression was used to forecast the missing data. The suggested Deep Neural Networks (DNN) method with the optimized hyperparameters achieved the greatest accuracies of 0.972 and 0.973.

III. PROPOSED METHODOLOGY

This task comprises several essential stages. The comprehensive workflow is illustrated in Figure 1. We began by amassing a publicly accessible benchmark datasets. Once that is determined, the datasets are examined for missing values. To convert the categorical values present in the datasets to numeric values. Following the conclusion of the preprocessing stage, ML and DL models are developed. Subsequently, the models undergo testing using the test data after being instructed using the training data.

A. DATASET

The dataset utilized in this study was first introduced in [51], and the dataset is publicly accessible [52]. The dataset was collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh, and approved by a doctor. It contains a total of 17 columns, including 16 features and 1 target column (identifies if a patient has diabetes or not).

B. DATA PREPROCESSING

Data preparation is the first step in creating an ML model, signaling the beginning of the process. Real-world data is frequently insufficient, inconsistent, inaccurate, and lacks essential attribute values. Data preparation involves cleaning, organizing, and formatting raw data to be suitable for ML models. In this study, we have employed a variety of data preparation approaches. The feature importance graph is presented in Figure 2.

1) CATEGORICAL DATA ENCODING

Category data encoding is converting categorical variables into numerical variables to make them usable in ML models. Categorical data comprises variables grouped into different categories, such as colors, locations, or types of items. Given that the majority of ML models rely on mathematical equations, it is essential to transform categorical data into numerical data to prevent any issues. We have converted the category data in the datasets into numerical values. We utilized the Label Encoder function from the sklearn package for this task.

2) FEATURE CORRELATION ANALYSIS

A critical aspect of statistical analysis and data exploration is the comprehension of the interrelationships among variables. In pursuit of this objective, we utilized the Pearson correlation coefficient, a commonly employed technique for quantifying the linear association between two continuous variables within our dataset. The range of values for the Pearson correlation coefficient (r) is -1 to $+1$ [53]. The coefficient's magnitude denotes the correlation's intensity, whereas the sign indicates the direction. The Pearson correlation method was utilized to examine the relationships between every pair of continuous variables in the dataset [54]. By identifying substantially correlated variables, one can acquire valuable insights into latent patterns and connections that may indicate the dataset's intrinsic causal relationships or interdependencies.

3) FEATURE IMPORTANCE ANALYSIS

Feature Importance Calculation: When constructing predictive models, it is critical to comprehend the significance of every feature concerning the target variable. An efficacious approach to assess this level of importance is by employing the computation of Mutual Information (MI) scores. In contrast to more straightforward linear metrics, MI offers a broader measure that can encompass any relationship between variables, whether nonlinear or linear [55]. If two variables are independent, then the score is zero. Conversely, a larger score signifies an enhanced interdependence or correlation among the variables. In the context of feature selection, the MI score of a feature concerning the target variable indicates the degree to which knowledge of the feature reduces uncertainty regarding the target.

4) FEATURE SCALING

An integral part of our research was preparing the data, with a special emphasis on feature scaling. Feature scaling is essential in ML models to standardize the range of independent variables, which aids algorithms in converging faster and achieving higher performance. Scaling the characteristics of our dataset assures that no one feature has a dominant influence because of its scale, considering the dataset's heterogeneous nature. We tested four distinct feature scaling methods in comparison to a baseline model without any feature scaling to assess their influence on the efficacy of ML models in predicting heart disease. Table 1 illustrated the features of data set and their possible values.

The standard scaler method was used to standardize the feature distribution by removing the mean and dividing by the standard deviation, resulting in features centered around zero and with a standard deviation of one [56].

This approach is very efficient when the features are normally distributed, although it may be used with any distribution [57]. Standard scaling helps stabilize the convergence of gradient descent algorithms by ensuring all features are on a comparable scale. It can be represented as:

$$z = \frac{\boldsymbol{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{1}$$

Min-Max Scaling was used to standardize the features to a certain range, usually [0, 1] [58]. We standardized each characteristic by deleting its minimum value and dividing by the range to guarantee equal contribution to the final forecast.

This method is beneficial for situations where the parameters must fall inside a certain range and is commonly employed when the distribution is non-Gaussian.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

Normalization, specifically L2 scaling, was employed to adjust individual samples to have a norm of one [59], [60]. This method is beneficial in situations where the Euclidean distance between instances is significant [61]. This scaling approach improves the efficiency of algorithms that use distance between instances by normalizing each instance's feature vector to have an Euclidean length of one. Models were trained without applying any scaling to the dataset to evaluate the inherent impact of feature scaling.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3}$$

This method enabled a straightforward comparison to assess the efficacy of each scaling methodology in improving model performance. Integrating these feature scaling strategies into our preprocessing workflow allowed for a thorough assessment of their impact on model correctness and convergence. We compared the heart disease prediction models with and without feature scaling to determine the benefits of each strategy and find the best preprocessing procedures for this particular application. The comparative analysis emphasized the significance of feature scaling in ML pipelines and offered guidance on choosing suitable scaling strategies depending on the data properties and model needs.

C. MACHINE LEARNING MODELS

In this investigation, we implemented ML techniques that are widely utilized across industries due to their straight forwardness and ability to generalize. The methodologies employed in this investigation are delineated as follows:

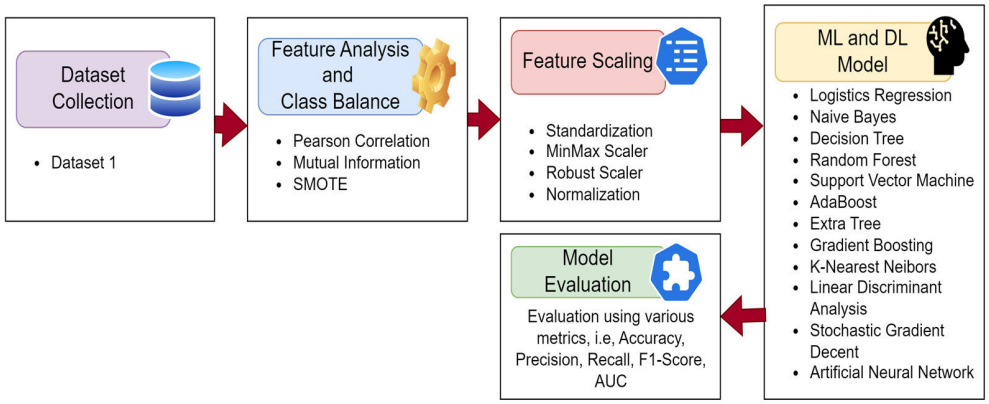

FIGURE 1. The workflow of the proposed methodology. The workflow begins with dataset collection. Then the data has been preproccessed. The various ML models has been developed after that. Finally, the models has been trained, and tested using various metrics.

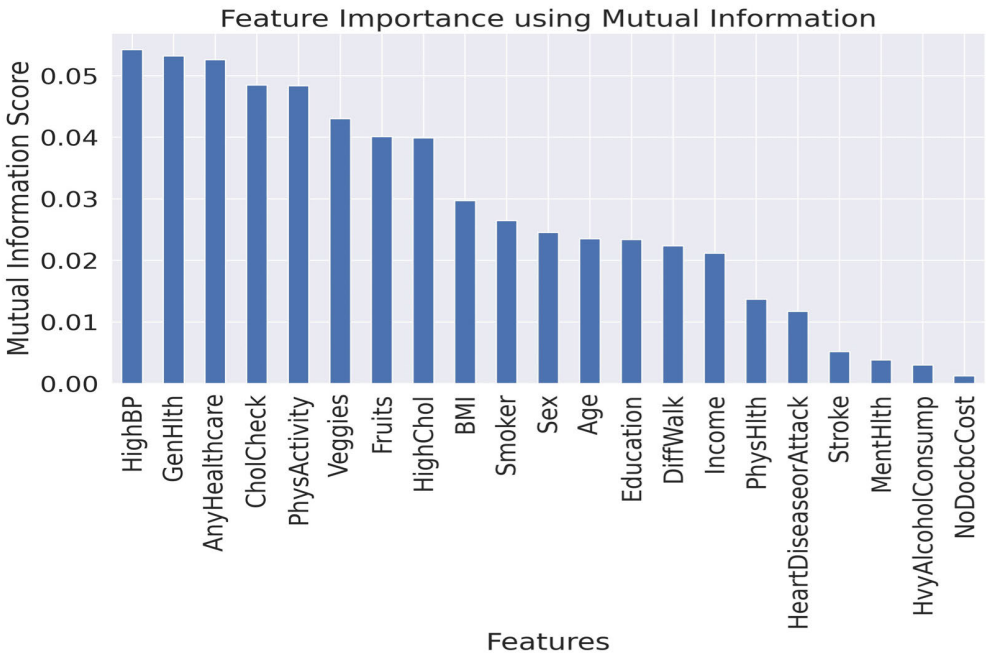

FIGURE 2. MI score of the features in the dataset.

1) LOGISTIC REGRESSION

As a predictive analysis algorithm, LR is predominantly applied to binary classification problems [62]. It employs a logistic function, which is bounded between 0 and 1, to estimate probabilities. This feature renders it an exceptional instrument in situations where the probability of an occurrence must be predicted, such as determining the spam status of an email or the malignant or benign nature of a tumor. Logistic regression is fundamentally concerned with generating probability scores for observations and classifying them into two distinct categories via a decision threshold (typically set at 0.5).

2) NAIVE BAYES

NB is an algorithm for ML and predictive modeling that is both straightforward and potent [63]. It operates under the assumption that the existence of a specific feature in a given class is independent of the existence of any other feature (hence the term "naive"). It is founded upon Bayes' theorem. Although NB simplifies the process, the resulting models can be quite accurate, particularly when used for sentiment analysis, spam detection, and document classification tasks. The algorithm is well suited for high-dimensional data because of its efficiency and speed. It is frequently employed in text classification tasks involving sizable datasets, where the algorithm's simplicity can confer a substantial benefit.

3) DECISION TREE

A DT is a nonparametric supervised learning algorithm utilized for classification and regression tasks [64]. At its essence, a DT employs a tree-like representation of choices and their potential ramifications, encompassing resource

TABLE 1. The features in the data set and their possible values.

Features Name	Possible Value
Regular Medicine	No
	Yes
	40-49
	Less than 40
Age	50-59
	60 or older
	High
	Normal
	Low
BP Level	Low
	High
	Normal
Family_Diabetes	No
	Yes
highBP	Yes
	No
BMI	15-45
	Sometimes
	Not at all
Stress	Very often
	Always
	One hr or more
	Less than half an hr
Physically Active	None
	More than half an hr
Pregnancies	0-4
Sleep	4-11
Pdiabetes	Yes
	No
Smoking	No
	Yes
Sound Sleep	0-11
Urination Freq	Not much
	Quite often
Alcohol	No
	Yes
	Occasionally
	Very often
Junk Food	Often
	Always
Gender	Male
	Female

expenses, utility, and random event outcomes. It commences with a solitary node that branches into potential outcomes; each of those branches subsequently connects to additional nodes that branch off into additional possibilities. The aforementioned procedure persists until it reaches a leaf node, which subsequently yields the DT's output pertaining to the corresponding input features. The routes connecting the root to the leaf signify classification rules or regression routes. At each node, a determination is executed using regression estimates or the attribute that provides the most effective separation of classes, as determined by a specific criterion. The straightforwardness of DTs facilitates their comprehension, representation, and elucidation, thereby substantially bolstering their prevalence in decision-making endeavors that demand transparency.

4) RANDOM FOREST

Data scientists frequently utilize the RF algorithm, one of the most widely recognized algorithms [65], [66]. A frequently employed Supervised ML Algorithm, it is utilized to tackle problems related to classification and regression. The algorithm is composed of a multitude of DTs, with each DT analyzing a unique subset of the dataset and computing the mean to enhance the prediction's precision. Multiple classifiers are integrated into the EL strategy to tackle a complex problem and improve the model's performance. By aggregating the results, RF is an ensemble technique that reduces overfitting and outperforms a single DT.

5) SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a potent supervised ML technique for classification and regression tasks [67]. Nevertheless, it is predominantly utilized in categorization difficulties [68]. SVM classifies data by identifying the hyperplane that most effectively separates the dataset into different groups. SVM aims to find the best hyperplane that optimizes the margin between various classes in the training dataset. Support vectors are the data points nearest to the hyperplane and significantly impact the hyperplane's location and orientation. SVM enhances classification accuracy by increasing the margin between classes through support vectors [69]. The SVM technique is adaptable and can handle linear and non-linear separations using kernel functions, making it suitable for a broad range of data formats and prediction problems.

6) K-NEAREST NEIGHBORS

K-Nearest Neighbors (KNN) is an instance-based or lazy learning technique that approximates the function locally and delays all computation until function evaluation [70]. It is a versatile tool utilized for classification and regression problems because of its simplicity and effectiveness. The KNN algorithm operates by calculating the distances between a query and all data instances, choosing the designated number ''K'' of the nearest examples, and then either picking the most common label (for classification) or averaging the labels (for regression). Choosing the parameter K is crucial. A lower K number increases the impact of noise on the output, while a large value leads to increased computational costs and the potential inclusion of points from other classes [71]. KNN is obvious and simple to understand however, its computational speed decreases notably as the dataset size increases.

7) AdaBoost

AB, also called Adaptive Boosting [71], [72], [73], is an ensemble learning technique mainly used for binary classification tasks. AB's fundamental concept is amalgamating several weak classifiers to form a robust classifier. A weak classifier is a classifier that performs marginally better than random guessing. AB gives weights to each training instance and adjusts them as training advances. Classifiers are trained successively, with each new classifier emphasizing the training cases misclassified by preceding classifiers. The final prediction is generated by combining the predictions of all classifiers using a weighted majority vote or total. Boosting's

adaptability stems from its emphasis on classifying challenging examples and its strategy of assigning greater weight to classifiers with superior performance. AB enhances the accuracy of weak learning models, making it a potent tool for enhancing model performance.

8) GRADIENT BOOSTING MACHINE

Gradient Boosting Machine (GBM) [72], [73] is a potent ML method that enhances DTs by adding weak learners sequentially to form a robust prediction model. Every new tree aims to rectify the mistakes of the trees constructed before it. Gradient Boosting Machine (GBM) employs the gradient descent approach to reduce errors in sequential models. It modifies the importance of a data point according to its prior categorization. When an observation is misclassified, its weight is increased, and vice versa. This method enables the model to focus more on challenging situations for classification, leading to enhanced accuracy. Gradient Boosting Machine (GBM) is versatile, applying to both regression and classification tasks, and has proven effective in addressing several real-world situations. GBM is a popular choice among data scientists for achieving excellent performance in predictive jobs due to its ability to handle different data types and produce strong predictions.

9) LIGHT GRADIENT BOOSTING MACHINE

LGBM, akin to alternative gradient-boosting methodologies, employs an incremental model construction approach [74]. It constructs a series of DTs, wherein each succeeding tree is designed to correct the errors committed by its antecedent. Assembled from the final model, this weighted sum of the individual trees. The model is denoted as follows:

$$\boldsymbol{F(x)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{f_i(x)} \tag{4}$$

where NN represents the quantity of trees and $f_i(x)$ denotes the forecast of the i-th tree. The process of training entails the minimization of a loss function. In LightGBM, the objective function is composed of two components: the regularization term and the loss function [75].

10) CatBoost

CatBoost (CB) is a gradient-boosting algorithm that has been purposefully developed to process categorical features [75] efficiently. The system integrates novel methodologies to attain optimal performance and resilience, specifically when confronted with situations involving diverse data types and extensive datasets. CB minimizes a differentiable loss function L(y_i,F(x_i)), where x_i is the feature vector, y_i is the target variable, and $F(x_i)$ is the predicted value for the i-th instance. To minimize the loss function, CB assembles an ensemble of DTs in a sequential fashion. The model acquires knowledge of the gradient of the loss function in relation to the preceding predictions at each iteration t.

$$g_i^{(t)} = \frac{\partial F(x_i)}{\partial L(y_{i,} F\,(x_i))} \tag{5}$$

CatBoost is a library for gradient boosting that has been purposefully developed to handle categorical features. It employs a gradient-boosting implementation with DT and integrates innovative methods to manage categorical variables [76] efficiently.

11) LINEAR DISCRIMINANT ANALYSIS

LDA [76], [77] is a supervised ML technique specifically designed for classification applications. It is a method employed to identify a linear combination of characteristics that most effectively distinguishes the different classes within a dataset. LDA functions by mapping the data onto a reduced dimensional space that optimizes the distance between the different classes. It does this by identifying a group of linear discriminants that optimize the variance ratio across classes to variance within classes. It identifies the directions in the feature space that most effectively distinguishes between the various data classes. LDA presupposes that the data follows a Gaussian distribution and that the covariance matrices of the various classes are identical. The assumption is that the data is linearly separable, allowing a linear decision boundary to categorize the various classes effectively.

12) ARTIFICIAL NEURAL NETWORK

A feed-forward Neural Network (FFN) is a type of network that creates a directed graph with nodes and edges. Data is transmitted along these edges from a node to the next without forming a cycle. The ANN [78], [79] is a variant of FFN with three or more layers: an input layer, one or more hidden layers, and an output layer. Each of these layers contains many neurons or units, as defined in mathematical notation. A hyper parameter tuning strategy is employed [80, 81] to determine the number of hidden layers in an ANN. Information is transferred from one layer to the next without taking into account previous values, and all neurons in each layer are connected, as documented in sources [15], [80]. ANN with n hidden layers may be expressed mathematically as follows:

$$H(x) = H_n(H_{n-1}(H_{n-2}(\cdots(H_1(x))))) \tag{6}$$

IV. EVALUATION

A. EVALUATION METRICS

A variety of metrics were utilized to assess the performance of the models, with the primary ones being accuracy, recall, precision, and F1-score. The accuracy metric quantifies the proportion of correct predictions relative to the total number of predictions. The accuracy metric quantifies the proportion of correct positive predictions relative to the overall number of positive predictions. Recall can be conceptualized as the ratio of precise positive forecasts to the sum of precise positive forecasts and erroneous negative predictions. The harmonic mean of recall and precision constitutes

the F1score. The Area Under the Curve (AUC) is a metric utilized to assess the binary classification model's performance. As the AUC increases, so does the model's ability to distinguish between positive and negative instances. It is a prevalent metric utilized in ML to evaluate the performance of classification algorithms.

B. RESULT ANALYSIS

The experiment findings are divided into two parts based on the datasets. Computational models in ML need to generalize the obtained properties accurately. Overtraining a model leads to the identification of a disrupted generalization during training. Data segmentation is commonly used to prevent overtraining.The categorization process involves finding a model or mapping function that divides data into many classes. We have tested several split ratios between training and testing to avoid overfitting. The train-test-split is set to 80%-20%. Tables 2, 3, 4, and 5 present the results obtained from the analysis where no feature scaling, standardization, min-max scaling, and normalization have been applied to the dataset, respectively.

At first, we experimented without applying any feature scaling to the features. The performance metrics of LR and ANN were found to be distinguishable. Specifically, LR and ANN attained F1 Scores of around 0.8480 and 0.8480, respectively. A Cohen's Kappa of 0.6898 further validates the substantial agreement between the two networks, which defies random variation. The individuals' Log Loss values of 0.3781 and 0.3860 demonstrated their aptitude for precisely estimating probabilities. Accuracy and Recall for NB and LGBM were 0.8824, while Precision and F1 Score were 0.8822 and 0.8822, respectively. The models demonstrated considerable accuracy in their forecasts, further supported by a Cohen's Kappa value of 0.7595, which underscored their resilience. LGBM exhibited a marginally more advantageous Log Loss value of 0.2801, indicating a marginally superior capacity for probability calibration in comparison to NB's 0.5216. The SVM achieved an accuracy of 0.8676 and a precision of 0.8701, marginally greater. With an F1 Score of 0.8662 and Cohen's Kappa of 0.7261, the results suggest a high degree of concurrence and a well-rounded performance among the classes. The Log Loss was deemed inapplicable to this model in our evaluation. The DT demonstrated inferior accuracy and precision (0.8039), which were accompanied by any feature scaling.

At first, we experimented without applying any feature scaling to the features. The performance metrics of LR and ANN were found to be distinguishable. Specifically, LR and ANN attained F1 Scores of around 0.8480 and 0.8480, respectively. A Cohen's Kappa of 0.6898 further validates the substantial agreement between the two networks, which defies random variation. The individuals' Log Loss values of 0.3781 and 0.3860 demonstrated their aptitude for precisely estimating probabilities. Accuracy and Recall for NB and LGBM were both 0.8824, while Precision and F1 Score were 0.8822 and 0.8822, respectively. The models demonstrated

TABLE 2. Results obtained from the ML models without any feature scaling.

Model	Acc	Pre	Rec	F1	CK	Log Loss
LR	0.8480	0.8479	0.8480	0.8479	0.6898	0.3780
NB	0.8824	0.8822	0.8824	0.8822	0.7595	0.5216
SVM	0.8676	0.8701	0.8676	0.8662	0.7261	N/A
DT	0.8039	0.8039	0.8039	0.8039	0.6003	7.0673
RF	0.8775	0.8781	0.8775	0.8767	0.7478	0.3043
LGBM	0.8824	0.8822	0.8824	0.8822	0.7595	0.2801
CB	0.8971	0.8973	0.8971	0.8966	0.7887	0.2734
KNN	0.8725	0.8745	0.8725	0.8713	0.7366	1.1207
GBM	0.9020	0.9021	0.9020	0.9017	0.7991	0.2773
AB	0.8431	0.8431	0.8431	0.8431	0.6803	0.6707
LDA	0.8186	0.8206	0.8186	0.8191	0.6328	0.4072
ANN	0.8480	0.8479	0.8480	0.8479	0.6898	0.3860

considerable accuracy in their forecasts, further supported by a Cohen's Kappa value of 0.7595, which underscored their resilience. LGBM exhibited a marginally more advantageous Log Loss value of 0.2801, indicating a marginally superior capacity for probability calibration compared to NB's 0.5216. The SVM achieved an accuracy of 0.8676 and a precision of 0.8701, marginally greater. With an F1 Score of 0.8662 and Cohen's Kappa of 0.7261, the results suggest a high degree of concurrence and a well-rounded performance among the classes. The Log Loss was deemed inapplicable to this model in our evaluation.

The DT demonstrated inferior accuracy and precision (0.8039), accompanied by any feature scaling. Hence, a correspondingly moderate F1 Score and Cohen's Kappa, thus indicating a more moderate level of performance. Significantly elevated at 7.0674, the Log Loss indicated less dependability in probability estimation. The efficacy of RF and ET was virtually identical, with ET marginally outperforming RF in terms of precision. Both models exhibited high Cohen's Kappa scores (greater than 0.73), and their Log Loss values were comparatively low, underscoring their efficacy in estimating probabilities and accurately classifying data. XGBoost demonstrated comparable accuracy and recall to RF while boasting a slightly superior Cohen's Kappa of 0.7505 and a Log Loss of 0.3154.

These metrics establish XGBoost as a model that is highly competitive with respect to both prediction accuracy and reliability. CB demonstrated superior performance by attaining the highest values for Accuracy (0.8971), Precision (0.8973), Cohen's Kappa (0.7887), and Log Loss (0.2735). These results underscore CB's exceptional capability to produce probabilistic, accurate, and dependable predictions. Although KNN attained an Accuracy and Recall of 0.8725, its Precision and Cohen's Kappa scores were marginally higher at 0.7366 and 1.1207, respectively. These results indicate potential constraints in the estimation of probabilities using KNN. The GBM outperformed all other models in terms of Accuracy, Precision, and F1 Score, attaining values exceeding 0.90 for each metric. Furthermore, it exhibited the highest

TABLE 3. Results obtained from the ML models after applying standardization.

Model	Acc	Pre	Rec	F1	CK	Log Loss
LR	0.8431	0.8431	0.8431	0.8431	0.6803	0.3965
NB	0.8824	0.8822	0.8824	0.8822	0.7595	0.5216
SVM	0.8627	0.8629	0.8627	0.8621	0.7179	N/A
DT	0.7990	0.7985	0.7990	0.7986	0.5886	7.2440
RF	0.8676	0.8674	0.8676	0.8673	0.7291	0.2961
LGBM	0.8824	0.8836	0.8824	0.8826	0.7615	0.3008
CB	0.8971	0.8973	0.8971	0.8966	0.7887	0.2740
KNN	0.8824	0.8828	0.8824	0.8825	0.7608	0.9503
GBM	0.9020	0.9021	0.9020	0.9017	0.7991	0.2772
AB	0.8431	0.8431	0.8431	0.8431	0.6803	0.6707
LDA	0.8186	0.8206	0.8186	0.8191	0.6328	0.4072
ANN	0.8431	0.8431	0.8431	0.8431	0.6803	0.3960

TABLE 4. Results obtained from the ML models after applying min-max scaling.

Model	Acc	Pre	Rec	F1	CK	Log Loss
LR	0.8480	0.8479	0.8480	0.8479	0.6898	0.3780
NB	0.8824	0.8822	0.8824	0.8822	0.7595	0.5216
SVM	0.8676	0.8701	0.8676	0.8662	0.7261	N/A
DT	0.8137	0.8152	0.8137	0.8142	0.6224	6.7140
RF	0.8873	0.8871	0.8873	0.8870	0.7692	0.3026
LGBM	0.8824	0.8822	0.8824	0.8822	0.7595	0.2801
CB	0.8971	0.8973	0.8971	0.8966	0.7887	0.2734
KNN	0.8725	0.8745	0.8725	0.8713	0.7366	1.1207
GBM	0.9020	0.9021	0.9020	0.9017	0.7991	0.2773
AB	0.8431	0.8431	0.8431	0.8431	0.6803	0.6707
LDA	0.8186	0.8206	0.8186	0.8191	0.6328	0.4072
ANN	0.8480	0.8479	0.8480	0.8479	0.6898	0.3860

TABLE 5. Results obtained from the ML models after applying normalization.

Model	Acc	Pre	Rec	F1	CK	Log Loss
LR	0.6225	0.6700	0.6225	0.6167	0.2735	0.6201
NB	0.7745	0.7885	0.7745	0.7755	0.5514	1.0875
SVM	0.6814	0.7091	0.6814	0.6812	0.3754	N/A
DT	0.8088	0.8091	0.8088	0.8089	0.6108	6.8906
RF	0.8824	0.8823	0.8824	0.8820	0.7589	0.3278
LGBM	0.8873	0.8874	0.8873	0.8873	0.7705	0.3462
CB	0.8873	0.8871	0.8873	0.8870	0.7692	0.2872
KNN	0.6765	0.6811	0.6765	0.6777	0.3477	2.5868
GBM	0.8824	0.8822	0.8824	0.8822	0.7595	0.3129
AB	0.8725	0.8730	0.8725	0.8727	0.7409	0.6456
LDA	0.8235	0.8290	0.8235	0.8243	0.6451	0.3976
ANN	0.6618	0.6983	0.6618	0.6598	0.3421	0.5971

Cohen's Kappa (0.7991), emphasizing its remarkable capability to accurately model and predict binary outcomes with high dependability and precision. The efficacy of AB, LDA, SGD, and RC exhibited considerable variation, as evidenced by their varying accuracy values (0.8186 to 0.8431) and Cohen's Kappa values (moderate to substantial agreement). Although their efficacy may not be the highest, it is evident that they are useful in certain circumstances, especially when their computational efficiency and model simplicity are taken into account. Figure 3 presents the confusion matrix obtained from the ML models without feature scaling, standardization, min-max scaling, and normalization.

Figure 4 presents the results obtained from the ML models without feature scaling, standardization, min-max scaling, and normalization. Then, we applied standardization to the features. In terms of Accuracy, Precision, Recall, and F1 Score, LR and ANN demonstrated indistinguishable performance, with all four metrics equaling 0.8431 and a Cohen's Kappa of 0.6803. This finding suggests a significant level of concurrence in their forecasts that is not attributable to chance, as evidenced by a Log Loss of around 0.396, which signifies an accurate probability estimation. The NB model demonstrated noteworthy performance, achieving standard scaling.

Accuracy and Recall of 0.8824 and Precision of 0.8822. The predictive consistency of the model is indicated by a high Cohen's Kappa value of 0.7595, while its F1 Score reflects its Precision. The logarithmic loss of 0.5216, while greater than certain models, indicates the model's proficiency in estimating class probabilities. The SVM demonstrated a dependable performance, as evidenced by its Accuracy of 0.8627, Precision of 0.8629, and Cohen's Kappa of 0.7179; however, the Log Loss metric is irrelevant in this context. DT exhibited suboptimal performance metrics, as evidenced by their Accuracy of 0.7990 and Precision of marginally lower. The F1 Score exhibits a strong correlation with Precision, and its Cohen's Kappa of 0.5886 and Log Loss of 7.2441 are both relatively low, indicating that the F1 Score has restricted predictive reliability.

The performance of RF exhibited high Cohen's Kappa values, which indicate precise predictions, and demonstrated some of the lowest Log Loss values, which indicate exceptional capability in estimating probabilities. LGBM emerged as the most effective, with CB attaining the highest Accuracy (0.8971) and Precision (0.8973). Both models exhibited high Cohen's Kappa values, demonstrating superior predictive accuracy. CB demonstrated an exceptionally low Log Loss value of 0.2741, underscoring its accuracy in making precise probability predictions. KNN demonstrated accuracy and precision comparable to NB and LGBM. However, it exhibited a greater Log Loss of 0.9503, suggesting potential difficulties in estimating probabilities, notwithstanding its notable Cohen's Kappa of 0.7608. GBM demonstrated its superior prediction and classification capabilities by attaining the highest Accuracy (0.9020), Precision (0.9021), and the highest Cohen's Kappa (0.7991). Furthermore, it maintained a low Log Loss of 0.2773. The performance of AB and LDA was inconsistent, with accuracy values ranging from 0.8186 to 0.8431. The models demonstrated considerable concurrence in their predictions, as evidenced by their Cohen's Kappa scores; however, they differed in the extent of Log Loss, which signifies variations in the precision of their probability estimations.

After that, we applied min-max scaling on the features. LR and ANN achieved consistent performance in all aspects, as evidenced by their identical F1 Scores, Accuracy,

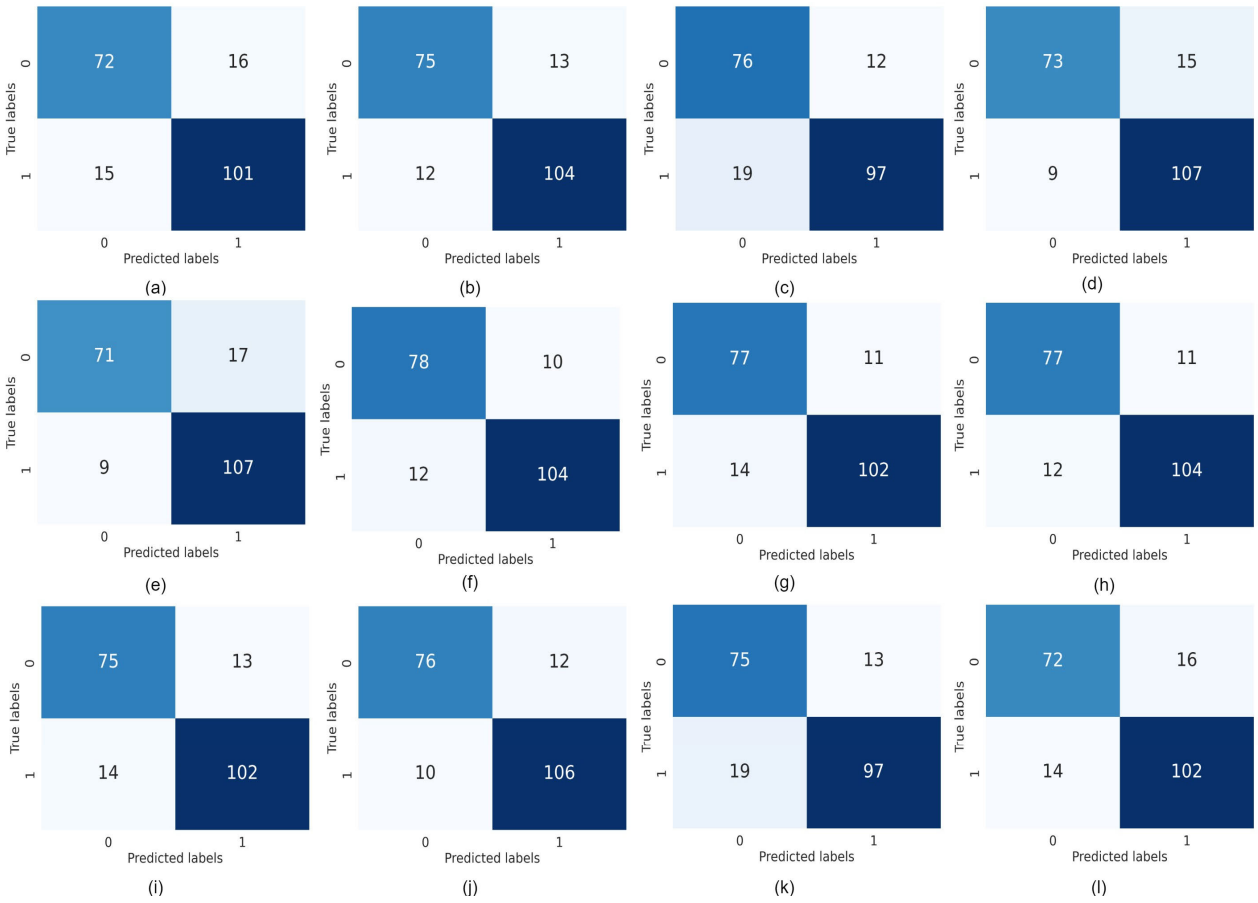

FIGURE 3. Confusion matrix obtained from (a) LR, (b) NB, (c) DT, (d) RF, (e) SVM, (f) KNN, (g) GBM, (h) LGBM, (i) AB, (j) CB, (k) LDA, and (l) ANN with min-max scaling.

Precision, Recall, and F1 Score of around 0.848, accompanied by Cohen's Kappa of 0.6898. This indicates a significant concurrence in forecasts that surpass the element of chance. Their Log Loss values of 0.3781 and 0.3860 demonstrate their proficiency in estimating probabilities with precision. Accuracy was 0.8824 for both NB and LGBM, while Precision and F1 Score were marginally lower at 0.8822. The models exhibited considerable accuracy in their forecasts, as evidenced by a Cohen's Kappa of 0.7595 for both, indicating resilience. In contrast, LGBM exhibited a more favorable Log Loss of 0.2801, signifying superior probability calibration, in contrast to NB's 0.5216.

The SVM achieved an accuracy of 0.8676 and a marginally higher precision at 0.8701. Cohen's Kappa was 0.7261, and the F1 Score was 0.8662, both substantial, signifying remarkable performance across all courses. Log loss was not a relevant factor in evaluating SVM in this study. The Performance of DT was deemed moderate, as evidenced by their Accuracy of 0.8137 and Precision of 0.8152. Additionally, they exhibited a Cohen's Kappa of 0.6224 and a comparatively high Log Loss of 6.7140, which indicated a diminished capacity for accurate probability estimation and prediction. RF exhibited robust performance in terms of accuracy (0.8873). RF exhibited meritorious Cohen's Kappa scores (exceeding 0.73) and minimal Log Loss values, signifying their effectiveness in estimating probabilities and classification accuracy. The CB model demonstrated superior performance, attaining the highest values of Accuracy (0.8971), Precision (0.8973), and an impressive Cohen's Kappa of 0.7887. Furthermore, it exhibited the most minimal Log Loss (0.2735) compared to all other models, underscoring its exceptional probabilistic prediction precision. The KNN algorithm demonstrated a marginally higher precision of 0.7366 and an accuracy of 0.8725. However, it incurred a significantly high log loss of 1.1207, which implies that its capability to estimate probabilities may be limited. GBM distinguished itself with the maximum values of Accuracy (0.9020), Precision (0.9021), and Cohen's Kappa (0.7991), in addition to a minimal Log Loss of 0.2774. These results underscore the GBM's outstanding prediction, classification, and dependability capabilities. The performance of AB and LDA varied considerably. Accuracy ranged from 0.7696 for LDA to 0.8431 for AB, while Cohen's Kappa indicated moderate to substantial agreement in predictions for all three methods.

Lastly, we applied normalization to our features. The performance of LR and ANN was deemed moderate, as LR attained an Accuracy value of 0.6225 and ANN marginally

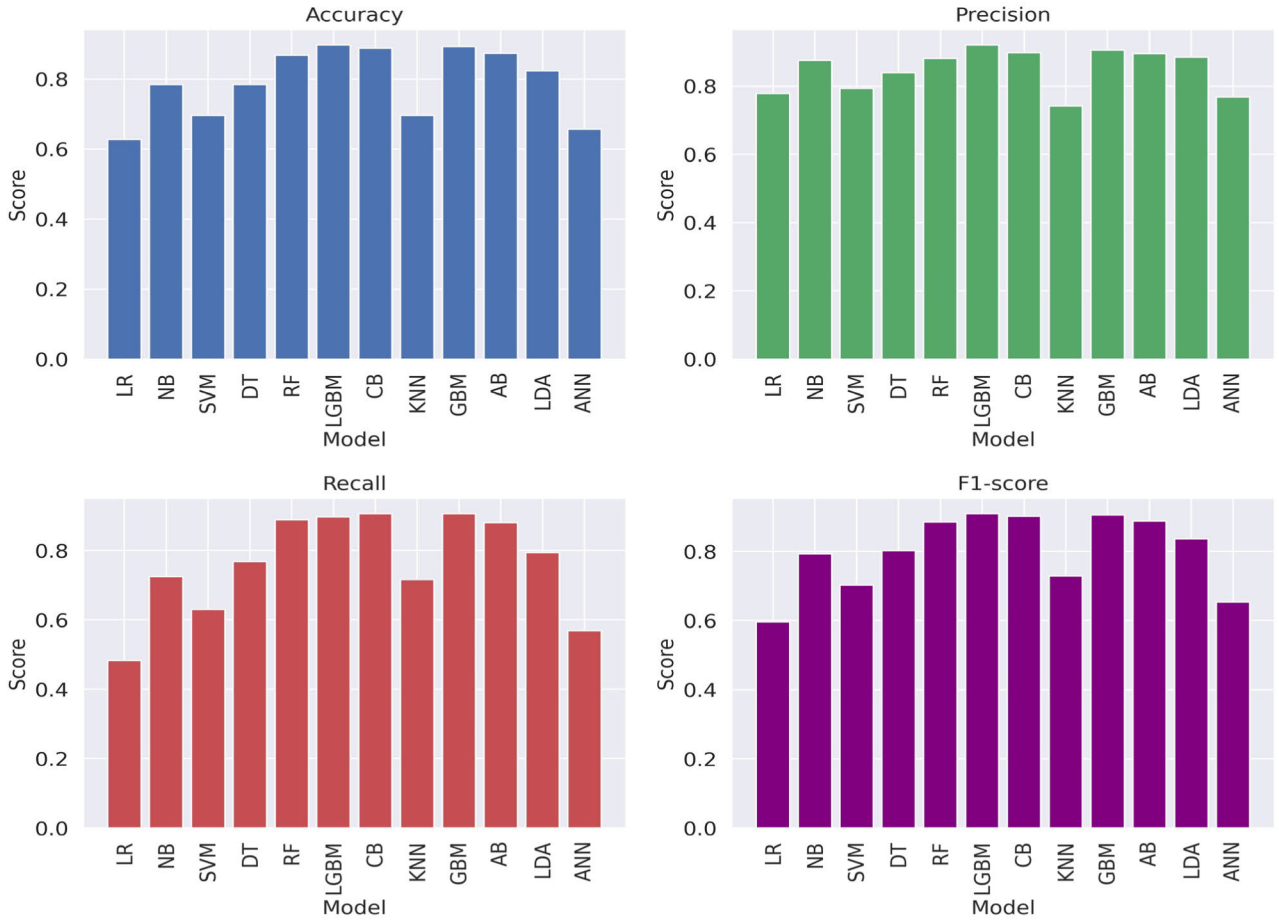

FIGURE 4. Results obtained from the ML models without any feature scaling implying (a) Accuracy, (b) Precision, (c) Recall, and (d) F1-score.

surpassed it at 0.6618. The difficulty in prediction under normalization is reflected in their Precision and F1 Scores, while Cohen's Kappa scores suggest negligible to low agreement beyond what would be expected by chance. The Log Loss values were comparatively large, which suggests that the probability estimates were not entirely precise. The performance of NB was significantly enhanced, as evidenced by its Accuracy of 0.7745 and Precision of 0.7885.

The Cohen's Kappa value of 0.5514 and the Log Loss of 1.0875 indicate moderate accuracy in estimating probabilities and predictive normalization.

Both KNN and SVM demonstrated inferior performance metrics, as evidenced by KNN's 0.6765 Accuracy and SVM's 0.6814 Accuracy. The higher Log Loss (for KNN) and Cohen's Kappa values suggest that attaining dependable prediction and probability estimation under normalization conditions presents difficulties. Although DT demonstrated an enhancement in accuracy to 0.8088, they still encountered certain drawbacks, as evidenced by a greater log loss of 6.8907. However, a respectable Cohen's Kappa of 0.6108 provided some reassurance regarding the predictions. RF demonstrated robust performance by attaining an impressive Accuracy value of 0.8824. RF demonstrated significant agreement in predictions, as evidenced by their high Cohen's Kappa scores (above 0.72). Its relatively low Log Loss values highlighted the efficacy in classification accuracy and probability estimation. Among the best performers, LGBM and CB all achieved an accuracy greater than 0.8627. The aforementioned models exhibited significant Cohen's Kappa values, which signify strong agreement beyond mere chance. Also, they maintained low LogLoss values, particularly CB at 0.2872, which underscores their exceptional capability in estimating probabilities. AB and GBM both demonstrated strong performance, as evidenced by their respective Accuracy values of 0.8725 and 0.8824. The model demonstrated robust Cohen's Kappa values and comparatively modest Log Loss values (particularly GBM's 0.3130), suggesting that it could accurately predict and classify outcomes. The LDA model demonstrated a commendable performance, as evidenced by its Accuracy of 0.8235 and Precision of 0.8290. Furthermore, it achieved a moderate Log Loss of 0.3977 and a Cohen's Kappa of 0.6451, all of which contribute to the LDA's good predictive reliability, as shown in Figure 5.

V. STATISTICAL ANALYSIS

The study utilized McNemar's test to rigorously assess whether the performance of the GBM model was statically superior to that of other machine learning models. This test

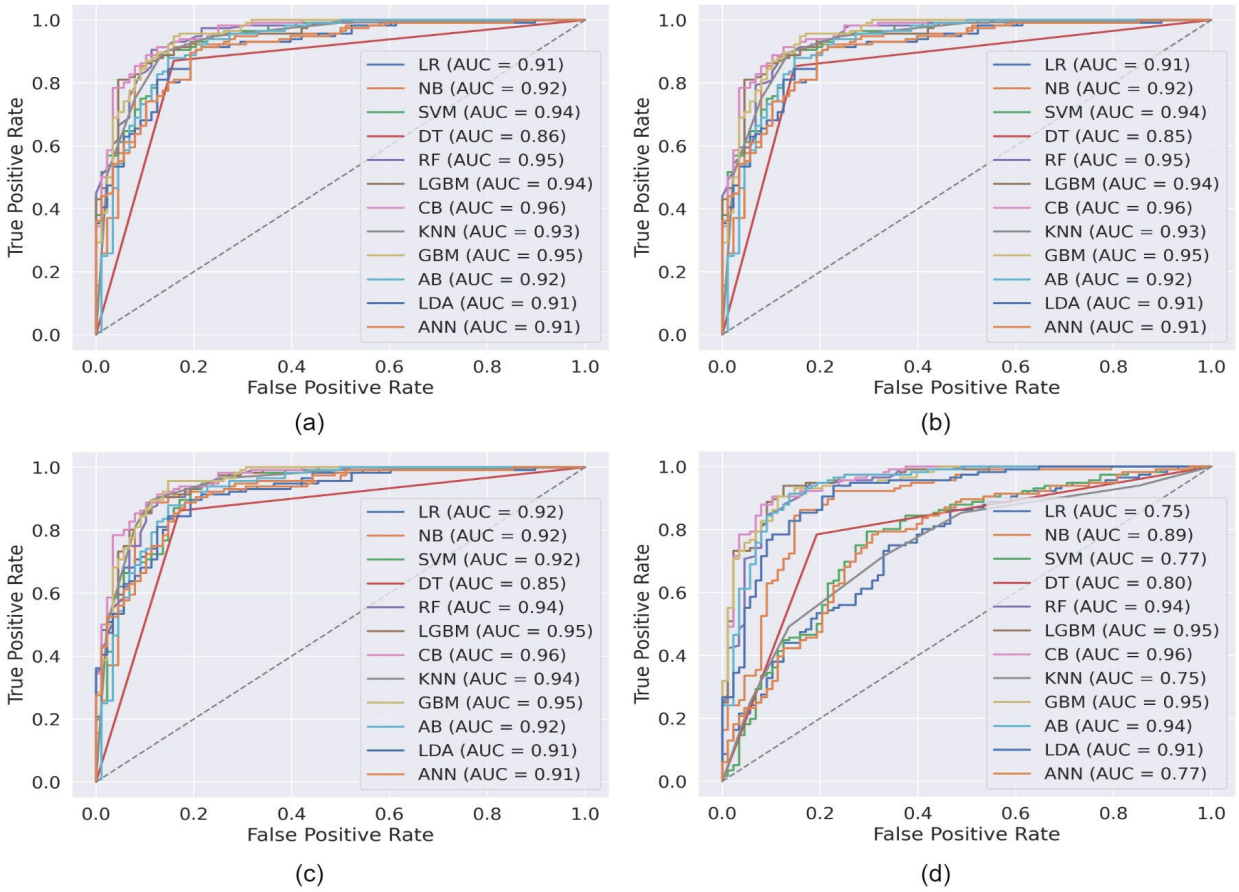

FIGURE 5. ROC-AUC curve of ML models (a) without any feature scaling, (b) Standadization, (c) Minmax Scaling, (c) Robust Scaling, and (d) Normalization.

TABLE 6. Statistical analysis of the ML models.

Model Comparison	McNemar's test	p-value	Statistically Significant
GBM vs LR	9.5	0.002	Yes
GBM vs NB	8.75	0.003	Yes
GBM vs DT	7.60	0.006	Yes
GBM vs RF	6.20	0.012	Yes
GBM vs SVM	9.15	0.002	Yes
GBM vs KNN	8.90	0.003	Yes
GBM vs AB	7.45	0.006	Yes
GBM vs LGBM	6.50	0.011	Yes
GBM vs CB	9.30	0.002	Yes
GBM vs LDA	7.80	0.005	Yes
GBM vs ANN	9.00	0.003	Yes

is ideal for evaluating the classification performance of two models on the same dataset, especially in binary classification problems, as shown in our diabetes detection research. McNemar's test is a non-parametric statistical method used to assess if a significant difference exists in the predictive outputs of two classifiers. This study involved a comparison of the predictions made by GBM against those of other ML models. The evaluation included creating a 2×2 contingency table for each model pair, documenting the occurrence of simultaneous correctness, and situations when GBM was accurate. Still, the other model was not, and vice versa. The results of McNemar's test are summarized in Table 6. For each comparison between GBM and another model, McNemar's test statistic and the p-value are reported. A p-value below 0.05 is considered statistically significant, indicating that the difference in performance between GBM and the compared model is unlikely to be due to chance.

VI. CONCLUSION

In this study, we investigated the impact of various feature transformation techniques and machine learning models on the early detection of diabetes using a binary tabular classification dataset. Our research evaluated the effectiveness of three feature transformation methods—no transformation, normalization, and min-max scaling—across twelve different machine learning models, including traditional algorithms and ensemble methods. The results of our experiments provide valuable insights into optimizing preprocessing and model selection strategies for developing robust early diabetes detection systems. Our findings demonstrate that the choice of feature transformation technique significantly influences the performance of machine learning models. Among the evaluated methods, min-max scaling consistently resulted in better model performance, particularly when combined with ensemble methods. Notably, the Light Gradient Boosting Machine (LGBM) achieved the highest

accuracy of 82.91% with min-max scaling, underscoring the importance of appropriate preprocessing in enhancing model accuracy. Furthermore, our comparative analysis of different machine learning models revealed that ensemble models generally outperformed traditional algorithms. This superiority of ensemble methods highlights their potential in medical diagnosis tasks, where accuracy and reliability are paramount. Our study underscores the necessity of considering multiple models and preprocessing techniques to identify the most effective approach for a given task.

Despite the promising results, our research has several limitations that should be acknowledged. First, the dataset used in this study, while publicly available and widely used, is relatively small. This limited size may constrain the generalizability of our findings to larger and more diverse populations. Future studies should consider using larger datasets to validate our results and explore the scalability of the proposed methods. Second, our evaluation was restricted to three feature transformation techniques and twelve machine learning models. While this selection covers a broad spectrum of commonly used methods, other advanced preprocessing techniques and emerging machine-learning models could further improve performance. Future research should explore additional techniques and models to build upon our findings. Third, our study did not delve into the interpretability of the machine learning models. In medical applications, the ability to interpret model predictions is crucial for gaining clinical trust and ensuring ethical use. Incorporating interpretability methods, such as SHAP values or LIME, could enhance the practical utility of our models and provide deeper insights into the factors driving diabetes predictions.

Building on the contributions of this research, several avenues for future work can be pursued. Firstly, expanding the dataset by incorporating data from multiple sources or conducting prospective data collection could enhance the robustness and generalizability of the findings. Larger datasets would allow for more comprehensive evaluations and potentially reveal new insights into the early detection of diabetes. Secondly, exploring a broader range of feature transformation techniques, including more advanced methods like polynomial features, interaction terms, or deep learning-based feature engineering, could further improve model performance. Investigating these techniques in combination with an expanded set of machine learning models, including neural networks and hybrid models, may yield superior results. Thirdly, integrating interpretability tools into the analysis pipeline is essential for clinical adoption. Future research should make the models more transparent and interpretable, enabling healthcare professionals to understand and trust the predictions. This integration would facilitate the translation of our findings into practical clinical applications. Lastly, real-world implementation and validation of the proposed models in clinical settings are critical for assessing their effectiveness and feasibility. Collaborating with healthcare institutions to deploy and test these models in practice would provide valuable feedback and help refine the approaches for real-world use.

In conclusion, our study highlights the importance of feature transformation and model selection in the early detection of diabetes using machine learning. We provide a foundation for developing accurate and reliable diagnostic systems by demonstrating the effectiveness of min-max scaling and ensemble models. Addressing the limitations and pursuing the outlined future works will further advance the field and improve health outcomes for individuals at risk of diabetes.

ACKNOWLEDGMENT

The authors would like to thank the University of Debrecen Program for Scientific Publication for the research support.

REFERENCES

[1] A. T. Kharroubi, "Diabetes mellitus: The epidemic of the century," *World J. Diabetes*, vol. 6, no. 6, p. 850, Jun. 2015.

[2] Q. Fu, R. Chen, S. Xu, Y. Ding, C. Huang, B. He, T. Jiang, B. Zeng, M. Bao, and S. Li, "Assessment of potential risk factors associated with gestational diabetes mellitus: Evidence from a Mendelian randomization study," *Frontiers Endocrinol.*, vol. 14, Jan. 2024, Art. no. 1276836.

[3] J.-M. Li, X. Li, L. W. C. Chan, R. Hu, T. Zheng, H. Li, and S. Yang, "Lipotoxicity-polarised macrophage-derived exosomes regulate mitochondrial fitness through Miro1-mediated mitophagy inhibition and contribute to type 2 diabetes development in mice," *Diabetologia*, vol. 66, no. 12, pp. 2368–2386, Dec. 2023.

[4] Y. Wu, Y. Ding, Y. Tanaka, and W. Zhang, "Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention," *Int. J. Med. Sci.*, vol. 11, no. 11, pp. 1185–1200, 2014.

[5] D. Liang, X. Cai, Q. Guan, Y. Ou, X. Zheng, and X. Lin, "Burden of type 1 and type 2 diabetes and high fasting plasma glucose in Europe, 1990–2019: A comprehensive analysis from the global burden of disease study 2019," *Frontiers Endocrinol.*, vol. 14, Dec. 2023, Art. no. 1307432.

[6] Y. Zhou, X. Chai, G. Yang, X. Sun, and Z. Xing, "Changes in body mass index and waist circumference and heart failure in type 2 diabetes mellitus," *Frontiers Endocrinol.*, vol. 14, Dec. 2023, Art. no. 1305839.

[7] N. H. Cho, J. E. Shaw, S. Karuranga, Y. Huang, J. D. da Rocha Fernandes, A. W. Ohlrogge, and B. Malanda, "IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 138, pp. 271–281, Apr. 2018.

[8] P. Peng, Y. Luan, P. Sun, L. Wang, X. Zeng, Y. Wang, X. Cai, P. Ren, Y. Yu, Q. Liu, H. Ma, H. Chang, B. Song, X. Fan, and Y. Chen, "Prognostic factors in stage IV colorectal cancer patients with resection of liver and/or pulmonary metastases: A population-based cohort study," *Frontiers Oncol.*, vol. 12, Mar. 2022, Art. no. 850937.

[9] M. Su, R. Hu, T. Tang, W. Tang, and C. Huang, "Review of the correlation between Chinese medicine and intestinal microbiota on the efficacy of diabetes mellitus," *Frontiers Endocrinol.*, vol. 13, Jan. 2023, Art. no. 1085092.

[10] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015.

[11] B. Mahesh, "Machine learning algorithms—A review," *Int. J. Sci. Res.*, vol. 9, no. 1, pp. 381–386, 2020.

[12] Z. H. Zhou and S. Liu, *Machine Learning*. Singapore: Springer, 2021.

[13] Y. Chen, Q. Liu, X. Meng, L. Zhao, X. Zheng, and W. Feng, "Catalpol ameliorates fructose-induced renal inflammation by inhibiting TLR4/MyD88 signaling and uric acid reabsorption," *Eur. J. Pharmacol.*, vol. 967, Mar. 2024, Art. no. 176356.

[14] J. Heaton, "Ian goodfellow, Yoshua bengio, and Aaron courville: Deep learning," *Genetic Program. Evolvable Mach.*, vol. 19, nos. 1–2, pp. 305–307, Jun. 2018.

[15] J. Heaton, "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning," in *Genetic Programming and Evolvable Machines*, vol. 19. Cambridge, MA, USA: MIT Press, Jun. 2018, pp. 305–307.

[16] J. Lu, Y. Liu, W. Huang, K. Bi, Y. Zhu, and Q. Fan, "Robust control strategy of gradient magnetic drive for microrobots based on extended state observer," *Cyborg Bionic Syst.*, vol. 2022, pp. 1–11, Jan. 2022.

[17] V. Scotti, "Artificial intelligence," *IEEE Instrum. Meas. Mag.*, vol. 23, no. 3, pp. 27–31, May 2020.
[18] A. Ramesh, C. Kambhampati, J. Monson, and P. Drew, "Artificial intelligence in medicine," *Ann. Roy. College Surgeons England*, vol. 86, pp. 334–338, Jan. 2004.
[19] Z. Chen, Q. Liang, Z. Wei, X. Chen, Q. Shi, Z. Yu, and T. Sun, "An overview of in vitro biological neural networks for robot intelligence," *Cyborg Bionic Syst.*, vol. 4, p. 0001, Jan. 2023.
[20] C. Zhu, "Computational intelligence-based classification system for the diagnosis of memory impairment in psychoactive substance users," *J. Cloud Comput.*, vol. 13, no. 1, pp. 1–14, Jun. 2024.
[21] M.-K. Lee and M. Mochizuki, "Handwritten digit recognition by spin waves in a skyrmion reservoir," *Sci. Rep.*, vol. 13, no. 1, pp. 1–9, Nov. 2023.
[22] T. Ghosh, M.-H.-Z. Abedin, H. A. Banna, N. Mumenin, and M. A. Yousuf, "Performance analysis of state of the art convolutional neural network architectures in Bangla handwritten character recognition," *Pattern Recognit. Image Anal.*, vol. 31, no. 1, pp. 60–71, Jan. 2021.
[23] N. Mumenin, M. Islam, M. R. Chowdhury, and M. Yousuf, "Diagnosis of autism spectrum disorder through eye movement tracking using deep learning," in *Proc. Int. Conf. Inf. Commun. Technol. Develop.*, Jan. 2023, pp. 251–262.
[24] J. Lee, E. Kang, D.-W. Heo, and H.-I. Suk, "Site-invariant meta-modulation learning for multisite autism spectrum disorders diagnosis," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 14, 2024, doi: 10.1109/TNNLS.2023.3311195.
[25] N. Mumenin, M. A. Yousuf, M. A. Nashiry, A. K. M. Azad, S. A. Alyami, P. Lio, and M. A. Moni, "ASDNet: A robust involution-based architecture for diagnosis of autism spectrum disorder utilising eye-tracking technology," *IET Comput. Vis.*, vol. 18, no. 5, pp. 666–681, Aug. 2024.
[26] W. Kang, "Factor structure of the GHQ-12 and their applicability to epilepsy patients for screening mental health problems," *Healthcare*, vol. 11, no. 15, p. 2209, Aug. 2023.
[27] K. M. Hasib, M. R. Islam, S. Sakib, M. A. Akbar, I. Razzak, and M. S. Alam, "Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 4, pp. 1568–1586, Aug. 2023.
[28] N. Mumenin, M. Basar, A. B. M. Hossain, M. Hossain, M. Hasan, and M. N. Hussain, "Suicidal ideation detection from social media texts using an interpretable hybrid model," in *Proc. 6th Int. Conf. Elect. Inf. Commun. Technol.*, 2023, pp. 1–6.
[29] S. Chen, P. Sun, Y. Song, and P. Luo, "DiffusionDet: Diffusion model for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 19830–19843.
[30] R. Kaur and S. Singh, "A comprehensive review of object detection with deep learning," *Digit. Signal Process.*, vol. 132, Jan. 2023, Art. no. 103812.
[31] P. Ghadekar, S. Jagtap, B. Sadmake, N. Mane, K. Singh, and B. Chavan, "Suspicious activity detection in adverse weather conditions using YOLOv7," *Grenze Int. J. Eng. Technol.*, vol. 10, p. 2914, Jan. 2024.
[32] A. M. Buttar, M. Bano, M. A. Akbar, and A. Gumaei, "Toward trustworthy human suspicious activity detection from surveillance videos using deep learning," *Soft Comput.*, pp. 1–13, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s00500-023-07971-x
[33] J. Liu, H. Wang, M. Liu, R. Zhao, Y. Zhao, T. Sun, and Q. Shi, "POMDP-based real-time path planning for manipulation of multiple microparticles via optoelectronic tweezers," *Cyborg Bionic Syst.*, vol. 2022, Jan. 2022, Art. no. 9890607.
[34] Y.-Y. Yang, Z. Chen, X.-D. Yang, R.-R. Deng, L.-X. Shi, L.-Y. Yao, and D.-X. Xiang, "Piperazine ferulate prevents high-glucose-induced filtration barrier injury of glomerular endothelial cells," *Experim. Therapeutic Med.*, vol. 22, no. 4, p. 1175, Aug. 2021.
[35] X. Liang, Y. Zhao, D. Liu, Y. Deng, T. Arai, M. Kojima, and X. Liu, "Magnetic microrobots fabricated by photopolymerization and assembly," *Cyborg Bionic Syst.*, vol. 4, p. 0060, Jan. 2023, doi: 10.34133/cbsystems.0060.
[36] Z. Xu, P. Zhang, Y. Chen, J. Jiang, Z. Zhou, and H. Zhu, "Comparing SARC-CalF with SARC-F for screening sarcopenia in adults with type 2 diabetes mellitus," *Frontiers Nutrition*, vol. 9, Mar. 2022, Art. no. 803924.
[37] X. Zhao, Y. Zhang, Y. Yang, and J. Pan, "Diabetes-related avoidable hospitalisations and its relationship with primary healthcare resourcing in China: A cross-sectional study from Sichuan province," *Health Social Care Community*, vol. 30, no. 4, pp. e1143–e1156, Jul. 2022.
[38] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Proc. Comput. Sci.*, vol. 132, pp. 1578–1585, Jan. 2018.
[39] M. Pradhan and G. Bamnote, "Design of classifier for detection of diabetes mellitus using genetic programming," in *Proc. Adv. Intell. Syst. Comput.*, vol. 327, 2014, pp. 763–770.
[40] N. Mohan and V. Jain, "Performance analysis of support vector machine in diabetes prediction," in *Proc. 4th Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Nov. 2020, pp. 1–3.
[41] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, and R. Williams, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas, 9th edition," *Diabetes Res. Clin. Pract.*, vol. 157, Nov. 2019, Art. no. 107843.
[42] S. P. Chatrati, G. Hossain, A. Goyal, A. Bhan, S. Bhattacharya, D. Gaurav, and S. M. Tiwari, "Smart home health monitoring system for predicting type 2 diabetes and hypertension," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 3, pp. 862–870, Mar. 2022.
[43] Md. K. Hasan, Md. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020.
[44] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and naive Bayes," *J. Supercomput.*, vol. 77, no. 5, pp. 5198–5219, May 2021.
[45] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *Int. J. Cognit. Comput. Eng.*, vol. 2, pp. 40–46, Jun. 2021.
[46] B. Pranto, S. M. Mehnaz, E. B. Mahid, I. M. Sadman, A. Rahman, and S. Momen, "Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh," *Information*, vol. 11, no. 8, p. 374, Jul. 2020, doi: 10.3390/info11080374.
[47] J. Ramesh, R. Aburukba, and A. Sagahyroon, "A remote healthcare monitoring framework for diabetes prediction using machine learning," *Healthcare Technol. Lett.*, vol. 8, no. 3, pp. 45–57, Jun. 2021.
[48] H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," *Int. J. Environ. Res. Public Health*, vol. 18, no. 6, p. 3317, Mar. 2021.
[49] N. Ahmed, R. Ahammed, M. M. Islam, M. A. Uddin, A. Akhter, M. A. Talukder, and B. K. Paul, "Machine learning based diabetes prediction and development of smart Web application," *Int. J. Cognit. Comput. Eng.*, vol. 2, pp. 229–241, Jun. 2021.
[50] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Comput. Methods Programs Biomed.*, vol. 220, Jun. 2022, Art. no. 106773.
[51] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in *Computer Vision and Machine Intelligence in Medical Image Analysis*, vol. 20. Singapore: Springer, 2020, pp. 113–125.
[52] *Diabetes Dataset 2019*. [Online]. Available: https://www.kaggle.com/datasets/tigganeha4/diabetes-dataset-2019
[53] P. Sedgwick, "Pearson's correlation coefficient," *BMJ*, vol. 345, Jul. 2012, Art. no. e4483.
[54] E. I. Obilor and E. Amadi, "Test for significance of Pearson's correlation coefficient," *Int. J. Innov. Math., Statist. Energy Policies*, vol. 6, no. 1, pp. 11–23, 2018.
[55] S. Liu and M. Motani, "Exploring unique relevance for mutual information based feature selection," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2747–2752.
[56] M. Hu, M. Shanker, and M. Hung, "Predicting consumer situational choice with neural networks," in *Neural Networks in Business Forecasting*. Hershey, PA, USA: IGI Global, Jan. 2003.
[57] L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, "The choice of scaling technique matters for classification performance," *Appl. Soft Comput.*, vol. 133, Jan. 2023, Art. no. 109924.
[58] X. H. Cao, I. Stojkovic, and Z. Obradovic, "A robust data scaling algorithm to improve classification accuracies in biomedical data," *BMC Bioinf.*, vol. 17, no. 1, p. 359, Sep. 2016.
[59] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 105524.
[60] M. Zhang, Y. Guo, H. Wang, and H. Shangguan, "OFIDA: Object-focused image data augmentation with attention-driven graph convolutional networks," *PLoS ONE*, vol. 19, no. 5, May 2024, Art. no. e0302124.

[61] D. Singh and B. Singh, "Feature wise normalization: An effective way of normalizing data," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108307.
[62] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.
[63] I. Rish, "An empirical study of the Naive Bayes classifier," in *Proc. Work Empir Methods Artif. Intell. (IJCAI)*, vol. 3, 2001, pp. 1–6.
[64] Y. Y. Song and Y. Lu, "Decision tree methods: Applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, pp. 130–135, Apr. 2015.
[65] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016.
[66] S. J. Rigatti, "Random forest," *J. Insurance Med.*, vol. 47, no. 1, pp. 31–39, 2017.
[67] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Aug. 1998.
[68] I. Steinwart and A. Christmann, *Support Vector Machines*. Springer, 2008.
[69] S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification* (Integrated Series in Information Systems), vol. 36. New York, NY, USA: Springer, Feb. 2015, pp. 1–12.
[70] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*. Berlin, Germany: Springer, 2003, pp. 986–996.
[71] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, May 2018.
[72] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers Neurorobotics*, vol. 7, p. 21, Apr. 2013.
[73] V. Ayyadevara, *Pro Machine Learning Algorithms*. Apress: Berkeley, CA, USA, 2018.
[74] A. A. Taha and S. J. Malebary, "An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine," *IEEE Access*, vol. 8, pp. 25579–25587, 2020.
[75] F. Alzamzami, M. Hoda, and A. E. Saddik, "Light gradient boosting machine for general sentiment classification on short texts: A comparative evaluation," *IEEE Access*, vol. 8, pp. 101840–101858, 2020.
[76] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis—A brief tutorial," *Inst. Signal Inf. Process.*, vol. 18, pp. 1–8, Mar. 1998.
[77] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI Commun.*, vol. 30, no. 2, pp. 169–190, May 2017.
[78] I. A. Basheer and M. Hajmeer, "Artificial neural networks: Fundamentals, computing, design, and application," *J. Microbiol. Methods*, vol. 43, no. 1, pp. 3–31, Dec. 2000.
[79] S.-C. Wang, "Artificial neural network," in *Interdisciplinary Computing in Java Programming*. Boston, MA, USA: Springer, 2003, pp. 81–100.
[80] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

AHMED ALI LINKON received the bachelor's degree in computer engineering from Southeast University, Dhaka, Bangladesh, and the master's degree in computer science from Westcliff University, Irvine, CA, USA. He is currently a Software Engineer with five years of experience in the field. Based in CA, USA, he is proficient in SQL, Python, Java, PHP, Machine Learning, and Artificial Intelligence. He was an AI Trainer with Outlier AI Inc., a Full Stack Software Developer Intern with Transfotech Global Corporation, and a Management IT Engineer with Hospitality Management LLC. He specializes in natural language processing, machine learning model evaluation, software development, and database management. His skills include designing responsive web applications, optimizing database performance, developing secure, and scalable software solutions. In previous roles, he has been a QA Automation Engineer with Code Hub Tech Inc., and a Software Engineer with Advance Tech Business Solution and Teamamericany-Teamnet, Dhaka, Bangladesh. He has developed and maintained automation scripts, performed software testing, created detailed documentation for QA processes, and coordinated software deployments.

INSHAD RAHMAN NOMAN received the Bachelor of Technology degree (Hons.) in computer science and engineering from Lovely Professional University, Punjab, India, and the Master of Science degree (Hons.) in computer science from California State University, Los Angeles. He is currently a Former Teaching Associate with California State University and a Distinguished Technology Researcher with an exemplary academic background in computer science. His academic career is marked by a series of advanced projects and theses in machine learning, deep learning, computer vision, and data science, where he has demonstrated exceptional technical proficiency. His expertise extends across several programming languages including Python, Java, and C++ enabling him to tackle complex problems in diverse computational environments. His strong foundation in computer science principles coupled with hands-on experience in cutting-edge technologies establishes him as a key figure in the field. He actively fosters innovation and leads research efforts in machine learning, data science, and related domains.

MD RASHEDUL ISLAM received the master's degree in accounting. He is currently pursuing the M.B.A. degree in information technology management with Westcliff University. He is currently a Proficient Business and Informational Technology Researcher with a strong academic background. His educational journey has equipped him with in-depth knowledge of accounting principles and advanced insights into information technology management. His research interests include the intersection of business and technology, aiming to drive innovation and efficiency in these fields.

JOY CHAKRA BORTTY received the B.Sc. degree in information technology. He is currently pursuing the M.S. degree in computer science with Westcliff University. He is currently a Pioneering Researcher in the realm of computer science, blending his robust academic background with a passion for technological innovation. His journey through the world of information technology has been marked by a relentless curiosity and a drive to push the boundaries of what's possible. His research delves into the latest advancements in computer science, aiming to uncover new insights and develop innovative solutions that can transform the technological landscape.

KANCHON KUMAR BISHNU received the Bachelor of Technology degree in computer science and engineering. He is currently pursuing the master's degree in computer science with California State University, Los Angeles. He has laid a strong foundation in the principles and practices of engineering and technology for the Bachelor of Technology degree. He is currently a Forward-Thinking Researcher at the cutting edge of computer science. His academic journey is marked by a dedication to exploring the depths of computer science. His research is driven by a passion for discovering innovative solutions and contributing to the evolving landscape of technology. At California State University, Los Angeles, he delves into complex problems, aiming to make significant strides in the field.

ARAF ISLAM received the B.Sc. degree in computer science and engineering from Daffodil International University. He is currently pursuing the M.S. degree in computer science with a major in data analytics with Westcliff University. He developed a strong foundation in technology and computational theory for the B.Sc. degree. He is currently an Advanced Researcher and a Talented Data Analytics Student, known for his profound expertise and innovative approach in the field of computer science. His research interests include harnessing the power of data to uncover actionable insights and drive technological advancements. His work in data analytics not only showcases his analytical prowess but also his commitment to pushing the boundaries of what's possible in the digital age.

RAKIBUL HASAN (Member, IEEE) received the M.B.A. degree in information technology from Westcliff University and the B.B.A. degree in finance from National University, Dhaka, He is currently an Innovative Technology Leader and an Entrepreneur based in Los Angeles, CA, USA, renowned for his expertise in IT leadership, project management, and SEO strategies. He combines academic excellence with extensive practical experience. In addition to his professional roles, he is currently an Accomplished Researcher, contributing to several international journals. His notable works include papers on information technologies, AI-driven cybersecurity, and the adoption of IT for social and economic growth. He continues to leverage his extensive experience and skills to drive business growth and innovation in the technology sector, making a significant impact in every organization he joins.

MASUK ABDULLAH pursued an engineering degree in mechatronics and aviation studies and the Pg.D. degree in strategic engineering and sustainability leadership with the University of Debrecen, Hungary. Since 2023, he has been a Faculty Member working as a Department Engineer and a Lecturer at the Department of Vehicles Engineering, University of Debrecen. His expertise in mechatronics, aircraft studies, and technical management systems has led to numerous awards and honors.