International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Enhancing Early Detection and Prediction of Diabetes Mellitus in Patients of Indian Origin through Rigorous Machine Learning Techniques with Comprehensive Models Evaluation

Prosanjeet Jyotirmay Sarkar*[1], Dr. Santosh Pawar[2]

Submitted: 03/11/2023 **Revised**: 22/12/2023 **Accepted**: 04/01/2024

Abstract: Worldwide, diabetes mellitus is considered to be the 2nd deadly disease. Diabetes mellitus is a severe medical condition characterized by an abnormality in blood glucose levels resulting from pancreatic dysfunction, namely the inability to produce insulin hormones. It is a potentially fatal condition that progresses gradually and often goes unnoticed. It has a high risk for harm, malfunction, and failure of human organs like the kidneys, heart, eyes, nerves, and hypertension. There are several researches for the prediction and detection of Diabetes mellitus. The medical practitioners confirm that there is no permanent cure for diabetes mellitus; it can be kept under control by early prediction and diagnosis. The impressive establishment of a public health care infrastructure for collecting crucial and delicate data. The uses of Machine learning algorithms and numerous interesting patterns are recognized for the early prediction and detection of diseases. The current research aims to create a reliable method for detecting and predicting diabetes mellitus at an early stage by utilizing machine learning (ML) techniques. ML algorithms were performed on the Pima India Diabetes Dataset (PIDD) to develop the model. In the experiment, we employed various machine learning models, including Naïve Bayes (NB), Logistic Regression (LR), decision tree (DT), Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbours (KNN), LightGBM (LGBM), and XGBoost (XGB), to identify cases of diabetes mellitus. Performance comparison of various ML models found that the XGBoost algorithm outperformed with an accuracy of 90.23%.

Keywords: *Accuracy, Classification, Diabetes, Machine Learning, Model, Prediction*

1. Introduction

Diabetes mellitus is normally known as diabetes by community people, a condition where the human body cannot produce enough insulin to maintain the blood glucose level [1]. According to a WHO report, around 700 million people will get infected with diabetes worldwide by 2045 [2]. Diabetes mellitus is a prevalent chronic condition that poses an essential challenge in both developed and developing nations. Presently, a large community of people are busy with their routine work and, due to this, unable to take care of their health [3]. Alcoholism, smoking, fast food, unhealthy food and poor nutrition affect the majority of individuals, resulting in a nutritional imbalance[4]. The primary energy source for the human body's physiological activities is blood glucose, derived from the consumption of carbohydrates[5]. The pancreas is a significant organ in the human body that releases insulin, which is essential for controlling blood sugar levels[6]. The imbalance of insulin levels in the body induces numerous severe complications in the body, including heart stroke, kidney failure, nerve damage, hypertension, blindness and pre-mature death[7], [8]. Diabetes mellitus is classified into three types: type-1, type-2 and gestational[9]. Type-1 is the pancreas's inability to produce insulin hormones, which leads to abnormal glucose levels. The precise cause, however, is still under investigation. As per studies, this condition is inherited and commonly manifests in child or young age[10]. The sole remedy entails providing patients with periodic health examinations, nutritious foods, and prescribed insulin doses. Type-2 in the worldwide 8 out of 10 diabetes patients suffer from type-2 diabetes, a condition characterized by reduced or absent production of insulin by the body[11]. Research indicates that type-2 diabetes develops when an individual's body weight exceeds 20% of the optimal weight for their height[12]. The remedy is augmenting insulin levels through consistent physical exercise and a nutritious diet. Gestational diabetes is a condition that occurs in certain women during pregnancy. The placenta hinders the body's insulin absorption during pregnancy, resulting in elevated blood sugar levels[13]. Gestational diabetes is not diagnosed prior to or following pregnancy. It can be readily treated with conventional medical treatment and a lifestyle modification.

The medical science industry uses various electronic technologies for recording, storing, and displaying distinct patient parameters. This advancement of intelligent systems in the medical domain provides medical practitioners with

1 Dr A. P. J. Abdul Kalam University, Indore, MP – 8023, INDIA
ORCID ID : 0000-0002-4659-6271
2 Dr A. P. J. Abdul Kalam University, Indore, MP – 8023,

vital information for early disease diagnosis. Current research work in the prediction of diabetes mellitus does not offer a comprehensive prediction for diabetes mellitus, and the results give much misclassification. The predictive analysis uses machine learning approaches to increase the accuracy of disease diagnosis, improve patient care, make the most significant use of available resources, and enhance clinical results. The paper aims to develop a machine learning models for accurate classification of diabetes mellitus using these eight algorithms: NB, LR, DT, RF, KNN, SVM, XGB and LGB and comparative study of proposed classification models for detection of diabetes mellitus.

The structure of this paper is outlined as follows: Section 2 provides a review of the existing literature on diabetes prediction. Section 3 details the proposed methodology employed in the development of various models. Section 4 outlines the statistical parameters used for evaluation. Section 5 delves into the experimental work and the results obtained, while Section 6 concludes the paper by discussing the implications and potential directions for future research.

2. Related Work

The team led by Umair Muneer Butt developed an innovative approach to classify diabetes mellitus by integrating long short-term memory (LSTM) and multilayer perceptron models. They employed the Pima Indian Diabetes Database from the UCI repository for their study. Their research revealed that while the multilayer perceptron model reached a peak accuracy of 86.08%, the LSTM variant showed a slightly higher accuracy of 87.26% [14].

Jobeda Jamal Khanam and colleagues focused on an artificial neural network (ANN) structure with two hidden layers, aiming for precise classification of diabetes mellitus. Utilizing the Pima Indian Diabetes Dataset (PIDD), they tested various machine learning algorithms including Decision Trees (DT), K-Nearest Neighbors (KNN), Random Forest (RF), Naive Bayes (NB), AdaBoost (AB), Logistic Regression (LR), Support Vector Machines (SVM), and Neural Networks (NN). Their findings highlighted the neural network model as the most effective, achieving an impressive 88.6% accuracy after 400 epochs [7].

Gaurav Tripathi and his team developed a predictive model for diabetes using four distinct machine learning classification methods: Linear Discriminant Analysis (LDA), KNN, SVM, and Random Forest (RF). They conducted experiments with the PIDD dataset to determine the relationship between features and classes, employing a confusion matrix to measure precision, accuracy, F1-score, and recall. The Random Forest model stood out, achieving a maximum accuracy of 87.66% [15].

Christobel Y. A and associates introduced a novel Class-Wise K-Nearest Neighbor (CKNN) classification. Their research, utilizing the Pima India Diabetes Database, aimed to evaluate the CKNN algorithm and compare it with traditional KNN performance metrics. Their innovative CKNN model surpassed the conventional KNN, reaching a peak accuracy of 78.16% [16].

Lastly, Quan Zou and his team applied five machine learning classification methods - Decision Tree (DT), Random Forest (RF), Artificial Neural Network (ANN), Principal Component Analysis (PCA), and Minimum Redundancy Maximum Relevance (mRMR) - to predict diabetes mellitus. They used the Pima India Diabetes Database (PIDD) for their analysis. Among the methods tested, the Random Forest approach was found to be most effective, achieving the highest classification accuracy of 77.21% [17].

3. Proposed Methodology

This research employs diabetes categorization and early prediction approaches in pattern recognition to categorize data into distinct groupings. The dataset is used to analyse diabetes mellitus is obtained from the PIDD dataset. Now, a more powerful classification algorithm supports days machine learning technology. These algorithms are also used in the medical field for early detection and treatment of the disease. The dataset trains machine learning models to predict future outcomes. To assess the precision of the model, employ mathematical statistical instruments. This study aims to construct a machine learning model that accurately predicts the beginning phase of diabetes by incorporating relevant factors closely associated with the condition in the medical domain. The proposed flow diagram in Figure 1 illustrates the basic steps employed for developing and evaluating ML models.

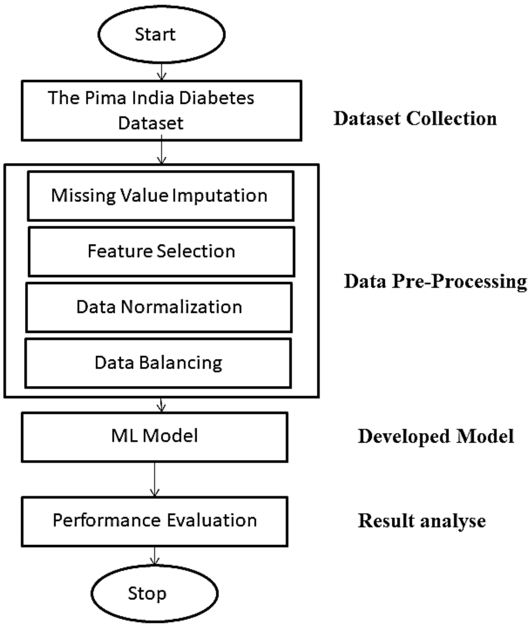


Fig. 1. Flow chart of proposed models

2.1 PIDD Dataset

This study employs the Pima Indian Diabetes Dataset (PIDD) sourced from Kaggle, originally gathered by the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset focuses on a specific group: adult females aged 21 and above, all of whom self-identify as Pima Indian. It includes 768 records, composed of 500 non-diabetic and 268 diabetic individuals. The dataset features eight predictive attributes and one binary outcome indicating the presence (1) or absence (0) of diabetes. The attributes include the number of times pregnant (PR), plasma glucose concentration (GL), diastolic blood pressure (BP), triceps skin fold thickness (ST), 2-hour serum insulin (IN), body mass index (BMI), diabetes pedigree function (PDF), and age (AG). An illustrative snippet of the data is provided in Table 1.

Table 1. Some examples of PIDD dataset

PR	GL	BP	ST	IN	BM	PDF	AG	Ou
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

2.2 Data pre-processing

It is the combination of data missing value imputation, feature selection, data normalization and data balancing process. If we use the dataset without pre-processing, the machine learning algorithm cannot make any proper model and give erratic output. To solve the above problem, we need to pre-process data before training the model.

2.2.1Missing Value Imputation

This approach involves verifying the presence of null and zero values within the dataset. We might employ a two-step approach to address the missing values in the dataset.

- Imputation method:- In this procedure, the feature class mean, median and mode assign the missing or null value in the dataset.
- Deleting the row:- This method can be used only when the dataset contains an extensive record; a few row deletions do not affect the data set records. Our record is small, so there are better approaches than deleting the record.

This research work used the mean method to impute missing values in the dataset. Consider the mean of the feature column and substitute in the missing or null place of the feature column.

2.2.2 Feature selection

It is the foundational method utilized to ascertain the optimal feature for a ML model. To simplify models, it is frequently applied to decrease computation time, memory consumption, dimension expansion, and predictive accuracy through the selection of essential features and prevent the inclusion of extraneous data that overfits the model. This article utilized an RF-based features selection methodology. Ranking features by RF is frequently determined by the node's impurity level, aiming to reduce impurities in the entire tree. The earliest nodes in the trees demonstrate the most significant reduction in impurity levels, while as the trees progress, a more progressive decline in impurity levels is observed. This results in the acquisition of a subset of the relevant attributes through the pruning of the trees beneath a particular node.

2.2.3 Normalization of features

This is a significant phase in the pre-processing procedure. The datasets consist of several feature columns, each with distinct scales. In order to create a high-quality machine learning model and achieve greater accuracy, it is necessary to normalize all characteristics, which involves scaling them to the same range. They commonly employed two strategies known as z-score and min-max methodology. We employed the min-max approach in this study, as shown in Equation 1.

$$x^{|} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

(1)

Where,

$x^{|}$ is the missing value in the feature sample.

x_{min} is the minimum value in the feature sample

x_{max} is the maximum value in the feature sample

x is the previous one value in the feature sample.

2.2.4 Balance and unbalanced data set

The PIDD dataset output is binary classification, and results dominated based on the majority. Assume that the positive class of the data sample comprises 90% of the data, whereas the negative class comprises 10%. Under those circumstances, there is unquestionably a presence of bias in favour of the positive sample side in the outcome. As a result, the model will exhibit high accuracy but provide misleading predictions. In order to address the issue above of imbalance, we employed the methodologies of random sampling and oversampling. The oversampling technique was employed in this study to address the issue of imbalanced datasets.

2.3 Algorithm used for modelling

2.3.1 Naive Bayes

The Naive Bayes method is a prevalent technique in supervised learning, particularly for tackling classification challenges. Known for its simplicity and effectiveness, it operates efficiently even with high-dimensional training datasets. The fundamental concept of this approach is a probabilistic classifier, which generates predictions by calculating the likelihood of a feature's presence. This method's mathematical formulation is detailed in equation 2.

$$P\left({}^{A}/_{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)} \qquad (2)$$

2.3.2 K-Nearest Neighbour (KNN)

This is a supervised machine learning technique applicable to both regression and classification problems. It determines the class of a new feature based on distance and similarity metrics. Known as a lazy learning algorithm, it retains the training data instead of immediately deriving insights from it. During the classification phase, the KNN algorithm executes certain operations on the dataset in accordance with a predefined series of procedural steps.

- In the training phase, load the feature data and class sample of the training sample.
- Choose the number of K of the neighbours included in the majority class as per Euclidean, Manhattan, and Minkowski's measuring distance. This paper used the Euclidean distance for the measuring distance between two samples, shown in Equation 3.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(yi - xi)\text{\^{}}2} \qquad (3)$$

- Place the new data sample in the category for which the neighbour number is farthest away.

2.3.3 Logistic Regression

Logistic regression is a prevalent and straightforward supervised learning technique for classification tasks. It addresses classification problems similarly to how linear regression predicts outcomes in a regression context. Focusing on the probability of an event's occurrence, such as voting or not voting, it calculates these probabilities based on the dataset of independent variables. The logistic function, which forms the core of this method, is mathematically articulated in equations (4) and (5).

$$\mathrm{f(x)} = \frac{1}{1+e^{-y}} \qquad (4)$$

$$y = b_0 + b_1x \qquad (5)$$

2.3.4 Decision Tree

A decision tree is employed for both classification and regression tasks but is especially preferred for classification problems. It graphically represents the dataset, delineating all possible decision paths based on specific criteria. A decision tree comprises two kinds of nodes: decision nodes and leaf nodes. While the decision nodes are responsible for making choices that branch out into various paths, the leaf nodes dictate the final outcome of the decision tree.

2.3.5 Random Forest

Random forest stands out as a highly effective supervised learning algorithm for both classification and regression tasks. While individual decision trees may exhibit significant variability, this fluctuation diminishes when they are collectively utilized in a parallel manner. As a result, the final decision is derived from an aggregate of multiple trees instead of just one. As an ensemble technique, random forest integrates various decision trees through bootstrapping, aggregation, and bagging methods. The bagging aspect particularly involves randomly selecting rows and features from the dataset to construct a sample dataset, which is then used to train each tree in the random forest.

2.3.6 Support Vector Machine

It is the predominant supervised ML method for case regression and classification. In this technique creates an optimal decision boundary that can effectively categorize the new data sample into the proper class within the feature. This decision boundary is capable of dividing the n-dimensional space into distinct classes. The primary difficulty in SVM is in selecting the optimal hyperplane within the dimensional space. The optimal hyperplane is characterized by the most significant separation between data points belonging to different classes. A support vector is a hyperplane positioned closest to the feature samples of the different classes. The unidentified data point is classified according to the hyperplane corresponding to one of the classes along the hyperplane.

2.3.7 XGBoost

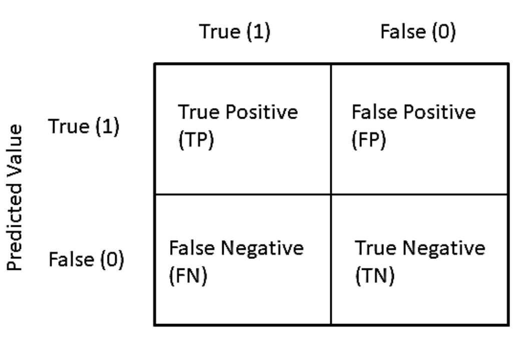
XGBoost, an abbreviation for Extreme Gradient Boosting, is a robust and dependable machine learning approach used for regression and classification. It operates as an ensemble method, incorporating both bagging and boosting techniques. The framework of XGBoost consists of three main elements: the loss function, a weak learner, and an additive model. While strong algorithms often grapple with data overfitting, gradient boosting, despite its greedy nature, can also overfit large datasets. To mitigate this, regularization techniques are implemented within XGBoost to curb overfitting and improve the overall performance of the algorithm.

2.3.8 LightGBM

LightGBM is a widely used supervised ensemble ML technique for regression and classification tasks. LightGBM is an abbreviation for Light gradient boosted machine, a machine learning algorithm designed to handle huge datasets. It enhances the gradient-boosting technique by giving more importance to situations with larger gradients and incorporating automatic feature selection. This could result in a rise in the effectiveness and precision of training.

2.4 Statistical tools for evaluation of performance

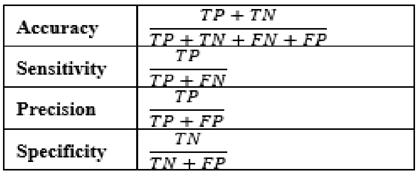
Figure 2 presents a confusion matrix that evaluates the effectiveness of different machine learning classification models. Performance metrics such as accuracy, precision, sensitivity, and specificity are derived using specific formulas, which are detailed in Table 2.



Predicted Value	Actual Value: True (1)	Actual Value: False (0)
True (1)	True Positive (TP)	False Positive (FP)
False (0)	False Negative (FN)	True Negative (TN)

Fig. 2. Confusion Maatrix

Table 2. Performance metrix

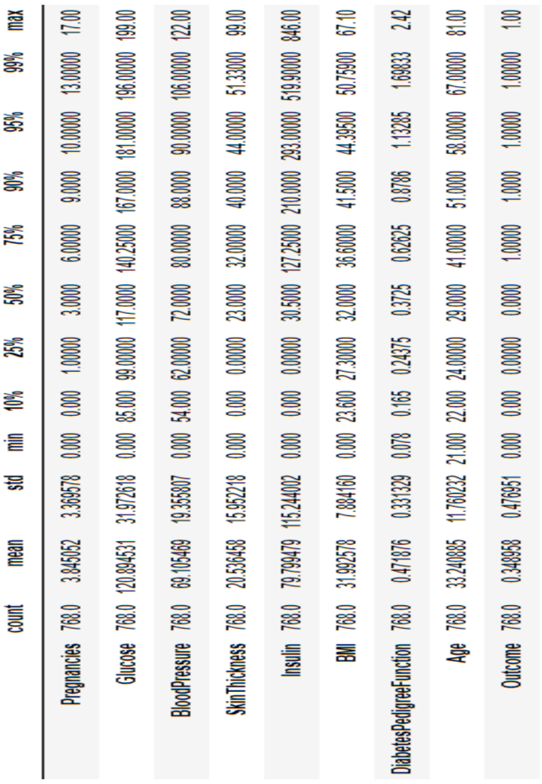


Accuracy	$\frac{TP+TN}{TP+TN+FN+FP}$
Sensitivity	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
Specificity	$\frac{TN}{TN+FP}$

Accuracy	$\frac{TP+TN}{TP+TN+FN+FP}$
Sensitivity	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
Specificity	$\frac{TN}{TN+FP}$

4. Experimental work and Result

This study utilizes eight practicable classification algorithms to construct a model that assists in the early-stage prognosis of diabetes mellitus, relying on notable attributes linked to this ailment. The algorithmic models used are NB, KNN, LR, DT, RF, SVM, XGB, and LGBM. This experiment used the PIDD dataset. In the model analysis, we have to check the description of the data set; we have to find missing values in the features column, imbalance of data, mean, variance, etc., shown in figure 3 and Table 3 and Figure 4 shows the comparison output of all classification models without hyperparameter tuning.



	count	mean	std	min	10%	25%	50%	75%	90%	95%	99%	max
Pregnancies	768.0	3.845052	3.369578	0.000	0.000	1.00000	3.0000	6.00000	9.0000	10.00000	13.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	85.000	99.00000	117.0000	140.25000	167.0000	181.00000	196.00000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	54.000	62.00000	72.0000	80.00000	88.0000	90.00000	106.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.000	0.00000	23.0000	32.00000	40.0000	44.00000	51.33000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.000	0.00000	30.5000	127.25000	210.0000	293.00000	519.90000	846.00
BMI	768.0	31.992578	7.884160	0.000	23.600	27.30000	32.0000	36.60000	41.5000	44.39500	50.75900	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.165	0.24375	0.3725	0.62625	0.8786	1.13285	1.69833	2.42
Age	768.0	33.240885	11.760232	21.000	22.000	24.00000	29.0000	41.00000	51.0000	58.00000	67.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.000	0.00000	0.0000	1.00000	1.0000	1.00000	1.00000	1.00

Fig. 3. Summary of the PIDD dataset

Table 3. Accuracy of all classical machine learning model without hyper parameter tuning

Model	NB	LR	KNN	DT	RF	SVM	XGB	LGBM
Accuracy (%)	82.13	84.86	84.07	80.12	88.15	85.39	89.07	88.28

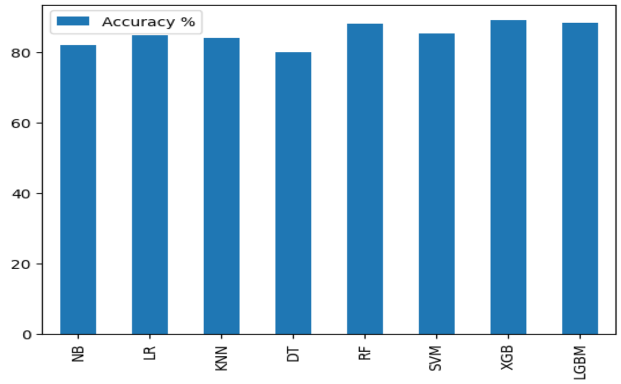


Fig. 4. Classification output of all Algorithms without hyper parameter tuning.

We will now select the top three models to enhance the classification model further. To obtain the most suitable model, it is necessary to conduct hyper-parameter tuning operations on the dataset. Table 4 displays the accuracy of various classification models, including LightGBM, XGBoost, and Random Forest. The Table 4 shows the accuracy rates for these models are 89.73%, 90.13%, and 89.73%, respectively and figure 5 displays the performance of all enhanced three model.

Table 4. Improved accuracy of classification model with hyper parameter tuning

Model	RF	XGB	LGBM
Accuracy (%)	89.73	90.23	89.60

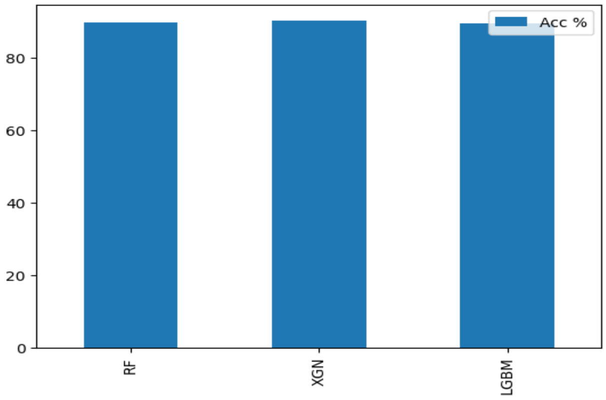


Fig. 5. Comparison performance of top three classification model with hyper parameter tuning

5. Conclusion and Future scope

Diabetes is a persistent medical illness that limits daily activities, diminishes quality of life, and rises mortality chances of the patients. Several researchers have studied the Pima Indian diabetes dataset to classify individuals as diabetic or non-diabetic. Medical researchers are proof that only early identification of diabetes can control the spreading of the disease. In this paper, we have modelled various classification algorithms and applied them to the Pima India diabetes dataset to find the best-fit model for maximum accuracy. The XGBoost algorithm achieves an optimal accuracy of 90.23%. In future endeavours, we can enhance the precision of the model by utilizing a substantial dataset. Additionally, machine learning algorithms may be utilized to accurately predict several diseases, including cancer.

Author contributions

Prosanjeet Sakar: Conceptualization, Methodology, Field study, Data Collection, Writing –original draft, Software, Validation.
Dr Santosh Pawar: Conceptualization, Methodology, Field study, Data Collection, Writing –original draft, Software, Validation.

Conflicts of interest

The authors do not have any conflict with other entitles or researches

References

[1] J. E. Sprietsma and G. E. Schuitemaker, ‘Diabetes can be prevented by reducing insulin production’, *Medical Hypotheses*, vol. 42, no. 1, pp. 15–23, Jan. 1994, doi: 10.1016/0306-9877(94)90029-9.

[2] P. Saeedi *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 Res Clin Pract*, vol. 157, p. 107843, Nov. 2019, doi: 10.1016/j.diabres.2019.107843.

[3] J. Lawton, N. Ahmad, L. Hanna, M. Douglas, and N. Hallowell, ‘“I can’t do any serious exercise”: barriers to physical activity amongst people of Pakistani and Indian origin with Type 2 diabetes’, *Health Education Research*, vol. 21, no. 1, pp. 43–54, Feb. 2006, doi: 10.1093/her/cyh042.

[4] P. Padrão, N. Lunet, A. C. Santos, and H. Barros, ‘Smoking, alcohol, and dietary choices: evidence from the Portuguese National Health Survey.’, *BMC Public Health*, vol. 7, p. 138, Jul. 2007, doi: 10.1186/1471-2458-7-138.

[5] T. M. S. Wolever, ‘Carbohydrate and the Regulation of Blood Glucose and Metabolism’, *Nutrition Reviews*, vol. 61, no. suppl_5, pp. S40–S48, May 2003, doi: 10.1301/nr.2003.may.S40-S48.

[6] P. V. Röder, B. Wu, Y. Liu, and W. Han, ‘Pancreatic regulation of glucose homeostasis.’, *Exp Mol Med*, vol. 48, no. 3, p. e219, Mar. 2016, doi: 10.1038/emm.2016.6.

[7] J. J. Khanam and S. Foo, ‘A comparison of machine learning algorithms for diabetes prediction’, *ICT Express*, vol. 7, Feb. 2021, doi: 10.1016/j.icte.2021.02.004.

[8] A. E. Kitabchi, G. E. Umpierrez, J. M. Miles, and J. N. Fisher, ‘Hyperglycemic Crises in Adult Patients With Diabetes’, *Diabetes Care*, vol. 32, no. 7, pp. 1335–1343, Jul. 2009, doi: 10.2337/dc09-9032.

[9] P. Bekkering, I. Jafri, F. J. van Overveld, and G. T. Rijkers, ‘The intricate association between gut microbiota and development of type 1, type 2 and type 3 diabetes’, *Expert Rev Clin Immunol*, vol. 9, no. 11, pp. 1031–1041, Nov. 2013, doi: 10.1586/1744666X.2013.848793.

[10] L. A. DiMeglio, C. Evans-Molina, and R. A. Oram, ‘Type 1 diabetes.’, *Lancet*, vol. 391, no. 10138, pp. 2449–2462, Jun. 2018, doi: 10.1016/S0140-6736(18)31320-5.

[11] U. Galicia-Garcia *et al.*, ‘Pathophysiology of Type 2 Diabetes Mellitus’, *Int J Mol Sci*, vol. 21, no. 17, p. 6275, Aug. 2020, doi: 10.3390/ijms21176275.

[12] S. N. Bhupathiraju and F. B. Hu, ‘Epidemiology of Obesity and Diabetes and Their Cardiovascular Complications’, *Circ Res*, vol. 118, no. 11, pp. 1723–1735, May 2016, doi: 10.1161/CIRCRESAHA.115.306825.

[13] H. D. McIntyre, P. Catalano, C. Zhang, G. Desoye, E. R. Mathiesen, and P. Damm, ‘Gestational diabetes mellitus’, *Nat Rev Dis Primers*, vol. 5, no. 1, Art. no. 1, Jul. 2019, doi: 10.1038/s41572-019-0098-8.

[14] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, ‘Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications’, *J Healthc Eng*, vol. 2021, p. 9930985, Sep. 2021, doi: 10.1155/2021/9930985.

[15] G. Tripathi and R. Kumar, *Early Prediction of Diabetes Mellitus Using Machine Learning*. 2020, p. 1014. doi: 10.1109/ICRITO48877.2020.9197832.

[16] T. S and D. C., ‘Classification using Convolutional Neural Network for Heart and Diabetics Datasets’, *IJARCCE*, vol. 5, pp. 417–422, Dec. 2016, doi: 10.17148/IJARCCE.2016.51296.

[17] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, ‘Predicting Diabetes Mellitus With Machine Learning Techniques’, *Frontiers in Genetics*, vol. 9, Nov. 2018, doi: 10.3389/fgene.2018.00515.