Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 6s (2025)
https://www.jneonatalsurg.com

Exploring an Efficient Machine Learning Technique for Diabetes Mellitus Prediction

Joshika Pradeep[1], S. Sathyanarayanan[1], S. Meenakshi[1]

[1]Department of Mathematical and Computational Sciences, Sri Sathya Sai University for Human Excellence, Navanihal, Gulbarga, 585316, Karnataka, India.

Email ID: joshika.p@ssslsg.org ; sathyanarayanan.s@sssuhe.ac.in, sathyanarayanan.brn@gmail.com ;

meenakshi.s@sssuhe.ac ;

Cite this paper as: Joshika Pradeep, S. Sathyanarayanan, S. Meenakshi, (2025). Exploring an Efficient Machine Learning Technique for Diabetes Mellitus Prediction. *Journal of Neonatal Surgery*, 14 (6s), 852-859.

ABSTRACT

Diabetes Mellitus is the most common disease characterized by hyperglycemia resulting from a deficit in the production or action of insulin. Untreated diabetes can lead to a myriad of complications. In 2021, it was estimated that 537 million individuals have diabetes, which is expected to increase to over 783 million by 2045. Preventing and managing this epidemic poses a significant challenge owing to various issues and barriers such as inadequate access to healthcare and lack of surveillance data. However, the advent of machine-learning (ML) techniques can easily address these critical problems. The main goal of this project is to develop a model capable of accurately predicting diabetes in patients. Five machine-learning classification algorithms were employed in this analysis to detect diabetes. The experiments were performed using a diabetes prediction dataset sourced from Kaggle. The efficiency of all five algorithms was assessed using various metrics such as the Kappa statistic, F-measure, precision, true positive rate, and recall. Accuracy was evaluated based on correctly classified and incorrectly classified instances. The results specify that the bagging algorithm outperformed the various other algorithms with the highest accuracy of 97.16% compared to other different algorithms..

Keyword: *Machine learning, Kappa statistic, Bagging, Diabetes, Confusion matrix*.

1. INTRODUCTION

The applications of Artificial Intelligence (AI) are pervasive in many domains such as healthcare, education, transport, image recognition, and many more. The goal of AI is to generate human-like intelligence in machines. This can be achieved through learning algorithms that try to imitate the human brain and its learning methods. ML is a field that enables systems to obtain human-like intelligence and exhibits outcomes without explicit programming. In particular, in healthcare, AI technology and ML techniques play a pivotal role in solving many barriers. AI methods are widely used to analyse heart sounds and phonocardiograms to detect heart diseases [1]. AI in the early detection of cardiovascular diseases (CVD) is beneficial in treating patients more effectively and preventing disease complications. This approach significantly enhances the long-term quality of life of patients. [2]. It is also used in other fields such as radiology, oncology, diabetic retinopathy, cardiac diseases, robotic surgical tools, clinical decision support systems, and medical imaging [3]. AI in the medical field is highly beneficial in various aspects such as keeping medical records, aiding surgeons during treatment, medication, operation, reviewing thousands of medical records and enhancing the speed of treatments with superior outcomes [4]. In this analysis, the ML technique was used to predict diabetes. This helps doctors to diagnose the disease in less time with high accuracy and treat it accordingly.

Diabetes mellitus (DM), a deadly chronic disorder, is characterized by increased blood sugar levels. Currently, it is one of the most-wide spread non-communicable disorders. The annual number of deaths due to diabetes is estimated to range from 2 to 5 million [5]. Based on the action and production of insulin, diabetes mellitus is classified as type 1, type 2, or type 3. A cross-sectional study of 202 individuals showed a greater risk of knee or hand osteoarthritis (OA) in patients with diabetes mellitus than in individuals who were not affected by diabetes mellitus. Patients with poorly controlled or severe diabetes are prone to increased levels of advanced glycation end products (AGEs) and hyperglycemia. As a result, these patients are expected to have a severe course of OA. [6]. Diabetes mellitus and coronary artery disease (CAD) are closely associated with each other. The risk of CAD was higher in patients treated for diabetes mellitus with insulin than in those treated for diabetes mellitus with oral agents and diet alone. Thus, it is essential to effectively treat diabetes mellitus effectively [7].

Therefore, progress in cutting-edge medical techniques is significant for the diagnosis of diabetes. Currently, there is a dire need for ML techniques to eliminate human effort by utilizing the automation ability of the system with minimum flaws

[8]. ML algorithms are categorized into three groups, supervised learning, unsupervised learning, and reinforced learning. ML techniques are also employed for diverse purposes in the healthcare and other sectors. In the healthcare field ML is used for data analysis, to predict uncertain disorders, mortality rates etc [9]. Weka, an auto ML tool, provides several algorithms

implemented in Java, for data mining and prediction models. It is a platform in which real-world problems are solved in an easier manner using ML techniques. This software is efficient in taking up large volumes of data and performs well

2. MOTIVATION

Over the past few decades, India has encountered a drastic shift in the contribution of primary factors of mortality from communicable to non-communicable diseases. Diabetes is a health concern because of its rising prevalence. Diabetes diagnosis and management demand urgent attention as it is an easily measurable disorder. ML techniques provide a beneficial approach for addressing this challenge. By utilizing data-driven insights, these techniques enhance the diagnostic process and reduce reliance on manual analysis and decision making. Additionally, they handle complex, multi-dimensional datasets and enhance the accuracy and efficiency of diabetes prediction.

Despite many advancements, current medical diagnostics occasionally exhibit errors, such as classifying diabetic individuals as non-diabetic or vice versa, or encountering problems where the diagnosis is inconclusive [5]. To mitigate these issues, there is a pressing need to develop AI-powered systems capable of accurately predicting the diabetes status based on robust training. AI-driven diagnostics play a crucial role in the healthcare sector, in various aspects such as prediction and analysis. For example, cardiovascular disease (CVDs) is a major issue that humanity is currently facing. AI methods have a great potential for early detection [10].

Weka is an open-source platform, that offers diverse ML algorithms for activities related to predictive modeling and data mining tasks. It has emerged as a particularly effective solution among the many varieties of ML tools. Its extensive feature set comprises features such as regression, clustering, and classification, providing users with the flexibility to analyze data and comprehensively assess the model performance. Through efficient model training and evaluation, Weka stands out as a valuable tool for accurate diabetes prediction and has the potential to significantly impact healthcare outcomes.

3. DIABETES PREDICTION DATASET

The dataset chosen here encompasses medical and demographic data from patients, including attributes such as age, sex, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. Each attribute uniquely contributes to the development of diabetes. Body mass index coexists with an enhanced risk of type 2 diabetes [11] when it falls into obese categories leading to an increase in insulin resistance. Common risk factors such as obesity, unhealthy diet, physical inactivity, hypertension and high blood pressure often overlap with diabetes. Smoking is a known risk factor for type 2 diabetes as it contributes to insulin resistance and impairs glucose metabolism. HbA1c or glycated haemoglobin levels were used to measure long-term blood glucose level control. An increase in HbA1c level indicates poor glycemic control with a high risk of developing diabetic complications.

This dataset is utilised in ML models to predict the likelihood of diabetes in patients based on their medical history and demographic information. Weka analyses the dataset to determine the prediction accuracy using the Kappa statistic, and other relevant measures using various data mining algorithms. These algorithms were employed to train models that predicted diabetes based on the influence of the aforementioned attributes. Kaggle is the best platform for both ML engineers and data scientists, to create ML models for specific problems and analyse certain datasets. The dataset used for the analysis was sourced from Kaggle and had a sample size of 100,000 instances. The machine learning tool Weka was used for analysis and modeling purposes.

4. EVALUATION METRICS

The present study provides a comprehensive assessment of the prediction performance of various algorithms. This helps to identify the most suitable algorithm for accurate diabetes prediction, which guides healthcare practitioners in selecting the most effective prediction model. The evaluation metrics were analysed using confusion matrix, which provides the prediction of the model determined by accuracy, precision and recall. The computation of Kappa statistics assesses the agreement between two variables, namely the model’s judgment and actual outcome. The Kappa values lie in the range of [-1,1] where ’1’ denotes total agreement and ’0’ represents no agreement and no independence. A negative statistic indicates that it is worser than random. A kappa value ≥ 0.75 marks the classifier is good and acceptable. Other statistical measures include Mean Absolute Error (MAE), which is computed as the average absolute error. Therefore, the difference between the expected and projected values may be positive or negative, but it will always be positive when calculating the MAE. The root mean square error (RMSE) is the square root of the residual variance, representing the standard deviation of the unexplained variance. The reduced RMSE values indicate a better fit of the model [12]. The relative absolute error (RAE) is the ratio of the mean error to the error produced by the model. A reasonable model is characterized by a ratio of less than one, which indicates that it outperforms a simple model in terms of results [13].

The following are a few metrics, the true positive rate (TP) represents the samples that are correctly classified. A higher value denotes a better model. The false positive rate (FP) rate determines the samples that are incorrectly classified. A smaller value implies a better precision [14]. The F-measure is the ratio of the combination of recall and precision, instances correctly classified by the model to the instances classified as positive. The accuracy of the model was determined using all previously computed values. In addition, the computational time required by each algorithm to make predictions varied.

5. METHODOLOGY

The diabetes prediction dataset underwent eight steps in order to build the model. Figure 1 illustrates the procedure for building the model. The dataset comprised of two categories: individuals with and without diabetes. Subsequently, five machine learning (ML) models were developed by employing algorithms distinct from major classifiers.

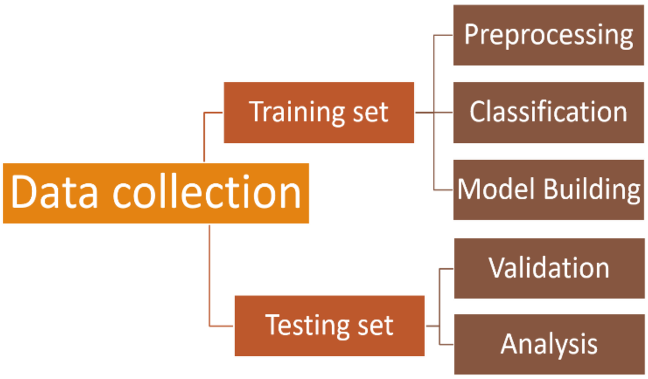

Figure 1 Steps involved in building a model.

a) **Data Collection**: The first step in building an ML model is to obtain accurate data for a specific problem. The dataset was taken from the online platform called Kaggle. The sample size of the dataset is 1,00,000 instances. Once the data is obtained it is divided into the following two stages.

b) **Training Data**: In the first stage, the data is trained using various algorithms to obtain an efficient output model. This includes the following steps.

i) **Pre-processing**: During the training stage, the data is prepossessed and the raw data is converted into structured data. It makes the dataset suitable for the analysis and also ensures uniformity. To enhance the algorithmic evaluation, certain attributes were converted from numeric to nominal type. Table 1 represents the type of the attributes converted using filters to upgrade the performance of the algorithms. The attribute types are listed in Table 1.

Table 1 Dataset Details

Attributes	**Type**
Gender	Nominal
Age	Numeric
Hypertension	Numeric
Heart disease	Numeric
Smoking History	Nominal
BMI	Numeric
HbA1c level	Numeric
Blood Glucose level	Numeric
Diabetes	Nominal

ii) **Classification**: The attributes in the data were classified using several available classification algorithms. The algorithms include Naive Bayes, Bagging, PART, Logistic regression and Random Forest, each utilises a unique ML classification technique for model building.

iii) **Model building**: The last step in the training set was model building. After choosing the algorithm the model was run. The model is trained for algorithms to learn the relationships between the input features and target variables.

c) **Testing set:** The next level is to test the data set. It includes the following steps.

i) **Validation**: After the model is trained, the performance is validated and examined using different techniques such as 'cross-validation', 'use training set', 'supplied test set', and 'percentage split'. Here, a 'cross-validation test' was used to test the model.

ii) **Analysis**: In post-training, the model's performance is analysed using a suite of performance metrics, including the kappa statistic, RAE, RMSE, MAE, *TP*, recall, precision, accuracy and confusion matrix analysis. The metrics precision, kappa statistic, and F-score were specifically calculated to estimate the models' ability to accurately identify and classify positive standards among all positive variables. Thus, the effectiveness and robustness of the model were recognized.

Finally, the confusion matrix summarises the predictions made by the model against the existing values in the dataset. The general form of the confusion matrix is given in Table 2.

Table 2 Confusion Matrix

Authentic grouping class	**Classifier defines category**	
	Categorised as class	**Not categorised as class**
Record falls within the class	True Positive (TP)	False Negative (FN)
Record falls apart from the class	False Positive (FP)	True Negative (TN)

6. ANALYSIS OF VARIOUS ML TECHNIQUES

The objective of this section is to perform a comprehensive assessment of five different ML models and identify the most suitable model for prediction purposes. Diabetes prediction was conducted by selecting algorithms that represented five major classes. From each class, one representative algorithm was chosen that was appropriate for the dataset for the analysis to determine the best prediction model. The chosen algorithms include Naive Bayes from Bayes classifier, logistic regression from functions, bagging, PART from rules and random forest (RF) from trees.

6.1 Metrics analysis of classification algorithms

Evaluating the performance of classification algorithms requires the use of appropriate metrics and indicators to evaluate their accuracy, reliability and overall efficacy. Several key metrics are commonly used in this analysis, each of which provides valuable information about the advantages and dis advantages of different algorithms. In the table, the downward arrow indicator signifies that a lower value of the indicator reflects better algorithm performance. The values highlighted in bold indicate the best values obtained by the respective algorithms for the evaluation criteria of the training set.

Table 3 Performance of the classification algorithms on the diabetes prediction dataset

Algorithms	**Naïve Bayes**	**Logistic Regression**	**PART**	**Bagging**	**Random Forest**
Kappa	0.6281	0.7082	0.7832	**0.7897**	0.7766
MAE ↓	0.0755	0.0636	**0.0454**	0.0456	0.0468
RMSE ↓	0.214	0.1781	0.1567	**0.1536**	0.1593
RAE ↓	0.485	0.4091	**0.2920**	0.2928	0.3011
TPR	0.966	0.991	0.997	**0.998**	0.995
FPR ↓	0.330	0.371	0.314	0.314	**0.308**

Precision	0.969	0.966	**0.972**	**0.972**	**0.972**
Recall	0.966	0.991	0.997	**0.998**	0.995
F measure	0.968	0.979	0.984	**0.985**	0.983

1) **Kappa statistic:** From Table 3, it is evidence that the kappa value generated by the bagging algorithm surpasses those of all other algorithms. This algorithm achieved the highest kappa value of 0.7897. Hence it can be concluded that it is the most suitable algorithm based on this criterion.

2) **Mean absolute error (MAE):** MAE, representing the average of all absolute errors is a crucial metric where the lower values signify better model effectiveness. Among all the algorithms considered, the bagging algorithm from the metaclassifier exhibited the most favorable performance with the lowest MAE.

3) **Root mean square error (RMSE):** RMSE, which indicates the standard deviation of the variance, is an essential feature for model fit. Lower RMSE values indicate a better model performance. Notably, the bagging algorithm exhibited the lowest RMSE, underscoring its superior model fit and performance.

4) **Root absolute error (RAE):** RAE is given as a ratio comparing the mean error to the error produced by the model. A value less than one indicates good performance of the model. The PART algorithm showed good performance.

5) **True positive rate (TPR)**: This is defined as the number of variables classified correctly for a given class.

$$TPR = \frac{TP}{TP+FN} \times 100\%$$

The true positive (*TP*) rate which was 0.998 served as a key indicator of the effectiveness of the bagging algorithm.

6) **False positive rate (FPR):** This represents slight deviations made by the model by classifying positive instances as negative. Thus, a lower value indicates enhanced model performance. Among all the algorithms, RF obtained the lowest value compared to other algorithms.

7) **Precision:** Precision is calculated by

$$Precision = \frac{TP}{TP+FP} \times 100\%$$

The three models, PART, bagging and random forest yielded a precision value of 0.972 indicating their fitness.

8) **Recall:** It is a performance metric employed in the evaluation of classification models. It assesses the potential of the model to capture all positive instances within the dataset. A higher recall value indicates that the model effectively classifies the positive instances.

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

Here, the recall value of the bagging algorithm was 0.998.

9) **F-measure:** The F-measure within the range of [0, 1], acts as a pivotal metric for evaluating the accuracy of the model. It is calculated as the symmetric mean between the precision and recall. The F-score provides insight into both the precision and robustness of the model's predictions. In this instance, the F-value was 0.981 which highlights robustness and effectiveness the bagging model.

10) **Timing:** Although it required more time than the other algorithms, the bagging algorithm achieved the best result indicating that it is suitable for high dimensional data. The time required by each algorithm to build the model is presented in Table 4.

Algorithms	**Time taken (in sec)**
Naive Bayes	0.56
Logistic	2.72

Table 4 Duration table

Bagging	12.34
PART	19.01
Random Forest	74.05

6.2 Comparative analysis of classification algorithms

All algorithms achieved an accuracy exceeding 94% as shown in Table 5. When evaluating the execution of each algorithm according to the aforementioned features, the bagging algorithm demonstrated the highest level of accuracy, making it the optimal choice for this dataset.

Table 5 Accuracy table

Algorithms	**Accuracy**
Naive Bayes	94.133 %
Logistic	96.026 %
Bagging	**97.165** %
PART	97.059 %
Random Forest	96.926 %

Table 5 indicates that the bagging algorithm emerged as the top-performing choice across all the examined metrics followed by the PART and Random Forest algorithms. Thus, it demonstrates superior accuracy, effectiveness and model fit compared to alternative algorithms.

The confusion matrix for the bagging algorithm is given in Table 6 which meticulously details the execution of the bagging algorithm underscoring the resilience of the model. This demonstration of superiority emphasizes its potential for real-world applications and ability to provide valuable insights into data classification and prediction tasks.

Table 6 Confusion Matrix Table

	Diabetic	**Non- Diabetic**
Diabetic	**91331**	169
Non- Diabetic	2666	**5834**

7. CONCLUSION

In this research, various ML techniques were used in the data collection to scrutinize and evaluate the classification ability and applicability of each class of ML algorithms. Remarkably, the bagging algorithm yielded the highest accuracy of 97.165%. In addition, the PART algorithm can be considered a notable alternative to the bagging algorithm because of its high accuracy. These trained models can assist doctors in making predictions based on the diagnoses. However, in reality doctors must combine their empirical knowledge with patient information to comprehensively assess the illness. Thus, the problem of surveillance or privacy of data and multi modal fusion can be encountered. Furthermore, this approach can be extended to analyse the probability of diabetic progression in non-diabetic patients at an early stage. A graphical representation of the effectiveness of each model, using various measures is shown in Figure 2. These measures provided insights into the accomplishment of the model.

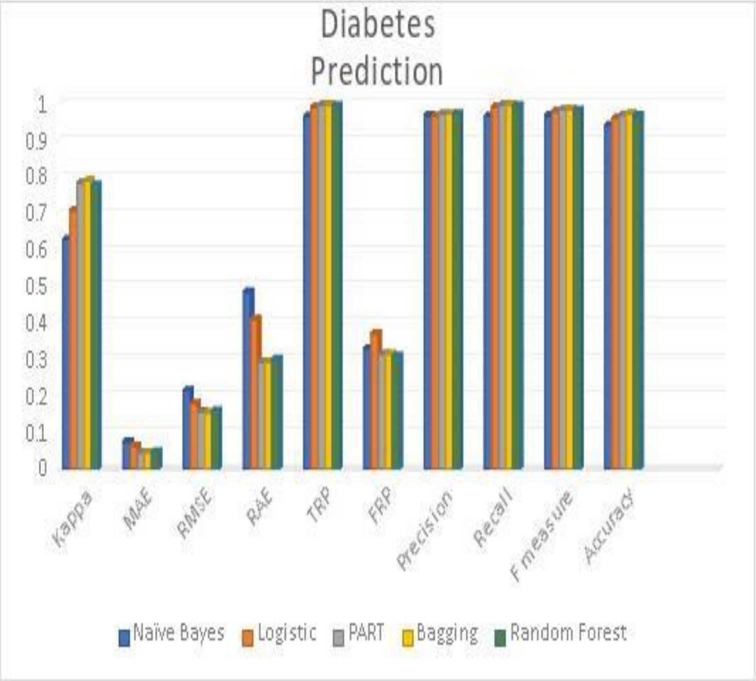

Figure 2 Performance of classification algorithms

REFERENCES

[1] Satyanarayanan. S, Srikanta Murthy K, S. Chitnis. "A comprehensive survey of analysis of heart sounds using machine learning techniques to detect heart diseases." J. Popul.Ther. Clin. Pharmacol, vol.30, no. May, pp.375-384, 2023.

[2] Swaminathan. S, Krishnamurthy, S. M., Gudada, C., Mallappa, S. K., and Ail, N. "Heart Sound Analysis with Machine Learning Using Audio Features for Detecting Heart Diseases". International Journal of Computer Information Systems and Industrial Management Applications, 16(2), 17-17 (2024).

[3] Sathyanarayanan. S and Sanjay Chitnis. "A survey of machine learning in healthcare." Artificial Intelligence Applications for Health Care, I., M.K. Ahirwal, N.D. Londhe, and A. Kumar, Eds. Boca Raton: Taylor and Francis Ltd, (2022): 1-22.

[4] Haleem, Abid, Mohd Javaid and Ibrahim Haleem Khan. "Current status and applications of Artificial Intelligence (AI) in medical field: An overview." Current Medicine Research and Practice 9.6 (2019): 231-237.

[5] Mujumdar, Aishwarya and Vaidehi. V. "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299.

[6] Nielen, Johannes TH, et al. "Severity of diabetes mellitus and total hip or knee replacement: a population-based case–control study." Medicine 95.20 (2016): e3739.

[7] Lemp, George F, et al. "Association between the severity of diabetes mellitus and coronary arterial atherosclerosis." The American journal of cardiology 60.13 (1987): 1015-1019.

[8] Rani, Jyoti. KM. "Diabetes prediction using machine learning." International

Journal of Scientific Research in Computer Science Engineering and Information Technology 6 (2020): 294-305

[9] Osuwa, Abdulhafis Abdulazeez and Hu¨seyin Oztoprak. "Importance of continuous¨ improvement of machine learning algorithms From A Health Care Management and Management Information Systems Perspective." 2021 International Conference on Engineering and Emerging Technologies (ICEET). IEEE, (2021).

[10] S. Swaminathan, S.M.K, C. Gudada, and S.K. Mallappa, "Use of Audio Transfer Learning To Analyse Heart Sounds For Detecting Heart Diseases", Nanotechnology Perceptions, vol. 20, no. S11 (2024), pp. 1863-1878,

2024

[11] Bekkering, Pjotr, et al. “The intricate association between gut microbiota and development of type 1, type 2 and type 3 diabetes.” Expert Review of Clinical Immunology 9.11 (2013): 1031-1041.

[12] Karunasingha, Dulakshi Santhusitha Kumari. “Root mean square error or mean absolute error? Use their ratio as well.” Information Sciences 585 (2022): 609-629.

[13] Pandey, Anand Kishor, and Dharmveer Singh Rajpoot. “A comparative study of classification techniques by utilizing WEKA.” 2016 International Conference on Signal Processing and Communication (ICSC). Ieee, (2016).

[14] Obuandike, Georgina. N, Isah Audu and Alhasan John. “Analytical study of some selected classification algorithms in weka using real crime data.” (2015)