Predicting Diabetes Mellitus in Healthcare: A Comparative Analysis of Machine Learning Algorithms on Big Dataset

***Corresponding author:**

Chandrakanth Rao Madhavaram

Infosys, Technology Lead

And

Eswar Prasad Galla[2], Mohit Surender Reddy[3], Manikanth Sarisa[4], Venkata Nagesh Boddapati[5], Siddharth Konkimalla[6]

[2]Infosys, Senior Support Engineer
[3]Motorola Solutions, Sr Network Engineer
[4]Sr Application Developer, Bank of America
[5]Microsoft, Support Escalation Engineer
[6]Amazon Com LLC, Network Development Engineer

SPECIAL EDITION

2021

Published in the Journal of

Global Journal of Research in Engineering & Computer Sciences

ISSN: 2583-2727 (Online)

Volume: 1 Issue: 1 (2021)

Published by GJR PUBLICATION

OPEN ACCESS

Global Journal of Research in Engineering & Computer Sciences
ISSN: 2583-2727 (Online)
Volume 01| Issue 01 | 2021
Journal homepage: https://gjrpublication.com/gjrecs/

Research Article

Predicting Diabetes Mellitus in Healthcare: A Comparative Analysis of Machine Learning Algorithms on Big Dataset

*Chandrakanth Rao Madhavaram [1], Eswar Prasad Galla [2], Mohit Surender Reddy [3], Manikanth Sarisa [4], Venkata Nagesh Boddapati [5], Siddharth Konkimalla [6]

[1]Infosys, Technology Lead
[2]Infosys, Senior Support Engineer
[3]Motorola Solutions, Sr Network Engineer
[4]Sr Application Developer, Bank of America
[5]Microsoft, Support Escalation Engineer
[6]Amazon Com LLC, Network Development Engineer

DOI: 10.5281/zenodo.14010835

***Corresponding author:** Chandrakanth Rao Madhavaram
Infosys, Technology Lead

Abstract

As the leading cause of death and morbidity among non-communicable diseases, diabetes impacts millions of people worldwide. Diabetes mellitus, commonly referred to as diabetes, poses a significant worldwide public health issue. According to the International Diabetes Federation, by 2040, there would be 642 million individuals living with the disease, up from 415 million now. Early risk prediction is crucial for diagnosis and prevention of this chronic condition since it impairs the body's capacity to absorb glucose. This study presents a comprehensive evaluation of machine learning techniques for diabetes outcome prediction using data from the UCI Machine Learning Repository. Training (80%) and testing (20%) subsets of the dataset are used to evaluate several classifiers, including Support Vector Machines (SVM), Multi-Layer Perceptrons (MLP), and Gradient Boosting Machines (GBM). When performance is measured using accuracy and F1-score, the results demonstrate that the GBM model outperforms the MLP and SVM by a significant margin, with an accuracy of 96.92% compared to 76% and 77.73%, respectively. This study highlights the superior predictive capability of the GBM model, emphasizing its potential to enhance diabetes management and support healthcare professionals in making informed clinical decisions. These findings contribute to the growing body of evidence supporting the integration of machine learning in healthcare settings for improved patient outcomes.

Keywords: *Healthcare, prediction, Machine learning, diagnosis, Diabetes mellitus, Data mining, diabetes dataset.*

Introduction

The healthcare data are produced in a variety of forms and from a variety of sources. In order to generate useful information, integrating health data and bringing it to a shared platform for additional analysis calls for sophisticated tools and procedures.

The variability, inconsistency, incompleteness, etc. of health care prevents the healthcare workers from gaining useful knowledge for usable clinical intelligence; Healthcare providers use a variety of methods to predict diabetes mellitus[1]. Hyperglycemia is a hallmark of diabetes mellitus, a chronic illness [2][3]. Numerous difficulties might result from it. In 2040, there will be 642 million diabetic patients worldwide, meaning that one in ten persons will have the disease, according to rising morbidity in recent years. Without a doubt, this concerning statistic requires careful consideration. Numerous facets of medical health have benefited from the quick growth of machine learning[4].

Diabetes mellitus, often referred Known as diabetes, this chronic metabolic illness impairs the body's ability to use food as fuel, which raises blood sugar levels. Hyperglycemia toxicity, in which the bloodstream is saturated with sugar, heart disease, gum disease and tooth decay, renal failure, and various diseases that can be fatal are among the problems that diabetes produces in the human body. Diabetes has no known cure, however early identification allows patients to prevent or halt the progression of the condition [5].

An examination of the most popular machine learning methods for determining the number of people with diabetes mellitus. AI and ML [6] contribute significantly to the control of diabetes by empowering patients to choose wise dietary and physical activity choices. Given that diabetes can strike anybody and that its symptoms are difficult to identify, early identification is crucial and confirms the necessity of routine checkups.[7]. The diabetes prediction system diagnoses diabetics using machine learning. In order to diagnose diabetes, the supervised learning algorithm is also used to train the diabetes prediction system[8].

1.1 Motivation and Contribution paper

The rising incidence of diabetes mellitus throughout the world presents serious health issues, making the creation of efficient prediction models for early detection and treatment necessary. Rapidly identifying those who are at risk is essential to enhancing patient outcomes and reducing the strain on healthcare systems, since millions are impacted. This study aims to harness machine learning algorithms to enhance diabetes prediction capabilities, addressing the urgent need for data-driven approaches in healthcare that facilitate proactive management and targeted prevention strategies. This paper makes several key contributions to the field of healthcare analytics and ML. In following contributions are:

- Utilize the diabetes dataset for predicting diabetes Mellitus.
- Implements label encoding to convert categorical variables into numerical format, facilitating the application of ML algorithms.
- Employs the ETC for feature importance analysis, identifying key predictors of diabetes, which helps in reducing dimensionality and improving model performance.
- Applies normalization techniques, specifically Standard Scaler, to standardize the dataset, enhancing the performance.
- Conduct a comparative analysis of various ML models, including GBM, MLP, and SVM, to evaluate their effectiveness in predicting diabetes outcomes.
- Utilizes performance matrix like AUC-ROC, accuracy and f1-score.

1.2 Structure of paper

The remainder of the paper is organised in this manner. Research on diabetes mellitus prediction in an industrial setting is presented in Section 2. The approach is described in depth in Section 3. The findings, analysis, and discussion are contrasted and compared in Section 4. The study's findings and recommendations for more research are presented in Section 5.

Literature Review

Researchers have recently demonstrated an increasing interest in the development of Predicting Diabetes Mellitus. Some background studies are provided in below:

This study Agarwal and Saxena, (2019) creates a model for One well-known diabetes dataset is the Pima Indians Diabetes Dataset research that includes information on Pima women, who are disproportionately affected by diabetes. The cardinal factor of this dataset is that the features are physical factors rather than dependent on region of the women. To successfully predict and diagnose diabetes, I worked on finding the best-suited algorithm for this purpose. Finding the best accuracy by comparing the various algorithms is the primary objective. DT, LR, Naïve Bayes, SVM, and KNN are the algorithms that are being compared. K-Fold and Cross Validation helped us achieve an accuracy of 81.1% in the end [9].

In this paper Yahyaoui et al., (2019), present the concept of a machine learning (ML) model for diabetes prediction using Decision Support Systems (DSS). We contrasted traditional algorithmic machine learning techniques with deep learning methods. We examined RF and the SVM classifier, or SVM, which are the two most often used classifiers for conventional ML techniques. In the suggested study, however, a fully CNN for DL was used to predict and identify the diabetes individuals. 768 samples with precisely 8 characteristics each were employed, along with the Indians Diabetes dataset Pima that is accessible through the Dew Media public repository, to evaluate the suggested process. The first 268 samples are classified as diabetic, while the remaining 500 samples are placed in the non-diabetic category. Accuracy was 76.81% for DL, 65.38 for SVM, and 83.67% for RF. According to the experimental research, RF was more effective in predicting diabetes than deep learning and SVM methods [7] .

This study Islam et al., (2019), have gathered 340 cases, each including 26 characteristics of individuals with diabetes that exhibit a range of symptoms divided into groups that are normal and those that are not. After the dataset was trained using the cross-validation approach, three ML algorithms—RF, LR, and Bagging—were applied for classification. At 90.29%, 83.24%, and 89.12%, respectively, Random Forest, Logistic Regression, and Bagging all have incredibly impressive accuracy rates [10].

This research Kowsher et al., (2019) 7 ML classifiers and an ANN approach are compared in order to recognize and treat diabetes patients as soon as feasible. Data from 9483 diabetic people make up our training and test dataset. The size of the training dataset prevents overfitting and yields very precise test results. We choose the best approach, deep ANN, by using performance indicators like accuracy and precision. With an accuracy of 95.14%, it outperforms all other studied ML classifiers. We anticipate that hospitals will be able to predict diabetes using our effective method, which will also stimulate research into more accurate prediction models [11].

In order to help with patient categorisation for intense This study examined case management among individuals with type 2 diabetes Seng et al., (2016), examines the use of predictive analytics to EHR data in a Singaporean academic health system. They have created a risk score for high healthcare EHR users using a multidisciplinary team approach. In order to forecast the top 10% of healthcare spenders in 2011, The Akaike Information to obtain this risk score, the backward stepwise variable selection model-building approach was combined with a multiple logistic regression model and criterion. Among the variables of the risk score were sociodemographic, biochemical, comorbid, and healthcare usage parameters. Compared to using the 2010's total cost as the sole prediction, the risk score's Area under the Curve (AUC) was greater at 0.708. If routine biochemistry measures were a part of the clinical practice for T2DM, the lack of them may be seen as either a sign that the patient's condition is being positively perceived or as a sign that they are not receiving frequent follow-up for treating their disease. In order to provide a comprehensive interpretation of a risk score, close cooperation across several disciplines is essential [12].

The background study of Comparative Analysis of Predicting Diabetes Mellitus with its dataset, models, performance, and contribution is provided in Table 1.

Table I. Comparative Study on Predicting Diabetes Mellitus using multiple approaches

Author	Methods	Data	Performance	Limitation/future work
Agarwal and Saxena,	SVM, KNN, Naïve Bayes, Decision Trees, and Logistic Regression	Pima Indians Diabetes Dataset (768 instances, 8 features)	Accuracy of K-Fold and Cross Validation: 81.1%Accuracy of K-Fold and Cross Validation: 81.1%	Limited to a specific demographic; further testing on diverse populations needed.
Yahyaoui et al.,	SVM, Random Forest, Convolutional Neural Network (CNN)	Pima Indians Diabetes Dataset (768 instances, 8 features)	SVM: 65.38%, RF: 83.67%, DL: 76.81%	RF outperformed; explore other deep learning architectures for better accuracy.
Islam et al.,	Bagging, Logistic Regression, Random Forest	Custom dataset (340 instances, 26 features)	Bagging: 89.12%, LR: 83.24%, RF: 90.29%	Limited dataset size; consider larger and more diverse datasets for generalization.
Kowsher et al.,	Various ML classifiers, Deep Artificial Neural Network (ANN)	Dataset of 9483 diabetes patients	Accuracy: 95.14%	High accuracy may not translate to clinical settings; explore real-world applicability.
Seng et al.,	Multiple Logistic Regression	EHR data for T2DM patients (not specified)	AUC: 0.708	Lack of biochemistry data; future work could integrate more clinical parameters.

Research Methodology

For Machine learning for diabetes mellitus prediction model, following steps of methodology workflow are present in figure 1. In this study, the methodology focuses on predicting diabetes mellitus by analyzing a large data extracted from UCI's ML repository. The dataset contains diabetes-related symptoms from 520 individuals. The initial stage involves comprehensive data preprocessing, including handling missing values through imputation and removing redundant data to ensure high-quality inputs. The Standard Scaler technique is applied to normalize features, ensuring all attributes are on a common scale, thus improving model accuracy. Categorical variables are transformed into numerical formats using label encoding to make them compatible with machine learning algorithms. Next, feature importance is assessed using the Extra Trees Classifier (ETC), which assigns importance scores according to each feature's role in forecast accuracy. The characteristics that have been found to have the greatest influence on diabetes outcome prediction are polyuria and polydipsia. To train and assess different machine learning models, the data is then separated into training (80%) and testing (20%) sets. The performance of the GBM, MLP, and SVM classifiers is assessed using confusion matrices and performance measures including accuracy and F1-score.

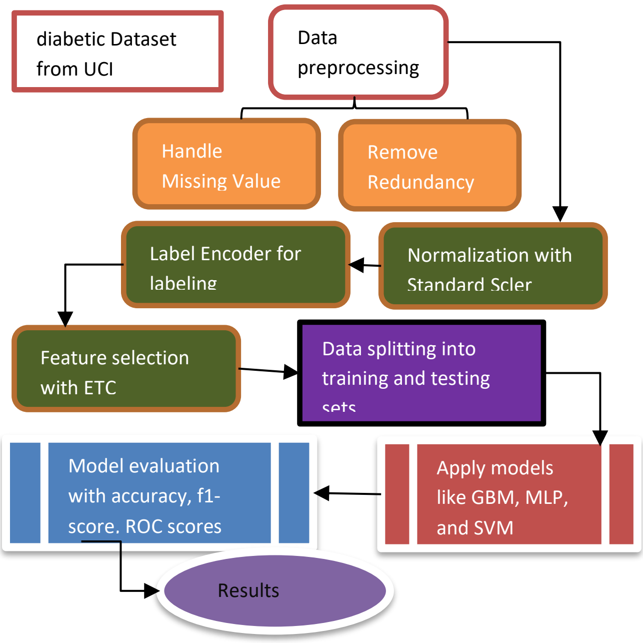
diabetic Dataset from UCI
Data preprocessing
Handle Missing Value
Remove Redundancy
Label Encoder for labeling
Normalization with Standard Scler
Feature selection with ETC
Data splitting into training and testing sets
Model evaluation with accuracy, f1-score, ROC scores
Apply models like GBM, MLP, and SVM
Results

Fig. 1. Flowchart for Predicting Diabetes Mellitus

In the following Figure 1 flowchart for Predicting Diabetes Mellitus, each graphic phase is briefly described.

3.1 Data Collection

The dataset, which comprises 520 people's reports of diabetes-related symptoms, was gathered from the diabetes dataset's UCI machine repository. It includes personal information about people, such as signs of diabetes. The preparation step involved extensive data quality tests on the dataset. The following visualization graph of the dataset are listed in below:

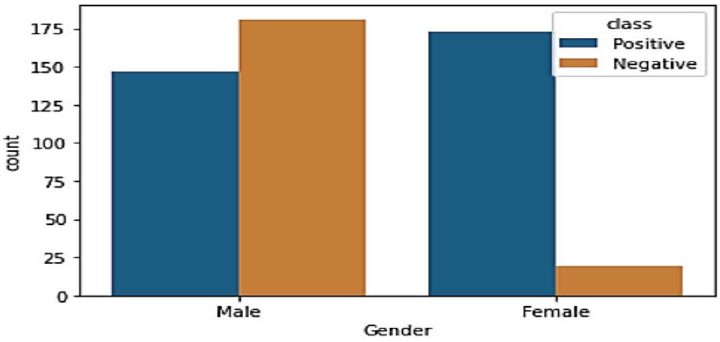

Fig. 2. Count plot for gender

The count plot in Figure 2 depicts the distribution of two classes (Positive and Negative) based on gender (Male and Female). For Males, compared for females, the Positive class far outnumbers the Negative class, with the Negative class having extremely few instances, but the Negative class is more common than the Positive class. This suggests a potential gender-based disparity in the data, where males have a higher representation in the negative class and females are predominantly in the positive class.

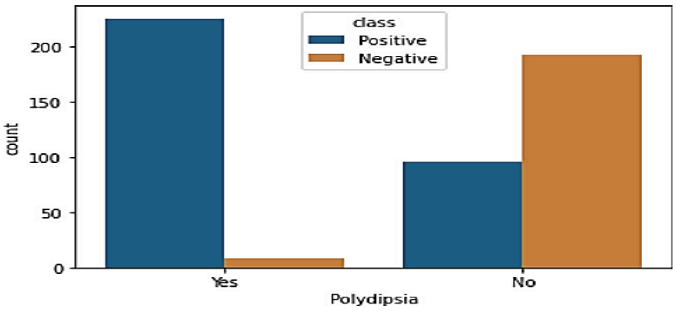

Fig. 3. Count lot for Polydipsia

Figure 3, the count plot shows the relationship between Polydipsia (excessive thirst) and the two classes (Positive and Negative). For individuals with Polydipsia ("Yes"), the Positive class is significantly higher, indicating a strong association between Polydipsia and the Positive class. On the other hand, for those without Polydipsia ("No"), the Negative class has a higher count, suggesting that the absence of Polydipsia is more commonly associated with the Negative class.

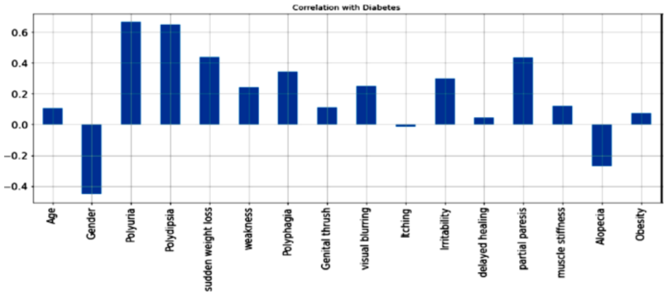

Fig. 4. Correlations of attributes with output class

The bar plot in figure 4 shows that attributes like polyuria, polydipsia, and sudden weight loss have strong positive correlations with diabetes, while age, alopecia, and obesity have weaker or negative correlations. Strongly correlated attributes are likely more important for predicting diabetes, whereas weakly correlated ones may be excluded to enhance model performance. However, correlation doesn't imply causation, so further analysis is needed to confirm any causal relationships.

3.2 Data Preprocessing

Preprocessing is the process by which unstructured data is transformed into intelligible representations suitable for machine-learning models [13]. Preprocessing is largely utilized to improve the quality of input data by minimizing the amount of noise, redundant data, and unnecessary data[14]. This phase of model deals with noise in order to arrive at better and improved results from the original data set which was noisy. This dataset also has some level of missing value present in it. Thus, most values are imputed on the basis of few chosen attributes such as Age, BMI, skin thickness, blood pressure, and glucose level, as well as because some characteristic values cannot be zero. The dataset should then be scaled so that all values fall between 0 and 1. Below are the main pre-processing terminology:

- **Handle missing values:** Deletion involves eliminating the rows that include missing data, whereas imputation involves substituting statistical measures such as mean, median, or model for the missing values.
- **Remove Redundancy:** Data redundancy may result in the need of extra storage space, particularly if that space is costly. So, remove all the redundancy from the dataset.

3.3 Normalization with standard Scaler

Normalisation is a data preparation method that shifts a dataset's features to a common scale to improve machine learning algorithms' accuracy and effectiveness[15]. The Standard Scaler approach, which uses the Z-score normalisation, standardises attributes and creates removing the mean from each value and dividing the result by the standard deviation of the attribute yields a distribution with zero mean and unit variance. A value xi may be changed into x 0 i using Equation 1, where ¯x is the x variable's mean.

$$x_i' = \frac{x_i - \bar{x}}{s} \ldots \ldots (1)$$

The sample mean of the property serves as the translational term in this instance, Although the standard deviation acts as the scaling factor.

3.4 Label Encoder

In data analysis and machine learning, label One technique for converting categorical information into numerical representation is encoding. One very useful technique The Label Encoder is used to transform categorical variables into a numerical representation during data preparation. This is accomplished by giving each input category is assigned a distinct number. A method for converting categorical variables into numerical representation in machine learning and data analysis is called label encoding [16].

3.5 Feature Importance using ETC

Feature importance using the Extra Trees Classifier (ETC) measures how the accuracy of the model's predictions is influenced by each feature. Following dataset training, the classifier gives each feature a significance score that indicates how relevant it is to the classification objective [17]. Higher-scoring features are thought to have more predictive power, whereas lower-scoring features could have less of an effect. This helps in identifying key predictors and potentially reducing the dataset's dimensionality by removing less important features, thereby improving model performance and interpretability.

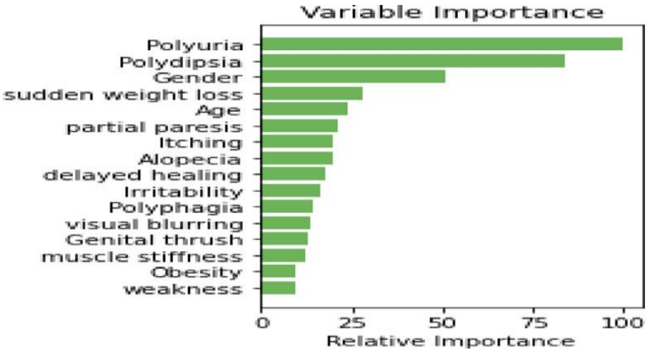

Fig. 5. Important features plot using extra trees classifier

In figure 5, the bar plot illustrates the relative importance of various variables in predicting the target variable of interest, likely the presence or absence of a disease. The y-axis represents the relative importance score, indicating. The variable's impact on the predictive ability of the model. Based on the plot, polyuria and polydipsia emerge as the most important variables, followed by gender and sudden weight loss. Other variables, such as age, partial paresis, and itching, exhibit moderate importance, while variables like obesity and weakness demonstrate relatively lower importance. These findings suggest that polyuria, polydipsia, gender, and sudden weight loss are key factors in determining the disease outcome.

3.6 Data Splitting

Therefore, consequently, the preprocessed data yields two sets: the training set and the testing set. The model is developed on or tested on the training set of 80 percent of the data and the accuracy in tested with the testing set of the 20 percent of data.

3.7 Classification GBM model

Gradient boosting machines, or GBMs, use gradients to determine the shortcomings of weak models[18][19]. This is accomplished by an iterative process that combines decision trees using an additive model, with the ultimate goal being to link base learners to reduce forecast mistakes. [20] while using gradient descent to lower the loss function. The sum of n regression trees (2) is known as the gradient boosting tree, or GBT, $Fn(xt)$.

$$F_n(x_t) = \sum_{i=1}^{n} f_i(x_t) \ldots\ldots\ldots (2)$$

In which each $fi(xt)$ is a regression-tree, or decision tree. The following equation (3) is used to estimate the new decision tree fn+1(xt) to strengthen the group of trees in a sequential manner:

$$argmin \sum_t L\left(y_t.F_n(x_t) + F_{n+1}(x_t)\right) \ldots\ldots (3)$$

When L, the loss-function L, is differentiable. The steepest descent approach is used to accomplish this optimisation [21].

3.8 Evaluation metrics

It is crucial to depict the confusion matrix and look at certain effectiveness indicators when evaluating any data mining classification system. The arrangement of the expected and actual classes is known as the confusion matrix. It displays how many samples fit into each of the model's four quadrants. Interpreting accurately and inaccurately projected model results, such as FN, TN, TP, and F, is made easier with its help. As a result, it is crucial for assessing how effectively the model performed the categorisation. A matrix of confusion is shown below figure (6).

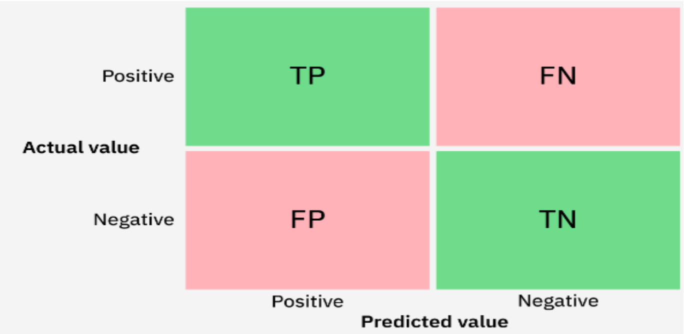

Fig. 6. Representation of confusion matrics

When comparing predicted and actual values, four different columns are produced: These consist of the quantity among false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN). For example, if an instance were predicted to have diabetes and it did not have diabetes, then it is classified as a false positive.

a) Accuracy

It is calculated by dividing the number of accurate predictions by the total number of input samples. It is offered as (4).

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{Fp} + \mathrm{TN} + \mathrm{FN}} \ldots\ldots(4)$$

b) F1 score

It is employed to gauge the correctness of a test. The F1 Score is the average of recall and accuracy. The F1 Score range is [0, 1]. It informs you of the robustness and precision of your classifier. It is expressed mathematically as (5).

$$\mathrm{F1} = \frac{2 * (\mathrm{precision} * \mathrm{recall})}{\mathrm{precision} + \mathrm{recall}} \ldots.\ldots\ldots\ldots(5)$$

c) ROC and AUC Score

ROC is the abbreviation for receiver operating characteristic and it is a probability curve which has been plotted with the FPR on X coordinate and TPR on Y coordinate. The quality of a binary classifier is summarized quantitatively by an ROC graph. In conclusion the ROC curve is expressed in terms of the area under the ROC curve abbreviated as AUC. For the model utilized, the greater value of AUC is desired. AUC is an evaluation measure and its maximum or perfect value is normative and contains a value of 1 always. In the meanwhile, the AUC of random classifier is 0.5.

Results and Discussion

The experiment result of the models is provided in this section. The following results are measured on f1-score, accuracy and ROC-AUC score. For the comparative analysis, use machine learning models like MLP[22], SVM[23], and GBM. The following Table 2 provides the performance of the GBM model with graphical results including confusion matrix, ROC, and AUC graphs.

Table II. GBM model performance for predicting Diabetes Mellitus on diabetes dataset

Performance matrix	Gradient Boosting Machine
Accuracy	96.92
F1-score	99.37
ROC-AUC	99.92

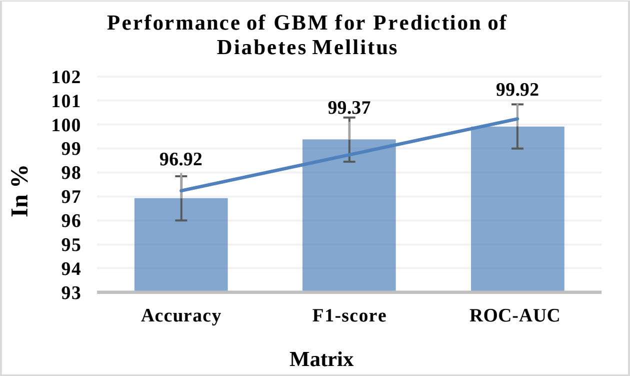

Fig. 7. GBM model performance on diabetes dataset

The following table 2 and Figure 7 show the GBM model performance data. In this figure, the GBM model for diabetes prediction showcases outstanding performance, achieving an accuracy of 96.92%. With an F1-score of 99.37, it effectively balances precision and recall, minimizing false positives and negatives. Additionally, a remarkable ROC-AUC score of 99.92 highlights the model's exceptional ability to differentiate between diabetic and non-diabetic cases, demonstrating its robustness for real-world applications in early diabetes detection.

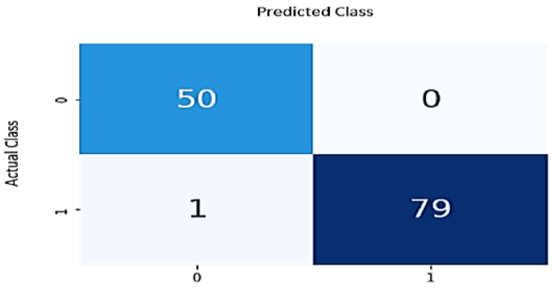

Fig. 8. Confusion Matrix for GBM model

Figure 8 illustrates the confusion matrix for the GBM model and its classification performance, revealing that the model accurately classified 50 instances of class 0 (TP) and 79 instances of class 1 (TN) while making only 1 FP and no FN. With a total of 129 correct predictions out of 130 instances, the model demonstrates strong performance. However, to gain a more comprehensive understanding of its efficacy, further evaluation using additional metrics such as precision, F1-score recall, and accuracy is recommended.

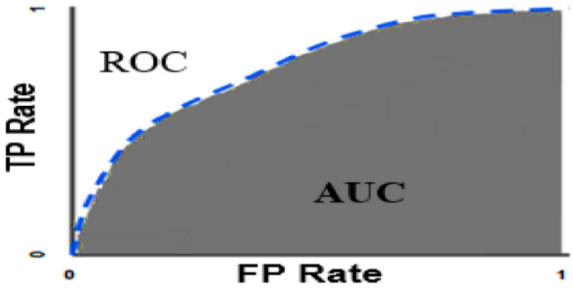

Fig. 9. ROC-AUC curve for the GBM model

The ROC curve for the GBM model in figure 9 illustrates its binary classification performance by plotting the TPR against the FPR at different thresholds. Improved discrimination is shown by a curve towards the upper-left corner, while the AUC quantifies this performance, with values close to 1 reflecting strong classification ability. In this case, the ROC curve suggests the GBM model performs well, likely achieving an AUC near 1, indicating effective identification of both positive and negative instances with a low FPR.

Table III. Accuracy Comparison between machine learning models on diabetes dataset

Model	Accuracy
Gradient Boosting Machine (GBM)	96.92
Multi-layer perceptron (MLP)	76
Support Vector Machine (SVM)	**77.73**

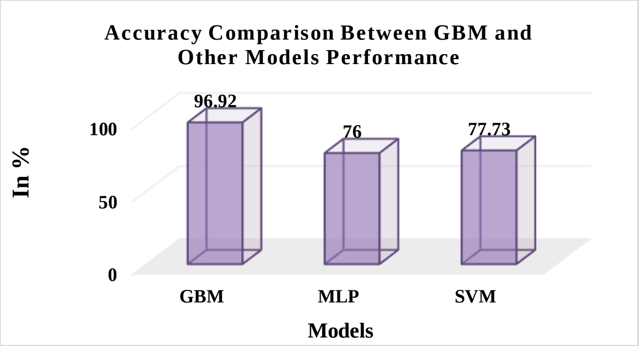

Fig. 10. Accuracy comparison between model performance

Figure 10 illustrates the accuracy comparison of models. in this comparison, The GBM outperforms both the MLP and the SVM in terms of diabetes prediction accuracy. The GBM achieves an impressive accuracy of 96.92%, significantly higher than the MLP, which reaches 76%, and the SVM, which achieves 77.73%. This comparison highlights the superior performance of GBM in handling this classification task, making it more reliable for accurate predictions compared to MLP and SVM models.

Conclusion and Future study

Diabetes Mellitus (DM) is a severe condition that affects a lot of individuals worldwide. Given the high incidence of diabetes mellitus, its detrimental effects on health, and the rising expenses of care and treatment, prevention, early identification, and better disease management are imperative. This study uses a dataset to illustrate the efficiency of machine learning algorithms, taken from the UCI Machine Learning Repository in diabetes mellitus prediction, underscoring the importance of careful data preparation and feature selection. According to the investigation, the Gradient Boosting Machine (GBM) performs noticeably better than other classifiers, including MLP and SVM, with an astounding 96.92% accuracy and f1-score of 99.37. The identification of key predictors, including polyuria and polydipsia, underscores the relevance of these symptoms in diabetes risk assessment. The robust performance of the GBM model, coupled with its interpretability through feature importance analysis, emphasizes its potential as a reliable tool for clinical decision-making in diabetes management. The findings advocate for the integration of machine learning methodologies in healthcare settings to facilitate early diagnosis and individualised therapeutic approaches. In order to further improve predicted accuracy, future studies should examine how well these models work in a variety of groups and think about adding other clinical characteristics.

References

1. M. H. Tanrıverdi, T. Çelepkolu, and H. Aslanhan, “Diabetes mellitus and primary healthcare,” J. Clin. Exp. Investig., vol. 4, no. 4, Dec. 2013, doi: 10.5799/ahinjs.01.2013.04.0347.
2. A. Petersmann et al., “Definition, classification and diagnostics of diabetes mellitus,” J. Lab. Med., 2018, doi: 10.1515/labmed-2018-0016.
3. S. C. R. Vennapusa, T. Fadziso, K. Sachani, V. K. Yarlagadda, and S. K. R. Anumandla, “Cryptocurrency-Based Loyalty Programs for Enhanced Customer Engagement,” Technol. Manag. Rev., vol. 3, no. 1, pp. 46–62, 2018.
4. Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques,” Front. Genet., 2018, doi: 10.3389/fgene.2018.00515.
5. A. K. Steck et al., “Predictors of slow progression to diabetes in children with multiple islet autoantibodies,” J. Autoimmun., 2016, doi: 10.1016/j.jaut.2016.05.010.
6. K. Mullangi, N. D. Vamsi Krishna Yarlagadda, and M. Rodriguez, “Integrating AI and Reciprocal Symmetry in Financial Management: A Pathway to Enhanced Decision-Making,” Int. J. Reciprocal Symmetry Theor. Phys., vol. 5, no. 1, pp. 42–52, 2018.

7. A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, “A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques,” in 1st International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC 2019 - Proceedings, 2019. doi: 10.1109/UBMYK48245.2019.8965556.
8. A. Gnana, E. Leavline, and B. Baig, “Diabetes Prediction Using Medical Data,” J. Comput. Intell. Bioinforma., 2017.
9. A. Agarwal and A. Saxena, “Analysis of machine learning algorithms and obtaining highest accuracy for prediction of diabetes in women,” in Proceedings of the 2019 6th International Conference on Computing for Sustainable Global Development, INDIACom 2019, 2019.
10. M. T. Islam, M. Raihan, F. Farzana, M. G. M. Raju, and M. B. Hossain, “An Empirical Study on Diabetes Mellitus Prediction for Typical and Non-Typical Cases using Machine Learning Approaches,” in 2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019, 2019. doi: 10.1109/ICCCNT45670.2019.8944528.
11. M. H. Tanrıverdi, T. Çelepkolu, and H. Aslanhan, “Diabetes mellitus and primary healthcare,” J. Clin. Exp. Investig., vol. 4, no. 4, Dec. 2013, doi: 10.5799/ahinjs.01.2013.04.0347.
12. A. Petersmann et al., “Definition, classification and diagnostics of diabetes mellitus,” J. Lab. Med., 2018, doi: 10.1515/labmed-2018-0016.
13. S. C. R. Vennapusa, T. Fadziso, K. Sachani, V. K. Yarlagadda, and S. K. R. Anumandla, “Cryptocurrency-Based Loyalty Programs for Enhanced Customer Engagement,” Technol. Manag. Rev., vol. 3, no. 1, pp. 46–62, 2018.
14. Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques,” Front. Genet., 2018, doi: 10.3389/fgene.2018.00515.
15. A. K. Steck et al., “Predictors of slow progression to diabetes in children with multiple islet autoantibodies,” J. Autoimmun., 2016, doi: 10.1016/j.jaut.2016.05.010.
16. K. Mullangi, N. D. Vamsi Krishna Yarlagadda, and M. Rodriguez, “Integrating AI and Reciprocal Symmetry in Financial Management: A Pathway to Enhanced Decision-Making,” Int. J. Reciprocal Symmetry Theor. Phys., vol. 5, no. 1, pp. 42–52, 2018.
17. A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, “A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques,” in 1st International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC 2019 - Proceedings, 2019. doi: 10.1109/UBMYK48245.2019.8965556.
18. A. Gnana, E. Leavline, and B. Baig, “Diabetes Prediction Using Medical Data,” J. Comput. Intell. Bioinforma., 2017.
19. A. Agarwal and A. Saxena, “Analysis of machine learning algorithms and obtaining highest accuracy for prediction of diabetes in women,” in Proceedings of the 2019 6th International Conference on Computing for Sustainable Global Development, INDIACom 2019, 2019.
20. M. T. Islam, M. Raihan, F. Farzana, M. G. M. Raju, and M. B. Hossain, “An Empirical Study on Diabetes Mellitus Prediction for Typical and Non-Typical Cases using Machine Learning Approaches,” in 2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019, 2019. doi: 10.1109/ICCCNT45670.2019.8944528.
21. M. Kowsher, M. Y. Turaba, T. Sajed, and M. M. Mahabubur Rahman, “Prognosis and treatment prediction of type-2 diabetes using deep neural network and machine learning classifiers,” in 2019 22nd International Conference on Computer and Information Technology, ICCIT 2019, 2019. doi: 10.1109/ICCIT48885.2019.9038574.
22. T. C. Seng et al., “Predicting high cost patients with type 2 diabetes mellitus using hospital databases in a multi-ethnic Asian population,” in 3rd IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2016, 2016. doi: 10.1109/BHI.2016.7455879.
23. S. Vijayarani, M. Ilamathi, and M. Nithya, “Preprocessing Techniques for Text Mining-An Overview Privacy Preserving Data Mining View project,” Int. J. Comput. Sci. Commun. Networks, 2015.
24. V. K. Y. Nicholas Richardson, Rajani Pydipalli, Sai Sirisha Maddula, Sunil Kumar Reddy Anumandla, “Role-Based Access Control in SAS Programming: Enhancing Security and Authorization,” Int. J. Reciprocal Symmetry Theor. Phys., vol. 6, no. 1, pp. 31–42, 2019.
25. R. P. Vamsi Krishna Yarlagadda, “Secure Programming with SAS: Mitigating Risks and Protecting Data Integrity,” Eng. Int., vol. 6, no. 2, pp. 211–222, 2018.
26. Z. Lin, G. Ding, J. Han, and L. Shao, “End-to-End Feature-Aware Label Space Encoding for Multilabel Classification with Many Classes,” IEEE Trans. Neural Networks Learn. Syst., 2018, doi: 10.1109/TNNLS.2017.2691545.
27. A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, “Permutation importance: A corrected feature importance measure,” Bioinformatics, 2010, doi: 10.1093/bioinformatics/btq134.
28. V. V. Kumar, A. Sahoo, and F. W. Liou, “Cyber-enabled product lifecycle management: A multi-agent framework,” in Procedia Manufacturing, 2019. doi: 10.1016/j.promfg.2020.01.247.

29. V. V. Kumar, F. T. S. Chan, N. Mishra, and V. Kumar, "Environmental integrated closed loop logistics model: An artificial bee colony approach," in SCMIS 2010 - Proceedings of 2010 8th International Conference on Supply Chain Management and Information Systems: Logistics Systems and Engineering, 2010.
30. V. V. Kumar and F. T. S. Chan, "A superiority search and optimisation algorithm to solve RFID and an environmental factor embedded closed loop logistics model," Int. J. Prod. Res., 2011, doi: 10.1080/00207543.2010.503201.
31. T. Chen et al., "Prediction of Extubation Failure for Intensive Care Unit Patients Using Light Gradient Boosting Machine," IEEE Access, 2019, doi: 10.1109/ACCESS.2019.2946980.
32. A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," in Procedia Computer Science, 2019. doi: 10.1016/j.procs.2020.01.047.
33. N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," J. Big Data, 2019, doi: 10.1186/s40537-019-0175-6.

CITATION
Chandrakanth R. M., Eswar P. G., Mohit S. R., Manikanth S., Venkata N. B., & Siddharth K. (2021). Predicting Diabetes Mellitus in Healthcare: A Comparative Analysis of Machine Learning Algorithms on Big Dataset. In Global Journal of Research in Engineering & Computer Sciences (Vol. 1, Number 1, pp. 1–11). https://doi.org/10.5281/zenodo.14010835