JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 06 Issue 02
2024

Research Article
https://doi.org/10.56979/602/2024

Boosting Early Diabetes Detection: An Ensemble Learning Approach with XGBoost and LightGBM

Faheem Mazhar[1], Wasif Akbar[1], Muhammad Sajid[2*], Naeem Aslam[1], Muhammad Imran[1], and Haroon Ahmad[2]

[1]Department of Computer Science, NFC-IET, Multan, Pakistan.
[2]Department of Computer Science, Air University Islamabad, Multan Campus, Multan 60000, Pakistan.
[*]Corresponding Author: Muhammad Sajid. Email: msajid@aumc.edu.pk

Received: January 25, 2024 **Accepted:** February 29, 2024 **Published:** March 01, 2024

Abstract: Given the increased prevalence of diabetes, early identification and prognosis of the condition are essential to avoiding long-term health consequences. Diabetes is an enduring medical illness that may have a role in the global health crises. The International Diabetes Federation estimates that 382 million people worldwide have diabetes. This number is expected to double by 2035, to reach 592 million. A medical condition known as diabetes is brought on by an excessively high blood glucose level. Diabetes is the main cause of renal failure, blindness, amputations, heart failure, and stroke. In order to develop a computerised approach for diabetes prediction, this work uses machine learning (ML) techniques on the Pima Indians dataset and private diabetes information. The aim of this project is to combine the findings from multiple machine learning techniques to create a system that can more accurately predict a patient's risk of developing diabetes in their early years. Techniques including logistic regression, SVM, RF, KNN, and decision trees are used. For every algorithm, the model's accuracy is computed. The model that predicts diabetes with the best accuracy is then chosen. We have achieved remarkable results in terms of accuracy, precision, recall, and F1-score for the models on the dataset by utilising several machine learning classifiers and putting feature removal techniques like feature permutation and hierarchical clustering into practice. This suggests that our characteristics or data are not limited to specific models.

Keywords: Machine Learning; Prediction; Diabetes; XGBOOST; Classification; Random Forest(RF).

1. Introduction

The processing of sensitive and crucial healthcare data has significantly progressed due to breakthroughs in biotechnology and the public health infrastructure. Through the utilisation of crucial characteristics, the application of sophisticated data analysis methods has enabled the detection and mitigation of certain chronic diseases at their first phases [1]. Diabetes is increasingly prevalent among individuals of all age groups, encompassing both young children and the elderly. If the incidence of this persistent ailment continues to increase, it has the potential to evolve into a global health concern. Diabetes is characterised by a variety of symptoms, including heightened thirst, fatigue, reduced appetite, weight loss, blurred vision, mood swings, confusion, difficulties focusing, and frequent infections. Diabetes presents a significant threat to an individual's life, elevating the probability of experiencing strokes, vision loss, pregnancy loss, limb removal, and renal failure, among other fatal diseases [2]. According to the IDF,

the prevalence of diagnosed diabetes is steadily spreading worldwide. Projections indicate that by 2030, the global diabetic population will reach 642.8 million. According to the 2023 statistics, the projected number of individuals with diabetes in Saudi Arabia by 2030 is estimated to exceed 5.61 million. Machine learning (ML), a burgeoning subject in artificial intelligence, focuses on the exploration of how computers might acquire knowledge and improve their performance through interactions with the environment.

To fully grasp the occurrence of diabetes, it is crucial to have a thorough understanding of how a healthy body functions. The meals we consume, particularly those that contain a significant amount of carbohydrates, supply our bodies with the necessary sugar (glucose). Both individuals, including the one with diabetes, rely on meals that are rich in carbs as their primary source of nutritional energy. Upon digestion, typical carbohydrate are converted into glucose [3].

A portion of the glucose is distributed throughout the body by the circulatory system to improve brain function, while the remainder is either stored in the liver or promptly used by cells as energy. Insulin is essential for the utilisation of glucose as an energy substrate and is synthesised by beta cells in the pancreas. Insulin acts as a receptor, binding to cellular gateways and facilitating the entry of glucose from the bloodstream [4]. Diabetes is the term used to describe conditions where insulin resistance or insufficient insulin production resulting from pancreatic malfunction prevent the body from using its own manufactured insulin properly. That being said, high blood glucose levels result in hyperglycemia, which is the first sign of diabetes [5].

As stated in [1], There are three primary categories into which diabetes falls. Diabetes type 1 is the most prevalent and is characterised by inadequate insulin synthesis by cells, leading to compromised immune function. There is a scarcity of definitive proof. The hallmark of type2 diabetes is the body's cells' incapacity to either generate sufficient insulin or effectively utilise it after its production. This particular type of diabetes accounts for 90% of instances and is prevalent among the majority of individuals diagnosed with diabetes.

Type2 diabetes arises from a mix of genetic and lifestyle factors. Elevated blood glucose levels in pregnant women can lead to the development of gestational diabetes, posing risks to both the mother and the foetus. There is a strong probability that gestational diabetes may happen again in future pregnancies, and women with this disease are more prone to developing type 1 or type 2 diabetes after giving child. Due to its inherent hazards, it is imperative to seek prompt medical intervention for all types of diabetes. Early detection can prevent complications arising from these illnesses [6].

Diabetes mellitus is a common and long-lasting medical condition. By 2045, the occurrence of diabetes is projected to rise to 10.9%. In China, diabetes affects 20–40% of people concurrently with renal problems; diabetic nephropathy (DN) is the primary cause of end-stage chronic kidney disease. [7]. On the other hand, those with diabetic nephropathy experience a significantly higher mortality rate, ranging from 20 to 40 times greater than those without the condition, across all causes of death. The adoption of innovative screening and treatment techniques bears noteworthy consequences for the country's efforts to alleviate diabetic nephropathy. [8].

Utilising metabolomic data to pinpoint specific pathophysiological mechanisms and find new prognostic and diagnostic biomarkers connected to the onset of diseases has garnered attention recently [8]. Being aware that this review article is the first to discuss the application of AI and ML to diagnosis, treatment customisation, and self-management of DM [9]. Review papers are valuable because they offer a comprehensive overview of the most recent research in a particular field of study [10]. Moreover, the authors have solely focused on machine learning processes; they have not addressed several crucial ML-associated subjects, such as databases, pre-processing techniques, and feature extraction and selection

strategies employed to find DM and AI answers to the requirement for intelligent DM assistants [11]. The selection of papers from the Scopus and PubMed databases is done using a systematic decision-making framework due to the complexity and variety of DM detection and diagnosis, as well as self-management and personalisation systems. The following goals are achieved by this approach [12].

The discussion of datasets, pre-processing methods, strategies for selecting DM features, and machine learning approaches for DM detection Artificial intelligence-based intelligent DM assistant; and (6) performance matrices [13]. 107 current, pertinent studies have been gathered from the Scopus and PubMed databases after intensive search techniques.[14]. This paper is anticipated to benefit the research communities investigating selfmanagement, personalised discipline, and DM diagnosis and detection [15].

2. Related Work

The authors [16] proposed a model that could be utilised to correctly diagnose patients with diabetes. This method relies on the expected accuracy of robust machine learning algorithms, which employ metrics like recall, precision, and F1-measure. The PIDD dataset is employed by the authors to predict the incidence of diabetes through diagnostic methodologies. The accuracy rates of the (KNN), (NB), and Logistic Regression (LR) algorithms were 89%, 79%, and 69%, respectively. Seven machine learning methods were employed by the researchers in the mentioned article [17] to forecast the prevalence of diabetes based on a particular dataset.

Several machine learning classification techniques, including Gaussian Naive Bayes, K-Nearest Neighbours, Artificial Neural Network, Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine, are applied to the PIID dataset in the study reported in paper [18]. The accuracy of logistic regression was shown to be higher. Predictive analytics in the healthcare sector is described by the authors in [19]. This research makes extensive use of machine learning techniques. For the aim of performing tests, a patient's medical dataset is obtained. We compare and discuss the accuracy and performance of the suitable algorithms. In their paper [20], the authors provide a diabetes prediction model. To appropriately categorise diabetes, in addition to traditional indicators like glucose, BMI, age, insulin, etc., outside factors that contribute to the disease's progression must be taken into account.

Remarkably, Random Forest classifier achieves an accuracy rate of 87.66% to win. Furthermore, the scientists [26] have developed algorithms to categorise and forecast diabetes-related events. This study categorised and predicted eight diabetes-related problems using a variety of supervised classification techniques. These results are influenced by several factors, such include obesity, diabetic foot, retinal degeneration, metabolic syndrome, dyslipidemia, which and nephritis. The authors of [27] outline two machine learning techniques for identifying diabetics. For the hybrid approach, use the XGBoost algorithm; for the classification strategy, use the Random Forest algorithm. XGBoost outperforms the other approaches, as evidenced by its accuracy rate of 74.10%.

A variety of machine learning approaches were examined by the authors of this work [28], including logistic regression, decision trees, random forests, gradient boost, K-nearest neighbour, support vector machines, and the NBayes algorithm. The results demonstrated that, with an accuracy of 80%, the Random Forest and NaivBayes classifiers performed better than the other algorithms.

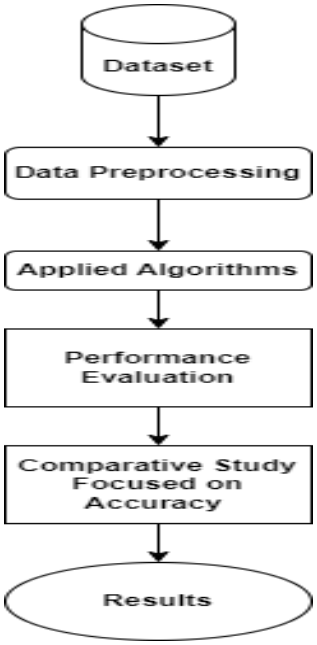

Figure 1. Proposed Model Workflow

3. Materials and Methods

3.1 Preprocessing

Data preprocessing is a crucial stage in data mining that transforms the data into an ideal and useable form when working with noisy, inconsistent, or missing data [29]. To regularly generate data in a coherent and accurate manner, data preparation involves a variety of operations such as data transformation, data reduction, data integration, data cleansing, and discretization [30]. For this study, nine different diabetes-related features of data were collected via the Kaggle platform, which houses a number of datasets. Nine parameters in the dataset under consideration reflect hospital and patient outcomes. It has been used to assess the prediction's accuracy using ensemble techniques.

3.2 Ensemble Model.

Because the number of people with diabetes is increasing, ensemble approaches have been used to analyse diabetes data, making it essential to forecast the risk of acquiring the disease in the future. In order to minimise bias and variation and enhance predictions, ensemble learning is a data mining methodology that integrates multiple approaches into a single ideal predictive model. This technique gives greater predictive performance as compared to a single model. The research employed AdasBoost, Bagging, and RF ensemble techniques to predict the probability of early-onset diabetes [31]. Data mining, statistical analysis, and exploration were the three main uses of Weka. Weka's default parameters were used [32]. An ensemble technique called AdaBoost is employed to address categorization issues. It is a part of the ensemble approaches known as the boosting family, which involve the sequential addition of new machine learning models, each of which seeks to correct prediction mistakes caused by earlier models. AdaBoost is the first successful use of this kind of model. Concise decision tree models with just one decision point per model were used in the development of AdaBoost.

Short trees are another name for decision trees [33]. Combining several models to enhance regression and classification tasks is known as bootstrap aggregation, or bagging in some cases. Using numerous random samples of data, the Bootstrap technique is used to generate a statistical measure, such as the mean. This approach is suggested when there is a lack of available data and a more reliable estimate of a statistical measure is needed. When used on models with high variation and little bias, this approach works well. As a result, the training set of data greatly influences their predictions. Frequently used as a Bagging technique

that meets the high variance criteria are decision trees. Decision trees are used in the Random Forest classification and regression method, which is based on bagging. One of the drawbacks of bagged decision trees is that they are constructed using a greedy algorithm that locates the best split point at each stage of the tree-building procedure. Because of this, the trees that are formed all have a similar appearance, which reduces both the variability and predictability of the forecasts from each bag. [34].

This section will explore the various classifiers used in machine learning to predict diabetes. We will also go over our suggested methods in an effort to increase accuracy. This paper used many approaches, Here is a description of the many methods. The output is the accuracy measurements of the machine learning models. Following that, the model can be used to make predictions.

Mathematical technique that use Random Forest to predict diabetes [35] are shown in the equations. Let Prob denote the probability that a patient has diabetes, and the input variables are X1, X2,..., Xn. Thus, Equation 1 can be employed to depict the Random Forest model:

$$Prob = RF(X) \quad (1)$$

The acronym RF stands for Random Forest, which is a model that combines many decision trees. Equation 2 is used to determine the contribution of each DT Ki in the RF to the prediction, based on the majority vote of the decision trees.

$$Prob(Ki) = Ki(X) \quad (2)$$

The ultimate probability is calculated using Equation 3, which computes the mean of the probabilities from all the decision trees.

$$Prob = 1/n * SUM(Prob(Ki)) \quad (3)$$

Every decision tree takes input variables X1, X2, ..., and Xn and generates a binary decision by comparing each node's value to a threshold.

4. Results and Discussion

4.1 Evaluation Metrics

Evaluating performance is an essential undertaking within the ML domain. Choosing the suitable parameters for evaluating the ML model is of utmost importance. Metrics are employed to evaluate and measure the effectiveness of machine learning algorithms. ML algorithms are evaluated using many performance metrics, such as RootMeanSquaredError, Root Relative Squared Error, F-Measure, ROC Area, Accuracy, Precision, Recall, and others.

4.1.1 Confusion Matrix

Every one of the lengthy studies was assessed using a variety of measures, each of which had a unique evaluation definition. The True-Positive (TP), False-Positive (FP) confusion matrix True-Negative (TN), False Negative (FN), and Positive (FP).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

$$Precision = \frac{TP + FP}{TP} \quad (5)$$

$$Recall = \frac{TP + FN}{TP} \quad (6)$$

$$F1\_Score = \frac{Precision + Recall}{2 \times Precision \times Recall} \quad (7)$$

It's important to choose the metrics that align with the goals of the specific application.

Table 1. Confusion Matrix

Predicted Class	Actual Class	
	Positive	Negative
	TP	FP
	FN	TN

4.2 Dataset

The diabetes dataset was initially acquired from the Kaggle website. The diabetes dataset contains 768 observations and consists of 9 variables. The objective is to ascertain If diabetes is present or not in the patient by utilising the measurements. The diabetes data collection consists of 768 data points, each containing 9 features. The feature we will predict is referred to as "Outcome," with a number of 0 signifying the absence of diabetes and a number of 1 signifying its existence.. The dataset is devoid of any null values.

Table 2. PIMA Indians Dataset

Preg	GL	BP	ST	Insulin	BMI	DPF	Age	Outcome
7	149	71	37	0	33.6	0.629	51	1
2	86	62	25	0	26.5	0.356	32	0
9	182	63	0	0	23.4	0.675	33	1
3	81	64	24	93	28.3	0.164	24	0
4	133	45	32	164	43.2	2.283	35	1

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Figure 2. Dataset Description

Upon applying a range of ML Algorithms to the dataset, we obtained the following accuracies. XGBOOST algorithm achieves a maximum accuracy of 90%.

Table 3. Accuracy Table

Classifier	Accuracy	F1 Score	Recall	Precision
SVM	85%	0.79	0.73	0.77
Random Forest	88%	0.75	0.66	0.69
LightGBM	86%	0.86	0.82	0.83
Logistic Regression	84%	0.78	0.87	0.89
XGBoost Classifier	90%	0.88	0.86	0.89
Decision Tree	85%	0.87	0.85	0.81
KNN	85%	0.89	0.84	0.82

4.3 Correlation Matrix

There is clear evidence that none of the individual attributes have a meaningful correlation with the worth of our outcome. Certain attributes exhibit a negative correlation with the outcome value, whereas other attributes show a positive correlation. Let's analyse the plotlines. It additionally illustrates the dispersion of each characteristic and label across different intervals, highlighting the need for scaling. Furthermore, each individual bar represents a distinct category variable in actuality. Before employing machine learning, it is necessary to consider and handle these categorical aspects. We employ two categorizations for our outcome labels: 0 shows that there is no disease, while 1 indicates that there is disease.

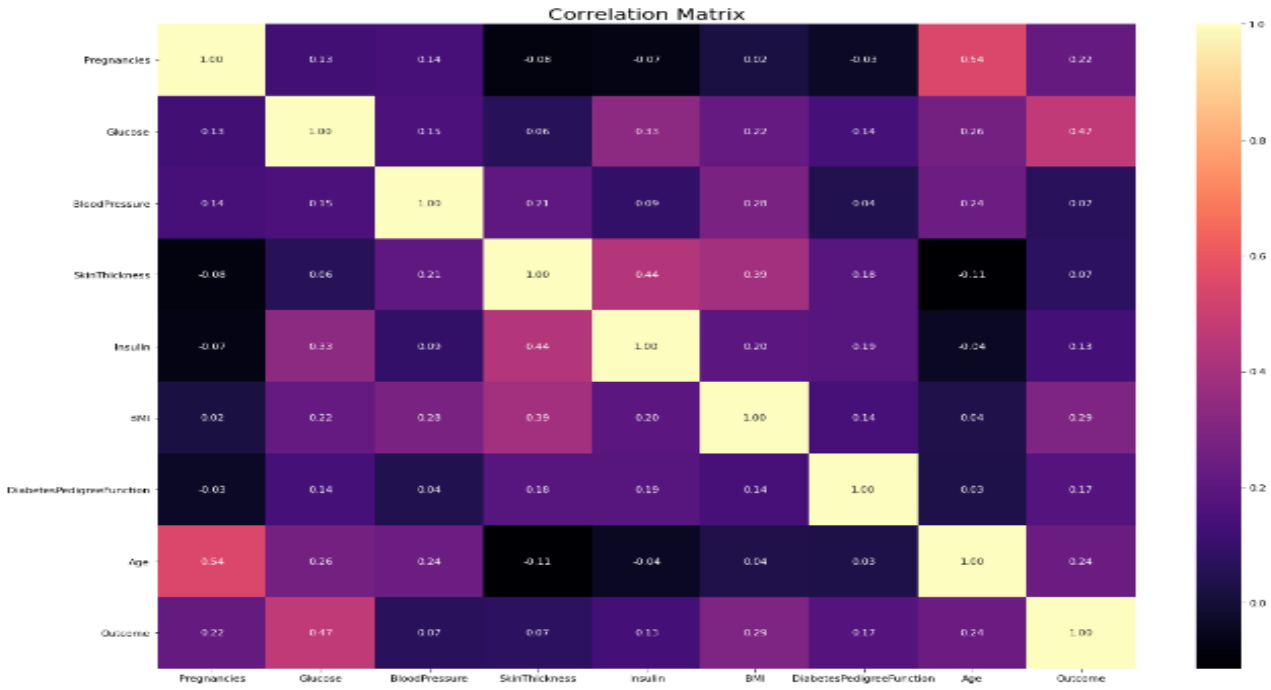

Figure 3. Correlation Matrix

There is clear evidence that no individual characteristic has a strong correlation with our outcome variable. Some characteristics display some exhibit a positive association and others a negative one with the outcome value.

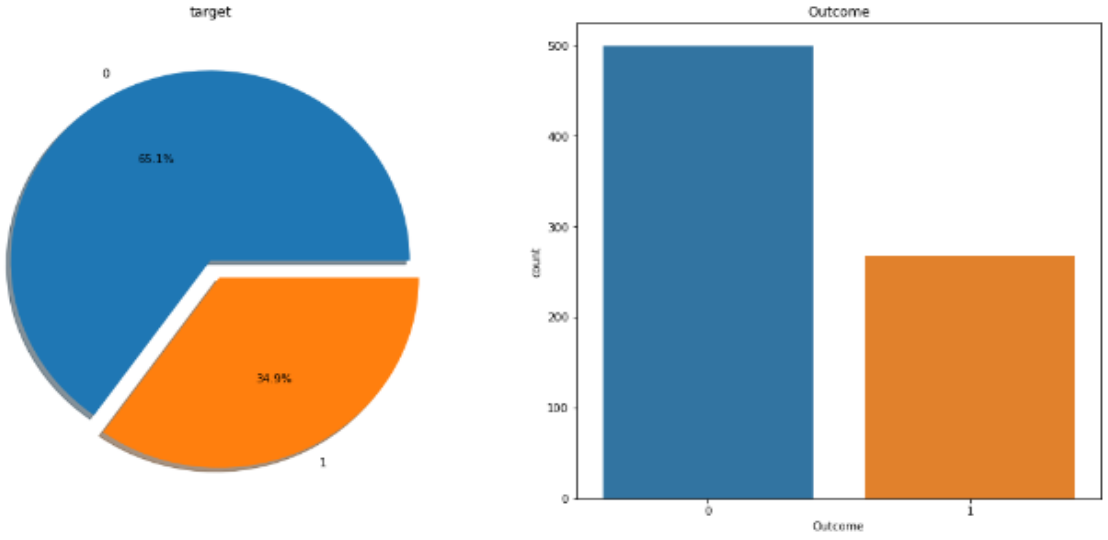

Figure 4. Predicted Results

The graph above illustrates the presence of a skew in the data, favouring data points with an outcome value of 0, indicating the absence of real diabetes. The number of non-diabetics is approximately double the number of diabetic individuals.

4.4 Machine Learning Classifiers

4.4.1 k-Nearest Neighbors

The k-NN technique is often considered to be the simplest straightforward machine learning algorithm. The sole stage in constructing the model is to save the training data set. The method identifies the nearest neighbours in the training data set that are closest to a new data point in order to make predictions, First, let's see whether we can validate the relationship between a model's accuracy and complexity. In the above picture, the n_neighbors parameter is displayed on the x-axis, while the accuracy of the training and test sets is shown on the y-axis.

4.4.2 Logistic regression

Logistic regression is one of the most commonly utilised algorithms for categorization. The default setting of C=1 yields a training set accuracy of 77% and a test set accuracy of 78% in the first row. When the value of C is set to 0.01, the second row achieves an accuracy of 78% on both the training and test sets. When it is applied, the accuracy on the training set decreases significantly, but the accuracy on the test set increases slightly.

4.4.3 Decision Tree

To give class values to each data point, this classifier makes use of a decision trees as DT. We can choose the maximum number of features here, that the model will consider.Importance of Decision Trees: Feature importance is a measure of the value of each feature in influencing the decision made by a tree. Each attribute is assigned a value ranging from 0 or 1, with 0 indicating no usage and 1 indicating a flawless prediction of target. This classifier promotes the concept of decision trees. It generates a collection of trees by randomly selecting features from the whole set of features that make up each tree.

4.4.4 Support Vector Machine(SVM)

There are several kernels from which to choose the hyperplane. We used the sigmoid, polynomial, radial basis function (RBF), and linear kernels in our experiments [29]. When comparing our approach to the current approaches, there are a few distinguishing characteristics that set it apart. A person's age, gender, height, weight, degree of physical activity, presence of hypertension, and other critical

characteristics are utilized to determine whether or not they have diabetes. Nonetheless, the majority of current methods make use of numerous features. For instance, [30] originally employed 123 characteristics in their study to predict diabetes. They remained significantly more feature-rich even after a great deal of laboratory testing was removed (the precise amount is unknown). Encountering all three characteristics in actual data is uncommon. Thus, we propose a mechanism that enables us to ascertain an individual's diabetes status just by considering a restricted set of characteristics. In the current inquiry, the technique utilized to ascertain each feature's contribution through feature importance is critical. In order to identify these fundamental patterns in the data, a correlation analysis is usually conducted immediately, as demonstrated in [31]. Thus, to carry out feature elimination, we had to implement a particular transformation that encompassed distance evaluation and clustering. Although correlation was not utilised for the prediction challenge, we can still derive insights from the ranking correlations discovered during the later stages of our model's development [32].

5. Conclusion and Future work

An important health concern that frequently arises in reality is the early detection of diabetes. This work employs a systematic approach to the development of a predictive system for diabetes. This study evaluates multiple machine learning categorization methods and assesses their performance using various measures. Investigation is carried out utilising the Diabetes Database. Results of a scientific investigation The XGBoost hyperparameter tweaking developed the model with the lowest value for the Cross Validation Score.. The value is 0.90. In addition, we identified several supplementary factors, such as body weight, levels of physical activity, and hypertension, that were indirectly associated with the prediction of diabetes. Furthermore, an association was found between the LDL/HDL measurement and diabetes. Conducting quick preliminary tests for diabetes, enhancing public knowledge and teaching for healthy lifestyles, and reducing government expenditures may all happen at the same time as a decrease in the significant strain that diabetes imposes on hospitals. Anticipating the progression of diabetes will enable the implementation of necessary measures to prevent millions of persons from receiving inadequate treatment because of few resources and insufficient awareness. This can have a positive impact on the healthcare system, as well as enhance people's quality of life. In the future, The system created using categorization methods derived from machine learning could enable the prediction or diagnosis a greater number of diseases. The technology can be further developed and improved to automate the analysis of diabetes, by combining more machine learning techniques.

References

1. G. Atlas, "Diabetes. International diabetes federation," IDF Diabetes Atlas, International Diabetes Federation, Brussels, Belgium, 10th edition, 2021.
2. S. Akhtar, J. A. Nasir, A. Sarwar et al., "Prevalence of diabetes and pre-diabetes in Bangladesh: a systematic review and metaanalysis," BMJ Open, vol. 10, no. 9, Article ID e036086, 2020.
3. R. Krishnamoorthi, S. Joshi, H. Z. Almarzouki et al., "A novel diabetes healthcare disease prediction framework using machine learning techniques," Journal of Healthcare Engineering, vol. 2022, Article ID 1684017, 10 pages, 2022.
4. F. A. Khan, K. Zeb, M. Al-Rakhami, A. Derhab, and S. A. C. Bukhari, "Detection and prediction of diabetes using data mining: a comprehensive review," IEEE Access, vol. 9, pp. 43711–43735, 2021.
5. K. J. Rani, "Diabetes prediction using machine learning," International Journal of Scientifc Research in Computer Science Engineering and Information Technology, vol. 6, pp. 294–305, 2020.
6. M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of diferent machine learning classifers," IEEE Access, vol. 8, pp. 76516–76531, 2020.
7. P. Saeedi, I. Petersohn, P. Salpea et al., "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 Research and Clinical Practice, vol. 157, article 107843, 2019.
8. D. R. Whiting, L. Guariguata, C. Weil, and J. Shaw, "IDF Diabetes Atlas: global estimates of the prevalence of diabetes for 2011 and 2030," Diabetes Research and Clinical Practice, vol. 94, no. 3, pp. 311–321, 2011.
9. Y. Cao, W. Li, G. Yang, Y. Liu, and X. Li, "Diabetes and hypertension have become leading causes of CKD in Chinese elderly patients: a comparison between 1990–1991 and 2009–2010," International Urology and Nephrology, vol. 44, no. 4, pp. 1269–1276, 2012.
10. S. Thomas and J. Karalliedde, "Diabetic kidney disease," Medicine, vol. 50, no. 11, pp. 704–710, 2022.
11. C. B. Newgard, "Metabolomics and metabolic diseases: where do we stand?," Cell Metabolism, vol. 25, no. 1, pp. 43–56, 2017.
12. G. Wu, "Amino acids: metabolism, functions, and nutrition," Amino Acids, vol. 37, no. 1, pp. 1–17, 2009.
13. J. Bene, M. Márton, M. Mohás et al., "Similarities in serum acylcarnitine patterns in type 1 and type 2 diabetes mellitus and in metabolic syndrome," Annals of Nutrition & Metabolism, vol. 62, no. 1, pp. 80–85, 2013.
14. Z. Sabouri, Y. Maleh, and N. Gherabi, "Benchmarking Classification Algorithms for Measuring the Performance on Maintainable Applications," in Advances in Information, Communication and Cybersecurity, Cham, 2022, pp. 173–179. doi: 10.1007/978-3-030-91738-8 17.
15. H. EL Massari, S. Mhammedi, Z. Sabouri, and N. Gherabi, "OntologyBased Machine Learning to Predict Diabetes Patients," in Advances in Information, Communication and Cybersecurity, Cham, 2022, pp. 437–445. doi: 10.1007/978-3-030-91738-8 40.
16. F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," Mater. Today Proc., Jul. 2021, doi: 10 .1016/j.matpr.2021.07.196.
17. J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432–439, Dec. 2021, doi: 10.1016/j.icte.2021.02.004.
18. P. Cıhan and H. Cos¸kun, "Performance Comparison of Machine Learning Models for Diabetes Prediction," in 2021 29th Signal Processing and Communications Applications Conference (SIU), Jun. 2021, pp. 1–4. doi: 10.1109/SIU53274.2021.9477824.
19. M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," in 2018 24th International Conference on Automation and Computing (ICAC), Sep. 2018, pp. 1–6. doi: 10.23919/IConAC.2018.8748992.
20. A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," Procedia Comput. Sci., vol. 165, pp. 292–299, Jan. 2019, doi: 10.1016/j.procs.2020.01.047.

21. M. Rady, K. Moussa, M. Mostafa, A. Elbasry, Z. Ezzat, and W. Medhat, "Diabetes Prediction Using Machine Learning: A Comparative Study," 334 H. El Massari et al. in 2021 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES), Oct. 2021, pp. 279–282. doi: 10.1109/NILES53778.202 1.9600091.
22. M. U. Emon, M. S. Keya, Md. S. Kaiser, Md. A. islam, T. Tanha, and Md. S. Zulfiker, "Primary Stage of Diabetes Prediction using Machine Learning Approaches," in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Mar. 2021, pp. 364–367. doi: 10.1109/ICAIS50930.2021.9395968.
23. T. Mahboob Alam et al., "A model for early prediction of diabetes," Inform. Med. Unlocked, vol. 16, p. 100204, Jan. 2019, doi: 10.1016/j. imu.2019.100204.
24. N. Yuvaraj and K. R. SriPreethaa, "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster," Clust. Comput., vol. 22, no. 1, pp. 1–9, Jan. 2019, doi: 10.1007/s10586-017-1 532-x.
25. G. Tripathi and R. Kumar, "Early Prediction of Diabetes Mellitus Using Machine Learning," in 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Jun. 2020, pp. 1009–1014. doi: 10.1109/ICRITO48877.2020 .9197832.
26. Y. Jian, M. Pasquier, A. Sagahyroon, and F. Aloul, "A Machine Learning Approach to Predicting Diabetes Complications," Healthcare, vol. 9, no. 12, Art. no. 12, Dec. 2021, doi: 10.3390/healthcare9121712.
27. S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of Prediction Accuracy of Diabetes Using Classifier and Hybrid Machine Learning Techniques," in Intelligent and Cloud Computing, Singapore, 2021, pp. 399–409. doi: 10.1007/978-981-15-6202-0 41.
28. K. Pavani, P. Anjaiah, N. V. Krishna Rao, Y. Deepthi, D. Noel, and V. Lokesh, "Diabetes Prediction Using Machine Learning Techniques: A Comparative Analysis," in Energy Systems, Drives and Automations, Singapore, 2020, pp. 419–428. doi: 10.1007/978-981-15-5089-8 41.
29. Chaki J, Ganesh ST, Cidham SK,. Machine learning and artificial intelligence based diabetes mellitus detection and self-management: a systematic review. Journal of King Saud 2020;
30. Ho-Pham, L.T.; Nguyen, U.D.; Tran, T.X.; Nguyen, T.V. Discordance in the diagnosis of diabetes: Comparison between HbA1c and fasting plasma glucose. PLoS ONE 2017, 12, e0182192.
31. Dinh, A.; Miertschin, S.; Young, A.; Mohanty, S. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med. Inform. Decis. Mak. 2019, 19.
32. Rodríguez-Rodríguez, I.; Rodríguez, J.V.; González-Vidal, A.; Zamora, M.Á. Feature Selection for Blood Glucose Level Prediction in Type 1 Diabetes Mellitus by Using the Sequential Input Selection Algorithm (SISAL). Symmetry 2019, 11, 1164
33. G. Mingrone, "Carnitine in type 2 diabetes," Annals of the New York Academy of Sciences, vol. 1033, no. 1, pp. 99–107, 2004.
34. D. Bzdok, N. Altman, and M. Krzywinski, "Statistics versus machine learning," Nature Methods, vol. 15, no. 4, pp. 233- 234, 2018.
35. Mahboob, K.; Ali, S.A.; Laila, U. Investigating learning outcomes in engineering education with data mining. Comput. Appl. Eng. Educ. 2020, 28, 1652–1670. [CrossRef]
36. UCIMachineLearning Repository: Early-Stage Diabetes Risk Prediction Dataset. Available online: https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset (accessed on 5 July 2021).
37. Patil, S.; Rajeswari, K.; Abin, D. Preprocessing and Classification in WEKA Using Different Classifiers. 2014. Available online: https://www.semanticscholar.org/paper/Preprocessing-and-Classification-in-WEKA-Using-Patil-Rajeswari/07899afe3 0164eea03245a5f05c4b47c1a50bfa7 (accessed on 29 June 2022).
38. Khan,S.U.; Khan, A.W.; Khan, F.; Khan, M.A.; Whangbo, T. Critical Success factors of Component-Based Software Outsourcing Development from Vendors' Perspective: A Systematic Literature Review. IEEE Access 2021, 10, 1650–1658. [CrossRef]
39. Wahome,P.; Bongo, W.; Maina, D.R.R. Towards Effective Data Preprocessing for Classification Using WEKA. Int. J. Sci. Res. 2016, 5, 1210–1214. [CrossRef]

40. Prema, N.S.; Varshith, V.; Yogeswar, J. Prediction of diabetes using ensemble techniques. Int. J. Recent Technol. Eng. 2022, 7, 203–205.
41. Weka 3: Machine Learning Software in Java-Weka 3—Data Mining with Open-Source Machine Learning Software in Java. Available online: https://www.cs.waikato.ac.nz/ml/weka/ (accessed on 11 January 2022).
42. Yang, H.; Luo, Y.; Ren, X.; Wu, M.; He, X.; Peng, B.; Deng, K.; Yan, D.; Tang, H.; Lin, H. Risk Prediction of Diabetes: Big data mining with fusion of multifarious physical examination indicators. Inf. Fusion 2021, 75, 140–149.