Journal of Distributed Computing and Systems(JDCS)
Vol.9, Issue 1, 2026, Pages 64-72

Predicting diabetes using machine learning with gradient boosting algorithms XGBoost and LightGBM and web-based data

Maryam Hajiee, Seyed Mohamadreza Soltanzadeh Solhdar
Department of Computer Engineering, ST.C., Islamic Azad University, Tehran, IRAN
Department of Computer Engineering, ST.C., Islamic Azad University, Tehran, IRAN

Article History:
Received: 25 January 2026
Received in revised form: 27 March 2026
Accepted: 31 May 2026
Available online: 15 August 2026

Abstract

Diabetes is a condition characterized by elevated blood glucose levels, which should not be overlooked as it can lead to severe complications and organ damage if left untreated. Effective management of diabetes relies on early prediction. Traditional diagnostic methods can be both financially and temporally burdensome for patients and may also be prone to inaccuracies. To address these challenges, modern approaches are proposed. In this project, we aim to achieve early prediction of diabetes in humans by employing various machine learning techniques. Two classification models, XGBoost and LightGBM, were developed based on the Pima Indians Diabetes Database, with hyperparameter tuning to optimize their performance. Both models were rigorously evaluated using various metrics, including Accuracy, Precision, Recall, F1-score, and False Negative Rate. Ultimately, this study achieved an accuracy of 83% with the LightGBM algorithm. After comparing the models, we observed that both demonstrated considerable accuracy in distinguishing diabetic patients, but the LightGBM model exhibited superior performance compared to XGBoost. Based on these findings, the LightGBM model was selected for the development of a web application prototype using the Django framework, which was subsequently deployed for pilot testing, yielding satisfactory results. Our future endeavors will focus on increasing the model's accuracy and completing the necessary steps for its real-world implementation.

Keywords: Machine Learning, Diabetes Prediction, Data Collection, Web, IQR, XGBoost, LightGBM

I. INTRODUCTION

Diabetes mellitus, a chronic metabolic condition frequently referred to as just diabetes, is characterized by elevated blood glucose levels. [1]. Defects in the synthesis or function of insulin cause this disorder. By secreting insulin, pancreatic beta cells [2], enables the body's cells to absorb glucose from the bloodstream for energy utilization. Untreated or poorly managed diabetes can lead to severe consequences, affecting a wide range of health aspects. From vision and kidney function to cardiovascular and nervous system health, diabetes poses a significant threat to individual well-being [3]. Moreover, the increasing prevalence of diabetes in recent years has raised serious concerns and triggered alarms within the World Health Organization [4]. About 536.6 million people between the ages of 20 and 79 had a diabetes diagnosis in 2021, and the International Diabetes Federation (IDF) projects this number to reach 694 million by 2045 [5]. Furthermore, another study predicts a global population increase of 25% and 51% by 2030 and 2045, respectively [6], suggesting that a substantial portion of the world's population is at a heightened risk of developing this disease.

The widespread prevalence of diabetes presents numerous challenges to humanity. Every year, 6.7 million people globally pass away from diabetes or its complications, according to the IDF. A death occurs every five seconds due to diabetes, and this number continues to rise [7]. From severe health implications to substantial economic burdens, diabetes has become a major global health crisis [8]. Healthcare professionals, researchers, and pharmaceutical companies are continuously striving to improve diabetes management and develop effective treatments. However, diabetes still incurs significant financial consequences, including medical expenses and reduced productivity. Therefore, finding innovative solutions to this health issue is crucial.

With remarkable advancements in machine learning, the advancement of powerful predictive models for diabetes diagnosis has become feasible. However, the quantity and quality of training data have a significant impact on these models' performance. A key challenge in this field is the acquisition of comprehensive and diverse datasets that accurately represent the diabetic patient population. The web, as a rich source of structured and unstructured data,

offers significant potential for collecting diabetes-related data. Electronic health records, research databases, patient websites, wearable devices, and social networks can all serve as valuable resources for extracting diabetes-related information.

Machine learning techniques, a branch of artificial intelligence, have significantly impacted progress in various fields, especially medicine [9]. With increasing access to health data and the availability of diabetes-related data [10], large datasets can be analyzed with machine learning algorithms to provide more accurate predictions and diagnose and manage diabetes [11]. These algorithms can be used to predict blood glucose levels to prevent harmful fluctuations, identify individuals at risk of developing diabetes, and personalize treatment plans. This paper explores the challenges and opportunities of web-based dataset acquisition for machine learning applications in diabetes prediction. Subsequently, with this objective, machine learning models are effectively employed for diabetes analysis. In this study, two different models for machine learning, XGBoost and LightGBM, are utilized. The experiments are conducted using the Indian Diabetes dataset 2 obtained from the Kaggle website [12]. The results are evaluated using various validation metrics such as accuracy, precision, recall, and the F1-score, and the two models are compared. Subsequently, the best-performing model is selected and presented. Finally, for the model demonstrating superior performance, a user-friendly web application version was developed using the Django framework.

This review is organized as follows: Section 2 provides background on diabetes, machine learning, and web and data. Section 3 discusses the approach used in this review by reviewing previous studies. Section 4 explains the proposed methodology in detail. Section 5 discusses the observations and results of this paper. Section 6 concludes the review.

II. BACKGROUND

A. Diabetes Mellitus

Primarily, three main types of diabetes mellitus can be identified: type 1 diabetes, type 2 diabetes, and gestational diabetes [13]. Type 1 diabetes is an autoimmune disease. In this condition, the immune system of the body targets the pancreatic cells that produce insulin, impairing the body's ability to produce insulin. Consequently, the absence of insulin, the hormone responsible for transporting glucose to the body's cells, leads to elevated blood glucose levels [14]. Type 2 diabetes is the most common form of diabetes and is associated with lifestyle factors [15], dietary habits [16], overweight and obesity, sedentary behaviors, lack of physical activity, and mental health [17]. In type 2 diabetes, insulin is present in the body but does not function effectively, leading to insulin resistance [18]. Consequently, the body's ability to regulate blood glucose levels is compromised, and individuals with type 2 diabetes require various treatments and lifestyle modifications to maintain proper management [19]. Gestational diabetes typically occurs during the later stages of pregnancy and is primarily associated with hormonal changes. Hormones produced by the placenta specifically affect insulin, impairing its function [20].

B. Machine Learning

Machine learning encompasses numerous algorithms, which are broadly categorized into three types: supervised learning, unsupervised learning, and semi-supervised learning [21]. Supervised learning algorithms are employed for constructing predictive models [22]. A predictive model forecasts missing values based on other available values within the dataset. A supervised learning algorithm uses a group of input data along with their matching outputs to create a model that can predict accurate answers for new sets of data. Unsupervised learning methods are used to develop descriptive models. In this approach, a defined set of inputs exists, but the outputs are unknown. Semi-supervised learning utilizes both labeled and unlabeled data within the training dataset. These are prominent techniques in machine learning [23].

C. Web and Data

Data is very important because it helps make machine learning models more accurate. so much so that it is often referred to as the “fuel” of these models. These data not only serve as input for training models [24] but also influence all stages of model development and evaluation. Given that this project focuses on diabetes prediction, having sufficient high-quality data is crucial for developing accurate predictive models for this disease. The web, as a vast and readily accessible source of information, plays a significant role in data collection [25] for various research endeavors, including health research and disease prediction. However, it presents challenges such as data quality, data volume, data dynamism, and legal and ethical considerations, which researchers must carefully address.

III. LITERATURE REVIEW

Substantial research efforts have been dedicated to the field of diabetes prediction, yielding notable advancements. These studies have employed diverse models and datasets, while introducing unique methodologies and innovations, ultimately achieving varying degrees of accuracy. In this

context, we will reference a selection of relevant research papers.

A method for diagnosing type 2 diabetic mellitus (T2DM) early is required since diabetes mellitus is a common chronic illness worldwide. Diabetes diagnosis has been aided by a number of machine learning and data mining approaches, such as ANN, SVM, KNN, decision trees, and Extreme Learning Machines. As a result, they provide Deep Learning, a branch of machine learning that uses effective data processing methods to manage smaller datasets. They develop a binary classification model to accurately predict if a person has diabetes or not using Keras and Theano as backends. The goal of their research is to improve diabetes diagnosis accuracy and dependability, which will ultimately aid in better healthcare decision-making [26]. Diabetes prediction is an active research topic where medical professionals are attempting to envision the issue more accurately. In order to diagnose diabetes in its early stages, this study examines multilayer perceptron classifiers, decision trees, K-nearest neighbors, and random forests. It also applies certain feature selection strategies to the classifiers. After preprocessing the raw data to eliminate outliers and impute missing values to the mean, hyperparameter optimization is carried out. Weka 3.9 was used to conduct experiments on the PIMA Indian diabetes dataset. The results showed that the accuracy of multilayer perceptron was 77.60%, that of decision trees was 76.07%, that of K-nearest neighbors was 78.58%, and that of random forests was 79.8%, the highest accuracy for random forest classifiers to date [27]. Diabetes is already one of the most common diseases in the world, and its prevalence is rising. However, if diabetes is detected early, its course can be considerably slowed. This research proposes a unique deep learning-based method for early diabetes identification. The PIMA dataset used in the study is composed entirely of numerical values, just as other medical data sets. Convolutional neural network (CNN) models that are commonly utilized have limited application to this kind of data. This work uses CNN models' strong representation to convert numerical data into visuals based on feature relevance for the purpose of early diabetes identification. The generated diabetes image data is then classified using three different approaches. In the first, convolutional neural network (CNN) models called ResNet18 and ResNet50 are used to sort photos of people with diabetes. Second, we use support vector machines (SVMs) to aggregate the deep features of the ResNet models for classification. The final method uses SVM to classify the chosen fusion features. [28]. Diabetes mellitus is a chronic illness that changes how proteins, lipids, and carbs are normally metabolized. Elevated blood sugar levels, referred to as hyperglycemia in medicine, are a hallmark of all forms of diabetes. The two primary components of the methodology presented in this work are normalization in the second portion and outlier removal, which is followed by jitter in the first. Principal component analysis (PCA) is the method used in this paper to analyze the feature component. At the conclusion of the classification process on the PIMA dataset, several classifiers, such as RF, NB, and DT, were employed in addition to SVM. The results of the experiment demonstrated that even though 80% of the data was used for training, employing the classifier in conjunction with the feature selection technique produced the best accuracy rate of 89.90% [29]. Diabetes is a dangerous medical disorder characterized by persistently elevated blood sugar levels. People worldwide are afflicted by this sickness, which is spreading quickly. Early disease identification is essential for successful therapy. In this work, a system that can be used by end users to identify the existence of diabetes utilizing a variety of health markers is proposed. For this system, they have created a predictive classifier using artificial neural networks. ANN is used because feature extraction is automatically carried out throughout the ANN model's training, negating the need for explicit use. This leads to a considerable decrease in training time, which is especially noticeable when using larger datasets. The model's classification accuracy of 0.8079 indicates that reducing training time has little or no impact on classification performance compared to similar studies using feature extraction. [30]. Using the Pima Indians Diabetes Database, this work investigates the use of machine learning (ML) models for diabetes prediction with an emphasis on improving model interpretability by applying Shapley Additive explanations (SHAP). Using test/train split and 10-fold cross-validation methods, the study assesses eight machine learning models: AdaBoost, k-NN, LR, MLP, NB, RF, SVM, and XGBoost. In this study, the RF model demonstrated remarkable performance. Insulin levels, glucose, BMI, and pregnancies are the main determinants in diabetes prediction, according to SHAP analysis, which was used to determine the most significant predictors. These findings are consistent with established clinical markers. Robust model performance is further enhanced by the application of the SMOTE for data scaling and balancing the classes. The work highlights the need for interpretable machine learning in healthcare and suggests SHAP as a useful tool for bridging clinical transparency and prediction accuracy in diabetes diagnosis [31].

IV. METHODOLOGY

This paper utilizes the supervised gradient boosting algorithms XGBoost and LightGBM to construct a diabetes prediction model. The primary objective of this research is to compare these two algorithms in the context of diabetes prediction, thereby investigating which method achieves superior results and higher accuracy.

The data used in this research comes from the Pima Indians Diabetes Database, which is also available online through Kaggle and is released by the National Institute of Diabetes and Digestive and Kidney Diseases. It comprises data from 768 patients, categorized into diabetic and non-diabetic groups. This dataset consists of nine features, the first eight of which are as follows: 1. The quantity of pregnancies, 2. The level of plasma glucose, 3. Blood pressure diastolic, 4. The thickness of the triceps skin folds, 5. Serum insulin after two hours, 6. BMI, or body mass index, 7. Pedigree function for diabetes, 8. Age. The features of this dataset are presented in Table I.

Table I: Features of Pima Dataset

#	**ATTRIBUTES**
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin Thickness
5	Insulin
6	BMI
7	Diabetes Pedigree Function
8	Age

The ninth feature, the dependent variable, represents the class label of each data point. This feature indicates a binary outcome of 0 or 1 for diabetic patients, denoting a negative or positive diagnosis for diabetes, respectively. A value of 0 signifies the absence of diabetes, while a value of 1 indicates the presence of the disease in the individual. The operational workflow of the project is in Figure 1.

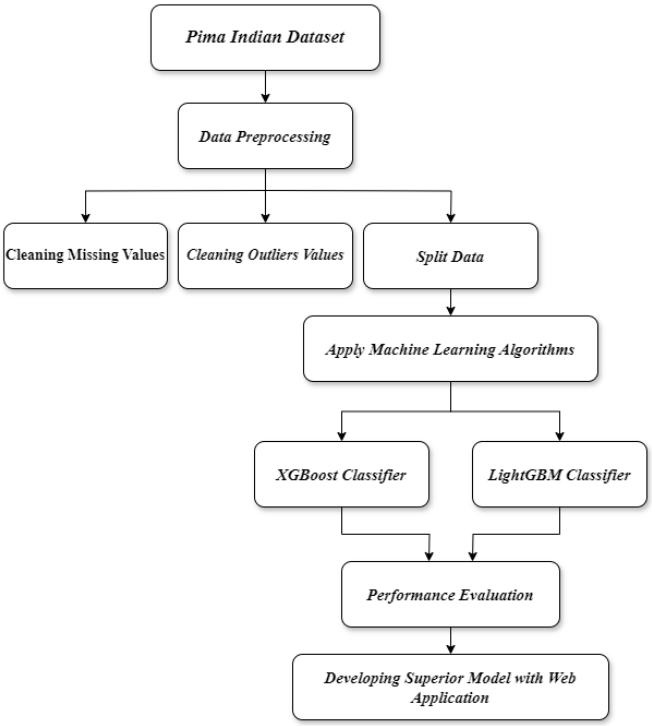

Fig 1: Overall Methodology

In this study, data preprocessing was initially performed. It is well-established that machine learning models are exclusively applicable to numerical data [32]. Therefore, as the first step of preprocessing, it was examined whether all data were in numerical format. Fortunately, this dataset contained no non-numerical data, with all data points being of numerical type. Subsequently, missing values were filled in, unusual data points were found and removed, the data was scaled to a standard format, and then the data was normalized. Finally, the dataset was split into training and testing portions using the train-test split method [33]. In the next step, the XGBoost method was used to train and assess the model. The model's performance was assessed using the evaluation metrics of accuracy, precision, recall, and F1-score. Subsequently, the model was trained using the LightGBM algorithm, and the same measurements were used to reevaluate its performance. Ultimately, the two models were compared. It is worth noting that during the construction of both models, hyperparameters were meticulously examined and incorporated into the model building process. The utilized hyperparameters had a significant impact on the performance results of both models, leading to performance enhancement. Finally, with the aim of deploying our project as a practical and user-interactive application, the model exhibiting the highest performance and accuracy was selected and implemented as a web application utilizing the Django framework.

Journal of Distributed Computing and Systems(JDCS)
Vol.9, Issue 1, 2026, Pages 64-72

In the data processing stage, specifically within the outlier handling section of the dataset, outliers are typically identified and subsequently removed, effectively discarding these data points from the dataset under consideration. However, an innovative approach employed in this paper was to avoid removing any outlier values for model development. This decision stems from the understanding that discarding a portion of the data can disrupt model performance and diminish model validity. Furthermore, a data point identified as an outlier may contain crucial information pertaining to other model features; its removal, therefore, would result in the loss of valuable information that could have been utilized in the model. Consequently, a specific technique was adopted in this stage. The employed technique involves utilizing the Interquartile Range (IQR) method for outlier identification [34]. In the IQR method, there are three quartiles: Q1, Q2, and Q3. The first quartile, or Q1, represents the value below which 25% of the data falls, and above which 75% of the data falls. Q2, the second quartile, denotes the median of the data. The value that divides the bottom and upper 50% of the data is known as the median. Q3 is the value below which 75% of the data falls, and above which 25% of the data falls. In many instances, the IQR can be utilized for analyzing data dispersion. The IQR effectively eliminates outliers below Q1 and above Q3, thereby providing a more realistic assessment of data spread. The innovation employed in this study involved replacing outliers below Q1 in a specific feature with the last acceptable value of Q1, and similarly, replacing outliers above Q3 with the last acceptable value of Q3. Figure 2 illustrates the dataset prior to outlier handling, while Figure 3 depicts the dataset after outlier handling and value replacement.

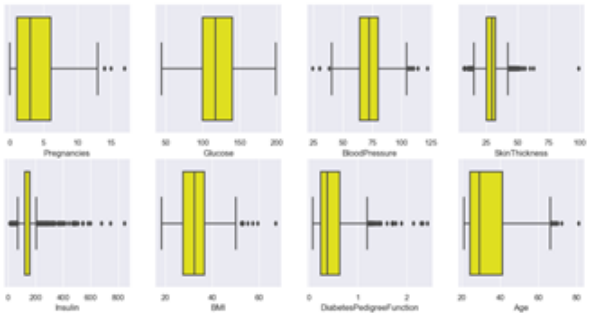

Fig 2: With Outliers

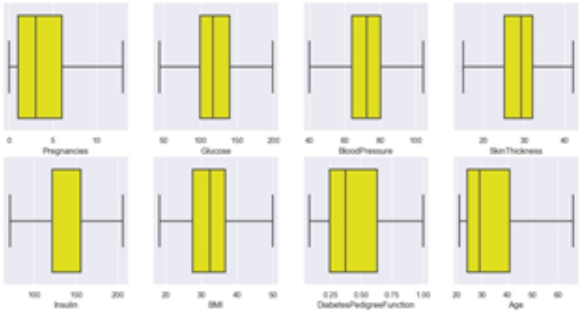

Fig 3: Without Outliers

V. RESULTS AND DISCUSSION

In this study, the dataset was initially examined, revealing a total of 768 data points. Among these, 500 data points had a result class of zero, indicating that these individuals were healthy. The remaining 268 data points had a result class of one, signifying that 268 individuals in this dataset had diabetes. The corresponding distribution is illustrated in Figure 4.

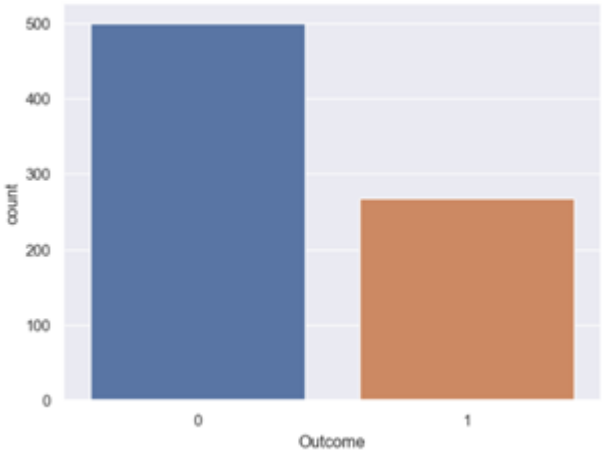

Fig 4: The quantity of the dataset's two features

The Subsequently, dataset preprocessing commenced. In this stage, an initial examination was conducted to ascertain the presence of missing values or meaningless zero values within the dataset; these values were subsequently imputed using the mean value of each respective feature. The next step addressed outliers. Outliers were identified using the IQR method and replaced with the lowest/highest acceptable values within the first/third quartiles, respectively. Following the handling of missing values and outliers in the diabetes dataset, data standardization and normalization were performed to constrain all data within a defined range. The data were then split up at random for the purpose of training and assessing the model. The model was trained using 80% of the data, while the remaining portion was retained for testing. Hyper parameters were meticulously examined and

employed in model construction using the XGBoost and LightGBM algorithms. These two algorithms were compared using five distinct metrics: Accuracy, Precision, Recall, F1-score, and False Negative Rate. The results are presented in Table II.

Table II: Diabetes prediction results using XGBoost and LightGBM models

Model	Accuracy	Precision	Recall	F1-Score	FNR
XGBoost	81.81	71.11	68.08	69.56	0.31
LightGBM	83.11	73.33	70.21	71.73	0.29

We utilized the accuracy ratio, which is the proportion of properly predicted cases out of all instances, as specified by Equation (1) to assess the model's overall accuracy. The LightGBM model achieved a higher accuracy with a score of 83.11% for the test set. Precision is one of the important metrics for model evaluation. The precision calculated using Equation (2) indicates the percentage of instances correctly labeled as positive relative to all instances predicted as positive; the LightGBM model demonstrated superior precision with a score of 73.33%. Equation (3) evaluates recall, which Measures how well a model can find real positive cases while considering both false negatives and true positives. In this metric as well, the LightGBM model exhibited a better score with 70.21%. The weighted average (harmonic mean) of precision and recall is the F1 score, which may be obtained by combining the precision and recall measurements. The F1-score, derived from the formula in Equation (4), establishes a balance between the amount of positives a classifier recall and its precision in identifying them; in this metric too, the LightGBM algorithm demonstrated better performance. Finally, the false negative rate of each model was calculated using Equation (5). A false negative error, or Type II error, refers to an error where the test result is incorrectly reported as negative; for instance, if we were to provide an example, we could say that an individual was ill, but the model predicted them as healthy. This error can incur significant costs, both financial and in terms of human life, for the patient. As can be observed in Table II, the rate of this error is lower for the LightGBM model compared to the XGBoost model. It is important to note that the LightGBM model demonstrated considerable accuracy in this research, indicating its adaptability and effectiveness in predicting diabetes diagnosis. Equations (1-5) can be found below:

(1) Accuracy = (TP+TN)/(TP+TN+FP+FN)
(2) Precision = TP/(TP+FP)
(3) Recall = TP/(TP+FN)
(4) F1 score = 2TP/(2TP+FP+FN)
(5) FNR = FN/(FN+TP)

All Equations (1-5) are derived from the confusion matrix [35]. It includes True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). A True Positive signifies that an individual is ill, and the model correctly predicts this. A False Positive signifies that an individual is not ill, but the model incorrectly predicts them as ill. A True Negative means that an individual is not ill, and the model correctly predicts them as healthy. A False Negative signifies that an individual is ill, but the model incorrectly predicts them as healthy. Based on the results obtained from the confusion matrix of this research, it was observed that the LightGBM model recorded the lowest number of False Negatives (FN) compared to the XGB model in the test set. This indicates that the LightGBM model is especially good at reducing the possibility that people with diabetes won't be identified. The confusion matrix results for these two models can be seen in Figure 5.

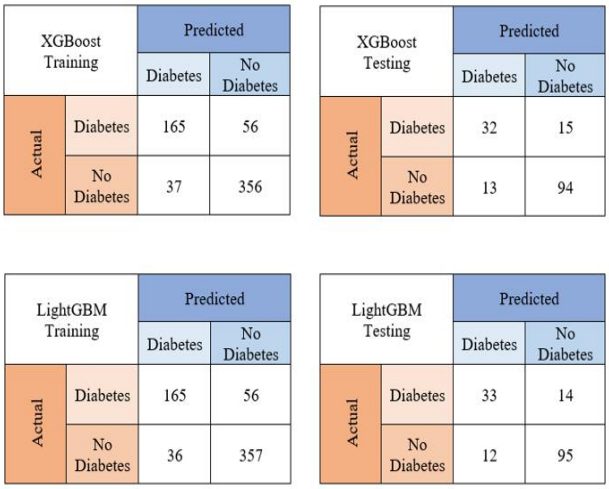

Fig 5: Confusion matrix results

Considering the findings from this investigation, it was determined that the LightGBM model demonstrated superior performance. Consequently, this model was selected, and with the aim of enhancing the interactive nature of such research with users, a web application version of this project was developed using the Django framework. Figure 6 illustrates an image of a novel interactive interface for diabetes prediction.

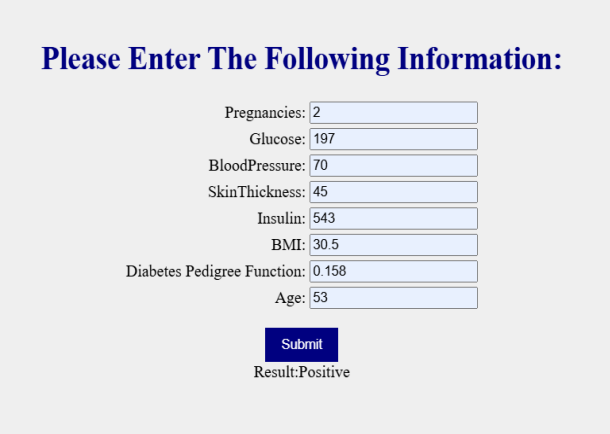

Fig 6: The new interface for diabetes prediction

There have been significant advances in diabetes prediction using machine learning on the Pima Indian diabetes database. Some of these advances are discussed in Section 3. Table III lists other papers that work on the Pima dataset but with different models. However, a notable difference between these papers and our paper is that they have completely removed outliers in the preprocessing section and also used different algorithms to train their models. A comparison between them is presented in Table III.

Table III: Comparative summary of recent PIDD studies for diabetes prediction

NO.	Authors	Year	Methods	Best Accuracy
1	Rahman, F., et al. [36]	2025	LightGBM, XGBoost	85.16%
2	Scholar, M., et al. [37]	2025	MLP	86.36%
3	Teimoory, G. K., et al. [38]	2025	Random Forest	88%
4	Vangoori, S., et al. [39]	2025	SVM, RF, AB, NN	89.90%
5	Shamal, M., et al. [40]	2024	SVM, NB, RF, DT	89.86%

VI. CONCLUSIONS

Nowadays, diabetes is one of the most dangerous illnesses, and its early diagnosis can significantly improve patient management outcomes. The primary objective of this project was to design and implement a diabetes prediction model using two machine learning methods and to enhance the accuracy of disease prediction, considering its importance as a major cause of death worldwide. This objective was successfully achieved. Eight distinct features (Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, family history of diabetes, and Age) were used in this study to preprocess the data. After both models were trained, tested, and compared, this study collected data to assess each model's performance. The outcomes showed that both models produced positive outcomes; however, the LightGBM model exhibited superior performance.

REFERENCES:

[1] *Singh, R., Gholipourmalekabadi, M., & Shafikhani, S. Animal models for type 1 and type 2 diabetes: advantages and limitations. Frontiers in Endocrinology. 2024; 15.*

[2] *Abdalla, M. Advancing diabetes management: Exploring pancreatic beta-cell restoration's potential and challenges. World Journal of Gastroenterology. 2024; 30.*

[3] *https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444*

[4] *Plebon-Huff, S., Haji-Mohamed, H., Gardiner, H., Ghanem, S., Koh, J., & LeBlanc, A. Contextualization of Diabetes: A Review of Reviews from Organisation for Economic Co-operation and Development (OECD) Countries. Current Diabetes Reports. 2025; 25.*

[5] *J.M. Lawrence, et al., Trends in prevalence of Type 1 and Type 2 diabetes in children and adolescents in the US, 2001-2017, JAMA 326 (8) (Aug 24 2021) 717–727, https://doi.org/10.1001/jama.2021.11165, in eng.*

[6] *International Diabetes Federation. Diabetes. Brussels: International Diabetes Federation; 2019.*

[7] *International Diabetes Federation. 2021. Available online: https://diabetesatlas.org/atlas/tenth-edition/ (accessed on 6 December 2021).*

[8] *Hatipoglu, B., & Pronovost, P. Role of Diabetes Self-management Education for Our Health Systems and Economy.. The Journal of clinical endocrinology and metabolism. 2025; 110 Supplement_2.*

[9] *S. Dev, H. Wang, C.S. Nwosu, N. Jain, B. Veeravalli, D. John, A predictive analytics approach for stroke prediction using machine learning and neural networks, Healthcare Anal. 2 (2022) 100032, https://doi.org/10.1016/j. health.2022.100032.*

[10] *Khokhar, P., Gravino, C., & Palomba, F. Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review. Artificial intelligence in medicine. 2025; 164.*

[11] *Khalifa, M., & Albadawy, M. Artificial Intelligence for Diabetes: Enhancing Prevention, Diagnosis, and Effective Management. Computer Methods and Programs in Biomedicine Update. 2024*

[12] *https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database*

[13] *Sweeting, A., Hannah, W., Backman, H., Catalano, P., Feghali, M., Herman, W., Hivert, M., Immanuel, J., Meek, C., Oppermann, M., Nolan, C., Ram, U., Schmidt, M., Simmons, D., Chivese, T., & Benhalima, K. Epidemiology and management of gestational diabetes. The Lancet. 2024; 404.*

[14] *Hm, Z., An, M., & Mm, S. Hyperglycemia, an abnormality that results from a breakdown of normal glucose control processes is also the result of molecular mechanisms to*

protect the insulin receptor? A hypothesis. Ibom Medical Journal. 2025

[15] *Gangani Dharmarathne a,b , Thilini N. Jayasinghe a,b , Madhusha Bogahawaththa c , D.P.P. Meddage c , Upaka Rathnayake, "A novel machine learning approach for diagnosing diabetes with a self-explainable interface", Healthcare Analytics, journal homepage: www.elsevier.com/locate/health, https://doi.org/10.1016/j.health.2024.100301, 13 January 2024*

[16] *Hye-Jin Kim 1 , Eunjeong Cho 2 and Gisoo Shin 2, Experiences of Changes in Eating Habits and Eating Behaviors of Women First Diagnosed with Gestational Diabetes, Int. J. Environ. Res. Public Health 2021, 18, 8774. https://doi.org/10.3390/ijerph18168774*

[17] *H. Li, et al., Genetic risk, adherence to a healthy lifestyle, and type 2 diabetes risk among 550,000 Chinese adults: results from 2 independent Asian cohorts, Am. J. Clin. Nutr. 111 (3) (2020) 698–707, https://doi.org/10.1093/ajcn/nqz310.*

[18] *Tan, S., Wong, J., Sim, Y., Wong, S., Elhassan, S., Tan, S., Lim, G., Tay, N., Annan, N., Bhattamisra, S., & Candasamy, M. Type 1 and 2 diabetes mellitus: A review on current treatment approach and gene therapy as potential intervention.. Diabetes & metabolic syndrome. 2019; 13 1.*

[19] *L.R. Saslow, et al., Psychological support strategies for adults with type 2 diabetes in a very low–carbohydrate web-based program: randomized controlled trial, JMIR Diabetes 8 (2023) e44295.*

[20] *P. Qian, et al., How breastfeeding behavior develops in women with gestational diabetes mellitus: a qualitative study based on health belief model in China, Front. Endocrinol. 13 (2022) 955484.*

[21] *Janiesch, C., Zschech, P., & Heinrich, K. Machine learning and deep learning. Electronic Markets. 2021; 31.*

[22] *Jiang, T., Gradus, J., & Rosellini, A. Supervised Machine Learning: A Brief Primer.. Behavior therapy. 2020; 51 5.*

[23] *https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained*

[24] *Altalhan, M., Algarni, A., & Alouane, T. Imbalanced Data Problem in Machine Learning: A Review. IEEE Access. 2025; 13.*

[25] *Sanjay Kumar Madria, Sourav S Bhowmick, W. -K Ng, E. P. Lim, Research Issues in Web Data Mining, Springer-Verlag Berlin Heidelberg 1999.*

[26] *Dey, A. Binary Classification of Diabetes in Pima Indian Dataset: A Deep Learning Perspective. International Journal for Research in Applied Science and Engineering Technology. 2023 https://doi.org/10.22214/ijraset.2023.57364.*

[27] *Saxena, R., Sharma, S., Gupta, M., & Sampada, G. A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods. Computational Intelligence and Neuroscience. 2022; 2022.*

[28] *G, A., Reddy, P., Budati, N., Rebecca, B., Perumalla, S., & Gujar, V. A Deep Learning Framework for Diabetes Prediction: PIMA Indian Dataset. 2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES). 2024*

[29] *Mishra, S., & Dash, S. Machine Learning Based Diabetes Prediction Using the PIMA Indian Dataset. 2024 2nd International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES). 2024*

[30] *Pyne, A., & Chakraborty, B. Artificial Neural Network based approach to Diabetes Prediction using Pima Indians Diabetes Dataset. 2023 International Conference on Control, Automation and Diagnosis (ICCAD). 2023*

[31] *Kirbaş, I., & Çifci, A. Leveraging SHAP for Interpretable Diabetes Prediction: A Study of Machine Learning Models on the Pima Indians Diabetes Dataset. Balkan Journal of Electrical and Computer Engineering. 2025*

[32] *Andrew Ng, book, Machine Learning Yearning, 2018*

[33] *https://scikit-learn.org/1.5/modules/generated/sklearn.model_selection.train_test_split.html*

[34] *Mramba, L., Liu, X., Lynch, K., Yang, J., Aronsson, C., Hummel, S., Norris, J., Virtanen, S., Hakola, L., Uusitalo, U., & Krischer, J. Detecting potential outliers in longitudinal data with time-dependent covariates. European Journal of Clinical Nutrition. 2024*

[35] *Erbani, J., Portier, P., Egyed-Zsigmond, E., & Nurbakova, D. Confusion Matrices: A Unified Theory. IEEE Access. 2024; 12.*

[36] *Rahman, F., Hossain, S., Tiang, J., & Nahid, A. (2025). Diabetes Prediction Using Feature Selection Algorithms and Boosting-Based Machine Learning Classifiers. Diagnostics, 15.*

[37] *Scholar, M., Singh, A. S., & Agrawal, K. K. (2025, April). Optimizing Diabetes Prediction with Multi-Layer Perceptron: A Comprehensive Analysis on the Pima Indians Dataset. In 2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI) (Vol. 3, pp. 1790-1795). IEEE.*

[38] *Teimoory, G. K., & Keyvanpour, M. R. (2025, April). An Explainable Ai Model for Diabetes Prediction Using Random Forest. In 2025 11th International Conference on Web Research (ICWR) (pp. 264-269). IEEE.*

[39] *Vangoori, S., Jerripothula, R. N., Gottumukkala, S. S., Venkatasai, V., & Srinivas, M. (2025, February). Classification of Early Onset of Diabetes with Machine Learning, Deep Learning Algorithms, and Bayesian Networks. In 2025 International Conference on Artificial Intelligence and Data Engineering (AIDE) (pp. 207-213). IEEE.*

[40] *Shamal, M., , S., Khalil, R., Zeebaree, S., Zebari, D., Abdulrahman, L., & Mahdi, N. (2024). Diabetic Prediction based on Machine Learning Using PIMA Indian Dataset. Communications on Applied Nonlinear Analysis.*

How to cite: M. Hajiee, S. M. Soltanzadeh. S, **Predicting diabetes using machine learning with gradient boosting algorithms XGBoost and LightGBM and web-based data,** Journal of Distributed Computing and Systems (JDCS), Vol 9, No 1, Pages 64-72, 2026.

Maryam Hajiee received her Ph.D. degree in computer engineering, with a focus on Software Systems. His research interests include wireless sensor networks, the Internet of Things, artificial intelligence, and quantum computers.

Email: Mhajiee@iau.ac.ir

Seyed Mohamadreza Soltanzadeh Solhdar received his MSc in Computer Engineering, majoring in Artificial Intelligence and Robotics from the Faculty of Artificial Intelligence, Cyberspace and Advanced Technologies, Islamic Azad University of South Tehran, Tehran, Iran. He received his BS in Computer Engineering from Arian Babol Institute of Higher Education in Science and Technology, Iran in 2023. His research interests include machine learning, neural networks, audio processing, and quantum computers.

Email: s.soltanzadehsolhdar@iau.ir