International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

Comparison of Classification of Different Machine learning Algorithms in the Diagnosis and Detect of Diabetes

Zainab N. Nemer[1], Sabreen Fawzi Raheem[2], and Maytham Alabbas[3]

[1] *College of Computer Science and Information Technology, University of Basrah, Iraq.*
[2] *Basra Technical Institute, Southern Technical University, Basrah, Iraq*
[3] *College of Computer Science and Information Technology, University of Basrah, Iraq.*

E-mail address: zainab.nemer@uobasrah.edu.iq, sabreen.fawzi@stu.edu.iq, ma@uobasrah.edu.iq

Received ## Mon. 20##, Revised ## Mon. 20##, Accepted ## Mon. 20##, Published ## Mon. 20##

Abstract: Diabetes, caused by a rise in blood glucose levels, can be detected using a variety of instruments that analyze blood samples. Heart attacks and kidney failure are among the serious complications that can arise from untreated diabetes. Consequently, the field of detecting and evaluating gestational diabetes requires more robust research and better learning models. The information system for detecting diabetes in this study is based on machine learning (ML) algorithms. In the study, various machine learning techniques are discussed, including Decision Trees (DT), Random Forest (RF), Logistic Regression (LR), Extreme Gradient Boosting (XGBoost), and K-Nearest Neighbors (K-NN). The data was collected from the Iraqi society, mainly from the laboratory of Medical City Hospital and the Specialises Centre for Endocrinology and Diabetes-Al-Kindy Teaching Hospital. On the basis of the Recursive Feature Elimination approach, research has been done to enhance the prediction index. The performances of all five algorithms are evaluated on the various measures like the Precision, Accuracy, F-Measure, Recall, Cohen Kappa, and AUC. Accuracy is measured over correctly and incorrectly classified instances. The Results obtained show XGBoost outperforms with the highest accuracy of 98% comparatively other algorithms. This study's findings can be inform a program for screening potential diabetes patients.

Keywords: Machine learning, random forest, Neural Network, XGBoost, Diabetes, Prediction, KNN, Decision tree.

E-mail:author's email

1. Introduction

Diabetes is a persistent and incurable condition. The enzyme that transports sugar into the platelets is decreased as a result of this illness. This raises the body's glucose levels, leading to serious issues like stroke, lung disease, eyesight loss, kidney failure, and mortality. Patients with diabetes exhibit weight loss, blurred vision, infections, frequent urination [1]. 2019 had 1.5 million diabetes-related deaths, 48% of which happened in adults under the age of 70 [2]. Machine learning techniques perform well. A reference importance can be derived from summarizing and contrasting the results of various classifiers when applied to their individual classification tasks. Evaluates and contrasts five classic ML classifiers—GMM, Random Forest, SVM, XGBoost, and Naive Bayes— Insidious and long-lasting, diabetes to demonstrate how they compute [3]. The performance of the "XGBoost model" on the RNA-seq and GEO datasets, and the comparison of the findings with other models. Studies revealed that the XGBoost model outperformed the current D-GEX algorithm, linear regression, and KNN approaches in terms of overall error. The XGBoost method beats current models and will significantly expand the toolkit for predicting gene expression value [4]. XGBoost-based multi-model combination forecasting approach. Through the outputs of the forecasting model's fitting and forecasting, the approach creates Introduce a novel time series as the set of characteristics. Researchers can choose from a number of models for feature reconstruction when using the results of the forecasting model's predictions. The effectiveness of the features can subsequently be evaluated using the score of the reconstructed features in the training set in the subsequent forecasting model. [5]. In this study, developing XGBoost leads to n a scalable tree boosting system that offers Cutting-edge results on a range of topics. For addressing sparse data, a unique sparsity aware technique is suggested, and for approximate learning, a theoretically supported weighted quantile sketch. A novel approach that takes into account the sparsity of the data is proposed, and for the task of building approximation trees, a weighted quantile sketch is recommended. To develop a tree boosting system that can be easily expanded, as well as provide analysis on cache usage patterns, data compression, and data sharing. XGBoost achieves scalability beyond billions of samples by effectively employing fewer resources compared to existing systems, thanks to the utilization of these findings[6]. In this study, data from the compact polarimetric (CP) RISAT-1 cFRS mode are classified. The Mumbai region was analyzed using the techniques of Artificial Neural Network (ANN) and Extreme Gradient Boosting (XGBoost). After preprocessing, the Raney decomposition approach was used to extract the image's R, G, and B channels. In order to obtain the best parameters for the classification, hyperparameter tuning of ANN was also carried out. In contrast of the two algorithms revealed that they performed almost equally accurately on the data. However, the precision of the XGBoost classifier was just 1% accurate in both the train and test sets. Since the ANN approach needed tuning, it took longer to compute than the XGBoost algorithm, which operates well without tuning [7]. In this work use of the XGBoost algorithm is the prediction of risk assessments in corporate finance. In this study, the data preparation technique is used for successfully beforehand and classify the enterprise revenue information source. Then, the technique of XGBoost is applied to determine the risk of the enterprise financial data. Finally, a set of models for assessing enterprise risk in finance is established, with an accuracy of. The study's findings demonstrate the great reliability of the XGBoost model in enterprises' financial risk assessments are being forecasted with an accuracy rate of less than 3%. The profit and loss of the business's income status accounts for the majority of the projection error, which is only 2.68%. The error is trustworthy enough for corporate use at 0.56%, which is the smallest possible mistake [8].

We developed a classification prediction model based on the aforementioned research background and suggested prediction algorithm based on XGBoost. The outcomes proven that, when compared to, DT, RF, LR and KNN, the XGBoost method greatly increased prediction accuracy. Additionally, it was more capable of generalization and prediction. Last but not least, the XGBoost method was easier to interpret than previous algorithms.

2. Methodology

The current work methodology is classified into the following phases:

A. Data Preprocessing

There are many steps in the preprocessing of data:

-Data filtering is the process of eliminating or deleting particular variables or observations from a dataset. This aids in concentrating the analysis on pertinent data. Extracting of data based on predetermined criteria. It is used for filtering data, erasing undesirable values or extracting data that meets specific criteria. It is crucial to improve the data quality before analyzing or using it in machine learning models.

- Visualization and descriptive statistical analysis involve utilizing illustrations and summary statistics to understand the main characteristics of the dataset, providing to identify patterns, trends, and potential outliers.

- Outlier management and detection: identifying and handling data points that significantly vary from the rest

E-mail:author's email

of the data. Outliers affect statistical analyses, managing them properly guarantees validity.

-Error detection and management is identifying and correcting errors in the dataset to ensure data accuracy and reliability.

-Missing data management includes dealing with missing values in the dataset either imputing values or removing incomplete observations.

-Data reduction refers to reducing the dataset's dimensionality by Principal Component Analysis (PCA) or feature selection to simplify analysis. PCA could reveal that patient age, BMI, and LDL have an essential effect on the first main component, that covers general health factors. The second primary component, that focuses cardiovascular health, may have more to do with VLDL and other issues.

-Data Reduction and Feature Relevance determination is the improvement of the dataset by recognizing important features of the patient and erase the redundancy and building more efficient models.

-Data scaling includes standardizing or normalizing the numeric values of a dataset. This ensures that variables with different units or scales have an equal effect on the analysis and training of the model.

B. Machine Learning Techniques

We used 5 techniques for classification, as described below

1. The XGBoost

XGBoost utilizes decision-tree approaches for arranging the data, following their application to a pre-existing dataset. The primary methodology employed in XGBoost is gradient-boosted trees, which relies on supervised learning. Basically, "supervised learning" refers to an approach where the training data is used as the input. target values are predicted using a model with numerous features. The model, or mathematical method, generates predictions. Use trained data for instance, in a linear model, a mixture of weighted input features is used to establish the prediction. There are numerous parameters in XGBoost that can be used to carry out particular tasks [11].

The strategy exhibits a significant speed improvement, surpassing well-established models commonly employed in machine learning (ML) and deep learning (DL), owing to its utilization of parallel, distributed, out-of-core, and cache-aware computing techniques. This technique also offers the advantage of being extremely efficient and easily adaptable to larger scales. This cutting-edge application of XGBoost was developed to tackle real-world challenges associated with sparse input data. The algorithm takes into account the presence of values that are unavailable, an excessive quantity of zero values in the dataset, and the results of feature engineering techniques that have been applied. The ensemble technique entails iteratively incorporating new models until their inclusion ceases to yield substantial enhancements to the performance of current models. [12].

By optimizing a loss function, the supervised learning method of gradient boosting creates ensembles of decision trees over time. In the current study, G-Boost, a unique instance of gradient boosting, is used. By using decision analysis on its structural components, decision trees have the benefit of interpretability. Because the constituent models are diverse, ensemble methods can learn higher order interactions between features and are scalable [13][14].

2. Logistic Regression (LR)

This function utilizes a solitary multinomial logistic regression model with a lone estimation and is based on class. Logistic regression is a method that determines the boundary between classes and calculates the probability of each class based on the distance from the boundary. As the size of the data set increases, it converges more rapidly into both ends of 0 and 1. These probabilistic statements elevate logistic regression above the level of a simple classifier. It makes more detailed predictions and can be fitted in a different way; however, those robust forecasts may out to be inaccurate. Logistic regression, similar to Ordinary Least Squares (OLS) regression, is a method used for making predictions. Nevertheless, logistic regression produces a dichotomous result when making predictions [15]. Logistic regression is a popular tool, regarding the field of study, it pertains to the practical use of statistical methods and the analysis of data that is not continuous in nature. Linear interpolation is used in logistic regression [16].

3. K-NearestNeighbor (KNN) classifier

Euclidian distance can be used to determine K.-Nearest Neighbor; other measures are also accessible, but Euclidian distance offers a beautiful combination of convenience, effectiveness, and productivity [19].

K-Nearest Neighbor can be conceptualized as a form of analogical learning, as it compares a specific test tuple with a set of training tuples that share similarities with it. The classification is determined by considering the class of the nearest neighbors. Multiple neighbors are often considered, which is why it is called K-Nearest Neighbor (K-NN), where "K" is the number of neighbors used in the classification. For more than 50 years, statisticians have used the K-NN method as a machine learning strategy. The K-NN is frequently referred to as a "lazy learner" because it does nothing more than store the provided training tuples while waiting to be presented with a test tuple, at which point it does generalization in order to learn [20].

4. Decision tree classifier

Each instance is represented by a decision tree utilizing the data consists of a collection of attributes (independent variables or features), where each instance is associated with just one class (dependent variable or outcome class). The type is expressed by the leaf nodes of the tree. The decision tree approach constructs a hierarchical structure of parameters that effectively forecast class labels by utilizing a training set of examples that have been assigned class labels. Every occurrence is transformed into a data point within the description space, and each characteristic serves as a decision point on the map. The description space is then divided into regions, each of which is associated with a different class, by the decision tree process. A new instance's class can then be predicted using the map given its set of specific attribute values.

A classification tree constructs a hierarchy of data structures consisting of nodes. The initial node in the tree is known as the root node, or starting point, while the subsequent nodes are referred to as internal nodes.

Individual points or vertices in a network or graph. Every inner node in the diagram represents a specific test that is used to categorize situations. Each potential result of a test is represented by a child node. For discontinuous characteristics, attribute A can have h potential outcomes, denoted as A = d1 . . . dh, where d1 . . . dh are known attribute values. For a continuous attribute, there are two potential results: Either A is less than or equal to t, or A is greater than t, where t is a value that needs to be determined at the node. The terminal nodes of the tree are referred to as leaf nodes, and their purpose is to determine the class to which the case instance will be allocated. It gives the user a tree that is simple to use in practice and enables them to see the rationale behind it. Decision trees are suitable in many contexts since they do not assume attribute independence like naïve Bayes classifiers do. According to earlier research, the classifier was effective at handling issues with traffic management, marketing, the health insurance sector, gene identification, and medical diagnostics [21][22][23].

5. RANDOM FOREST CLASSIFIER (RF)

Using decision trees constructed depending on the properties of the data. The Random Forest (RF) algorithm is a type of ensemble model that is based on decision trees. Each tree is generated by utilizing a random vector from the input. Every tree contributes one vote to the final forecast using the ensemble architecture. RF is capable of effectively processing sparse datasets, as well as data that contains missing values, noise, and errors. For text categorization issues, LR is frequently utilized. The link between dependent and independent variables can be explained by LR. For text categorization issues, LR is frequently utilized. The link between dependent and independent variables can be explained by LR. LR is a voting classifier that uses statistical estimation to forecast the class with the greatest predicted likelihood [24][25].

C. Proposed System Model

Initially, the dataset is divided into two separate datasets: the training dataset and the testing dataset. The training dataset is employed to create and refine the trained model, while the testing dataset is utilized to evaluate the finished model. Next, the classification method selects and prioritizes the most essential characteristic utilizing XGBoost, DT, RF, SVM, and KNN.

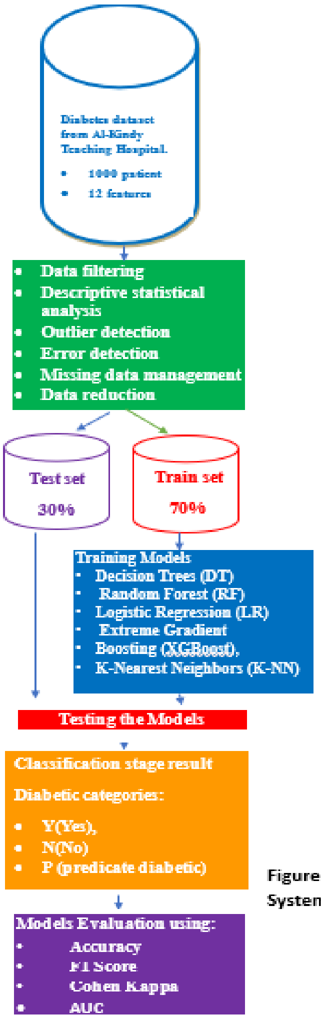

Figure 1: proposed System Methodology

Due to its extensive prior research, we concentrate on the dataset from uci.edu. The dataset, which is comprised of 1000 patients, is summarized in **Table 1.** We utilized correlation as a pre-processing step since correlation is excellent for imputed missing values in the dataset. Correlation can also be used to estimate the casual relationship that exists between the available data. The following phases will employ the pre- processed features.

A model created to predict specific, the main significant procedures used in this study are reviewed in the sections that follow:

Prior to training any model on a given dataset, preprocessing is required. For instance, when using gradient descent to train neural networks. We would have the issue of over-shooting the minima and our model would not converge if one dimension was significantly larger than the other. Standardization also makes it simpler to process, analyze. With this strategy, firms may use their data to make better decisions. Companies can compare and analyze data more easily to gain insights about how to improve their operations when it is standardized.

One of the most significant advantages of data standardization is that it assists businesses in avoiding making judgments based on inadequate or erroneous data. Companies may make better decisions to increase their bottom line by using data standardization to ensure that their data is comprehensive and correct. The z-score's foundation is made up of the mean and standard deviation. A technique for standardizing scores on the same scale is the z-score (also known as the standard score). It calculates the ratio of the departure of a score from the mean to the standard deviation of the data set. The standard deviation represents the measure of dispersion of a data point from the mean, and it is sometimes referred to as the final score. The summation of all z-scores inside a dataset yields a value of zero. A negative z-score indicates that the value is situated below the mean. A high z score indicates that the value is above the mean.

D. Model Evaluation Metrics

We use a set of normative measurements and indices, including metrics for accuracy, precision, the integral of the receiver operating characteristic curve (ROC), and detection rates, to test and analyse the proposed model. The confusion matrix (CM) was utilised to quantify the efficacy of the classifier. The testing dataset's records are evaluated using the metric.

1) Accuracy

Accuracy refers to the fraction of correctly detected observations [14]. The effectiveness of the classification algorithm model under test is determined by how accurate the findings are. Equation formulates accuracy is Equation (1).

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \qquad (1)$$

2) F-Measure (F1-Score)

Precision and recall are balanced, as seen by the F1-Score. The F1-Score, which ranges from 0 (worst score) to 1 (highest score), Equation 3.3.2 is used to formulate the F1-Score:

$$\text{F1-score}=2\times(\text{precision} \times\text{recall})/(\text{precision} +\text{recall}) \qquad (2)$$

Where, Precision is the proportion of observations in a batch of observations that the classifier correctly identifies as having positive data. Equation (3) formulates precision.

$$\text{Precision} = TP/(TP+FP) \qquad (3)$$

Recall is the proportion of observations that were successfully categorized and given a positive label. Equation (4) formulates recall.

$$\text{Recall}=TP/(TP+FN) \qquad (4)$$

3) Cohen Kappa

Cohen proposed the following interpretation of the Kappa outcome: rates less than or equal to 0 demonstrate inconsistency, values between 0.01-0.20 indicate little to no acceptance, values between 0.21-0.40 indicate fairness, values between 0.41 and 0.60 indicate moderateness, values between 0.61-0.80 indicate substantialness, and values between 0.81 and 1.00 indicate nearly perfect consistency [26].

4) AUC

The classifier's ability to detect classes is represented by the area under the curve (AUC). If the AUC is 1, the classifier detects class labels perfectly, whereas 0.5 represents random selection. AUC has been shown to be insensitive to an imbalanced dataset [27][28][29].

The variables TP, FP, and FN represent the number of correct positives forecasting, incorrect positives forecasting, and incorrect negatives forecasting, accordingly, for a particular category name.

3. Experimental Results

A. Dataset Description

These are techniques are particularly efficient in reducing the amount of data needed to filter out the noise, which reduces storage requirements, shortens processing times, and improves classifier accuracy.

The data was collected from the Iraqi society, mainly from the laboratory of **Medical City Hospital and the Specialises Centre for Endocrinology and Diabetes-Al-Kindy Teaching Hospital.** The patients' data were collected and extracted from them; it was entered into the database to create the diabetes dataset. The data involve medical information, laboratory analysis, and other relevant information of 1000 patients. The initial data entered into the system include: patient id, patient number, Blood sugar level, Age, Gender, Creatinine ratio (Cr), Body Mass Index (BMI), Urea, Cholesterol (Chol), Fasting lipid profile (including total, LDL, VLDL, Triglycerides (TG), and HDL Cholesterol), HBA1C, and Class (which correspond to the patient's diabetes diagnosis as Diabetic, Non-Diabetic, or Predict-Diabetic). Set of dataset features in Table 1. Train-Test split is dividing the dataset into two subsets, in 70% for training the model and 30% to test its performance.

TABLE 1. DATASET DESCRIPTION (1000 PATIENTS ,13 FEATURES)

Dataset features	**Features description**
N0_patient	patient number
Sex	F or M
Age	Patient age
Urea	
Cr	creatinine ratio
HbAlc	blood test that is used to diagnose type 2 diabetes
Chol	Cholesterol
TG	triglyceride
HDL	high density lipoprotein

Dataset features	Features description
LDL	low-density lipoprotein
VLDL	very low-density lipoprotein
BMI	body mass index
Class	Diabetic categories: Y(Yes), N(No), and P (predicate diabetic)

The Spearman correlation coefficient for each pair of features is shown in Figure 2. BMI, HbA1c, and age have the strongest correlation with the diagnosis.

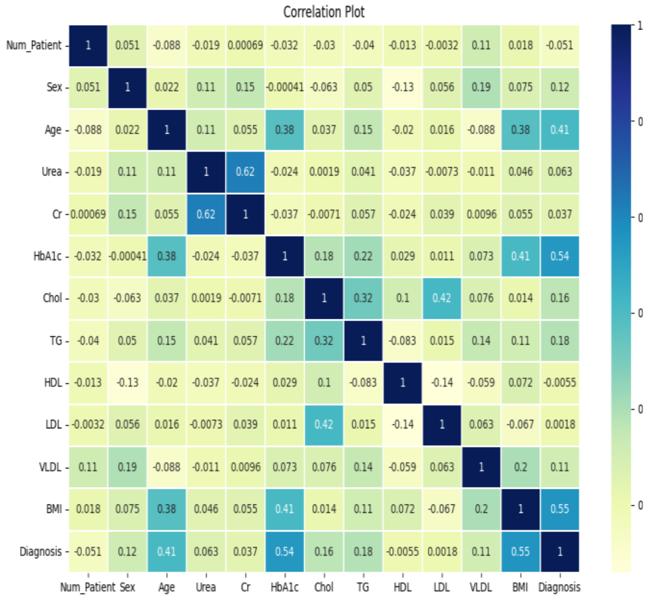

FIGURE 2. CORRELATION PLO==

B. Parameters Setting

The parameters that determine how a machine-learning model behaves are called hyperparameters. Training does not acquire these parameters; rather, they are pre-set. Hyperparameter optimization is the process of finding the best values for these parameters and is a crucial part in developing a Machine Learning model.

Hyperparameter optimization can be done in many different ways; however, grid search and randomized search happen to be the most popular.

In order to optimize the model's hyperparameters, we apply the grid search method, which requires constructing a list of all possible values for each hyperparameter and then training the model in each combination of those parameters. For an example, we would specify a list of values for the neighbors of values (3,5,7,9) if we wanted to optimize hyperparameters of the **KNN's,** using grid search. Next, the grid search algorithm would use such variables to train a model and compare how well each model performed. The models' performance is used to choose the hyperparameters' ideal values.

In order to optimize the hyperparameters, the models also use cross-validation. The process of cross-validation entails dividing the training data into several sets and then training the model repeatedly with a new set of data used as validation each time. This can aid in preventing overfitting and give a more accurate evaluation of the model's performance. the GridSearchCV technique's works through a grid of values for each hyperparameter and finds the combination that produces greatest model performance. we use GridSearchCV with Logistic Regression to optimize parameters that drive the behavior of the model as well as the ideal balance between model complexity and generalization. Typical **logistic regression** parameters include: regularization in order to avoid overfitting. Solver is an optimization algorithm. multiclass as Diabetic, Non-Diabetic, or Predict-Diabetic, as well as the size of the dataset, are factors that affect the choice of solver with LR.

Some of variables that help control the structure and behavior of the **DT** are the maximum depth and the minimum number of samples required to be in a leaf node. These parameters are adjusted based on the dataset and the performance that is we needed via the model.

Consider adjusting the minimum samples required to separate a node, the minimum samples required at a leaf node, and the maximum depth of trees when optimizing Random Forest parameters. Perform experiments to find out which combination of the parameters maximizes the impact of the model. bootstrap assess the impact that different parameter values have on generalization with model.

XGBoost is an effective powerful machine learning algorithm in our model. Some key XGBoost parameters are: Rate of Learning (eta), a slower learning rate makes the model more reliable, but it demands more trees. The number of trees refers to the total number of boosting rounds or trees to be built. The following are some of the parameters that were utilized to obtain the results (a brief explanation of each parameter's function follows).

In essence, the "learning rate" parameter is set to eliminate overfitting issues. Weights pertaining to the new features are extracted and the step size shrinking is performed. A tree's depth is determined by the "max_depth" option; the higher the value more complicated. The number of rounds or trees employed in the model is determined.

A learning parameter is the "random state" parameter. It is sometimes referred to as "seed" on occasion The dataset is divided into k sections.

The tree booster parameters described above were used to calculate the findings that are listed below. There are numerous settings that can be set up, but the model is mostly responsible for this. Although one can define the parameters in accordance with the intended model.

Tree Maximum Depth, this variable sets the maximum depth for every decision tree in the ensemble. Minimum Child Weight is the sets of a certain quantity of instance weight that a child must weigh at least. It can be utilized to manage over-fitting.

C.Results

The Python programming language was used in this work to implement the suggested techniques on a Windows 10-based computer system with an Intel® Core TM i7 CPU running at 7 GHz and 8.00 GB of total RAM. All of the dataset's data are first normalized, and then each dataset was split into two sets: The training set contains 70% of the data, and the testing set contains the remaining 30%.

There are several indicators of performance for classifiers and predictors. For the algorithms used in this study, we chose four indicators (more used), accuracy, F1 score, Cohen's kappa, and ROC AUC with the one-versus-rest method. The performance of the individual models is shown in a comparative Table (2).
From the Table 2 we can clearly see that XGBoost yields optimal outcomes with relation to accuracy, F1 Score, Cohen Kappa and AUC.
To discuss on the mean and standard deviation (sd) of DT, RF, LR, XGBoost, and k-NN models:

1.The range of mean AUC values, indicating prefect discriminative values across models, is 0.937953 to 0.997889. Low performance variability within each model is indicated by standard deviation.

2. The range of mean Kappa values, which show various levels of agreement beyond chance, is 0.631183 to 0.953714.Values for the standard deviation illustrate variations in the agreement among models.

3. The mean F1 score values range from 0.626668 to 0.971192, indicating the models' effectiveness in balancing precision and recall. Values for the standard deviation show that the harmonic mean of recall and precision can vary.

4. The mean accuracy values is in a range of 0.985925 to 0.899828, indicating the percentages of instances that are accurately classified. Variability in classification accuracy among models is represented by the values of the standard deviation.
By comparing and contrasting the distributions of various performance metrics among models, these talks shed light on the consistency and dependability of those measurements. This finding is joining our state of art study and confirming that the XGBoost outperform DT, KNN, RF and LR for Diabetes detection.
As shown in Figure 3, Among the suggested models include interaction terms, XGBoost performed the best, with 98.5% accuracy, 97.1% F1 Score, 95.4% Cohen Kappa, and a 99.7% AUC.

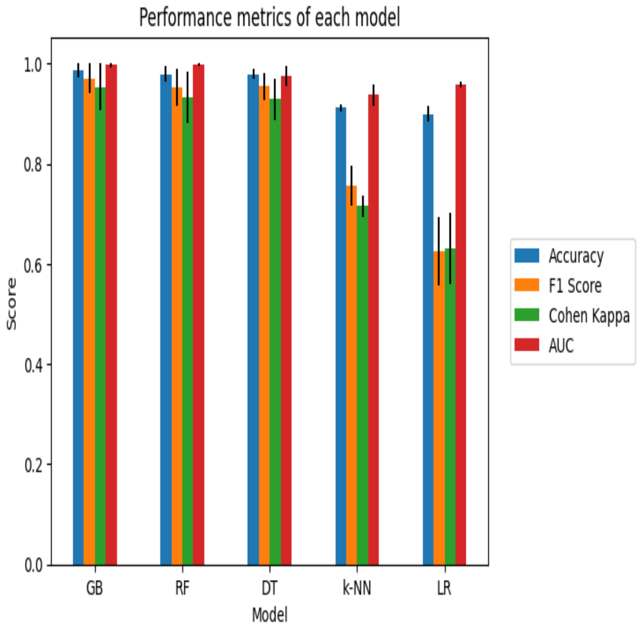

Figure 3.performance metrics

6. CONCLUSION

This research has been dedicated to the creation of computerized tools for decision-making that aid doctors in various aspects of patient care. Developers of these systems commonly assert that these programmed enhance the precision of healthcare diagnosis and lead to improved patient outcomes.
In this research, we have introduced work with multiple machines learning techniques, including Decision Trees (DT), Random Forest (RF), Logistic Regression (LR), Gradient Boosting (XGBoost), and K-Nearest Neighbors (k-NN).
Then, what distinguishes this work is the in-depth research to reach results that give a decision in directions to determine the efficiency.

The performance results show a good accuracy value, namely 98 % for XGBoost can be concluded as quite good in contrast with other model that are used in this research.

We prefer to refine the model in the future by incorporating more data from various sources and considering other machines learning techniques.

Models	Accuracy		F1 Score		Cohen Kappa		AUC	
	mean	sd	mean	sd	mean	sd	mean	sd
GB	**0.985925**	**1.1703E-16**	**0.971192**	**0**	**0.953714**	**0**	**0.99761**	**1.1703E-16**
RF	0.979652	0.001473	0.952213	0.00454	0.932438	0.005095	0.997889	0.000147
DT	0.97998	0.000663	0.955136	0.001786	0.930079	0.002323	0.978092	0.00265
k-NN	0.912365	0	0.756725	1.17028E-16	0.715168	1.17028E-16	0.937953	0
LR	0.899828	0	0.626668	1.17028E-16	0.631183	0	0.958123	0

TABLE 2.THE EFFECTIVENESS OF THE FIVE MACHINE LEARNING MODELING METHODS FOR CLASSIFYING DIABETICS

REFERENCES

[1] Ali S. Mosa, Zainab N. Nemer," COVID-19 Diagnosis based on Chest X-ray using Deep Convolution Neural Network and Testing the Software Complexity using Halstead Metrics and Artificial Neural Network", Advances in Mechanics Vol. 9, Issue 3, 2021 P: 780 - 802 780, 2021.

[2] Hanan Q. Jaleel a, Jane J. Stephan b, Sinan A. Naji "Textual Dataset Classification Using Supervised Machine Learning Techniques" و Engineering and Technology Journal 40 (04) (2022) 527- 538

[3] Haoyuan Tan," Machine Learning Algorithm for Classification", International Conference on Big Data and Intelligent Algorithms BDIA 2021, doi:10.1088/1742-6596/1994/1/012016.

[4] K.Koteswara Chari, M.Chinna babu, Sarangarm Kodati," Classification of Diabetes using Random Forest with Feature Selection Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Vol.9 Issue-1, November 2019

[5] Methaporn Phongying and Sasiprapa Hiriote," Diabetes Classification Using Machine Learning Techniques", Computation, may,2023 doi.org/10.3390/computation11050096

[6] Nimrabanu Memon, Samir B. Patel, and Dhruvesh P. Patel," Comparative Analysis of Artificial Neural Network and XGBoost Algorithm for PolSAR Image Classification", Springer Nature Switzerland, pp. 452–460, Vol.11941, AG 2019. https://doi.org/10.1007/978-3-030-34869-4_49.

[7] Rongyuan Qin, "The Construction of Corporate Financial Management Risk Model Based on XGBoost Algorithm", Volume 2022 | Article ID 2043369 | 13 Apr 2022. ttps://doi.org/10.1155/2022/2043369.

[8] Tianqi Chen,Carlos Guestrin ," XGBoost: A Scalable Tree Boosting System", KDD '16, August 13-17, 2016, San Francisco, CA, USA, 2016.ACM.,DOI:http://dx.doi.org/10.1145/2939672.2939785.

[9] Wei Li, Yanbin Yin, Xiongwen Quan1 and Han," Gene Expression Value Prediction Based on XGBoost Algorithm", Zhang1,3*ORIGINAL RESEARCH article, Vol.10, Nov. 2019,

[10] Zhen Li a, Tieding Lu a,b, Xiaoxing He c , Jean-Philippe Montillet d,e , Rui Tao," An improved cyclic multi model-eXtreme gradient boosting (CMM-XGBoost) forecasting algorithm on the GNSS vertical time series", advance in space research , ISSN: 1,P:912-935,Vol.71, Jan . 2023.

[11] Sukhpreet Sign, Abudlah Nahid, Roert Abbas," Effective Intrusion Detection System Using XGBoost", Vol 9, Issue 7, June 2018.

[12] Anna Palecsek,Dominik Grochala,Arthur Rydosz , "Artificial Breath Classification Using XGBoost Algorithm for Diabetes Detection", MDPI ,Vol.21, Issue12,18 June 2021.

[13] Hardik Rajpal, Madalina Sas, Chris Lockwood, Rebecca Joakim, Nicholas S Peters, Max Falkenberg, "Interpretable XGBoost Based Classification of 12-lead ECGs Applying Information Theory Measures from Neuroscience", PMC, Mar ,2021.

[14] Munisamy Eswara Narayanan, Balasundaram Muthukumar," Malware Classification Using Xgboost with Vote Based Backward Feature Elimination Technique", Turkish Journal of Computer and Mathematics Education Vol.12 No.10 2021.

[15] Newsom, I. (2015). Data Analysis II: Logistic Regression. Available at: http://web.pdx.edu/~newsomj/da2/ho_logistic.pdf

Logistic Regression pp. 223 – 237. Available at:https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf .

[16] Dakhaz Mustafa Abdullah, Adnan Mohsin Abdulazeez," Machine Learning Applications based on SVM Classification: A Review", Qubahan Academic Journal, Vol.1, No.2,2021

[17] Durgesh K. Srivastava, Lekha Bnamhu," Data Classification Using Support Vector Machine", Journal of Theoretical and Applied Information,

[18] F R Lumbanraja, E Fitri, Ardiansyah, A Junaidi, Rizky Prabowo," Abstract Classification Using Support Vector Machine Algorithm (Case Study: Abstract in a Computer Science",Journal of Physics: Conference Series 1751 (2021)doi:10.1088/1742-6596/1751/1/012042

[19] H Sain, H Kuswanto, S W Purnami and S P Rahayu," Classification of rainfall data using support vector machine ", Journal of Physics: Conference Series 1763 (2021) 012048 IOP Publishing, 2020. doi:10.1088/1742-6596/1763/1/012048 1

[20] Kataria1, M. D. Singh21P.G. Scholar, Review of Data Classification Using K-Nearest Neighbour Algorithm", international Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Vol. 3, Issue 6, June 2013.

[21] D.A. Adeniyi, Z. Wei, Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", Applied Computing and Informatics ,2016.

[22] Abdullah Awaysheh1, Jeffrey Wilcke1, Franc¸ois Elvinger," Review of Medical Decision Support and Machine-Learning Methods", Veterinary Pathology, Vol. 56(4) 512-525, 2019.DOI:10.1177/0300985819829524

[23] Anitha Juliette Albert • R. Murugan • T. Sripriya. "Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology ", Research on Biomedical Engineering ,2023.

[24] B A C Permana, R Ahmad, H Bahtiar, A Sudianto and I Gunawan, "Classification of diabetes disease using decision tree algorithm", Annual Conference on Science and Technology (ANCOSET 2020) Journal of Physics: Conference Series 1869 ,2021 doi:10.1088/1742-6596/1869/1/012082

[25] M. M. Imran Molla, Julakha Jahan Jui, Bifta Sama Bari," Cardiotocogram Data Classification Using Random Forest Based Machine Learning Algorithm ", © Springer Nature Singapore Pte Ltd. 2021.

[26] Nasir Jalal a, Arif Mehmood a, Gyu Sang Choi b, Imran Ashraf b," A novel improved random forest for text classification using feature ranking and optimal number of trees", Journal of King Saud University – Computer and Information Sciences Journal of King Saud University – Computer and Information Sciences 34 (2022) 2733–2742

[27] Mary L., “Interrater reliability: the kappa statistic”, Croatian Society of Medical Biochemistry and Laboratory Medicine, 22(3): 276–282.Published online 2012 Oct 15.

[28] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition. Lett., vol. 27, no. 8, pp. 861–874, 2006.

[29] Raheem, Sabreen Fawzi, and Maytham Alabbas. "A Modified Spider Monkey Optimization Algorithm Based on Good-Point Set and Enhancing Position Update." Informatica 47.4 (2023).

Dr. Zainab N. Nemer is currently an Assist prof in the Department of Computer Science at the University of Basrah where she has been a faculty member since 2003. She received her PhD degree (2009) in Computer Science from the Basrah University, Iraq. She received her MSc (2000) and BSc (1996) in Computer Science from the University of Basrah, Iraq. Her current research concerns are Artificial Intelligence and soft computing. She published more than 11 journals papers. She can be contacted at email: zainab.nemer@uobasrah.edu.iq

Sabreen Fawzi Raheem is currently an Assistant Lecturer in the Basra Technical Institute at Southern Technical University – Iraq. I got an M.Sc in Computer Science in the field of Artificial Intelligence from the University of Basrah at the College of Computer Science and Information Technology.

Dr. Maytham Alabbas is currently a Professor in the Department of Computer Science at the University of Basrah where he has been a faculty member since 2003. He received his PhD degree (2013) in Computer Science from the University of Manchester, UK. His PhD thesis has been awarded the 2014 Best Thesis Prize of the School of Computer Science at the University of Manchester. He received his MSc (2002) and BSc (1999) in Computer Science from the University of Basrah, Iraq. His current research concerns are Artificial Intelligence, NLP, Machine Learning, Computational Linguistic, and Language Engineering. He has served on different conference and workshop program committees such as IntelliSys 2021-2019, AMLTA 2019, ACLing 2018, and AISI 2018. He published more than 22 journal papers and 16 conference papers. He has ACM professional membership. He can be contacted at email: ma@uobasrah.edu.iq