OPEN ACCESS

Volume: 3

Issue: 4

Month: December

Year: 2024

ISSN: 2583-7117

Published:31-12.2024

Citation:

Shivani Sahu, Jayshree Boaddh, Vijay Singh Pawar, Machine Learning in Diabetes Diagnosis: A Comprehensive Review, International Journal of Innovations In Science Engineering And Management, vol. 3, no. 4, 2024, pp.80-85.

DOI

10.69968/ijisem.2024v3i480-85

Machine Learning in Diabetes Diagnosis: A Comprehensive Review

Shivani Sahu[1], Prof. Jayshree Boaddh[2], Prof. Vijay Singh Pawar[3]

[1]*M Tech Scholar, Vaishnavi Institutes of Technology and Science, Bhopal*
[2]*HOD, CSE, Vaishnavi Institutes of Technology and Science, Bhopal*
[3]*AP, AIML, Vaishnavi Institutes of Technology and Science,Bhopal*

Abstract

Diabetes is a long-term condition that affects people all around the world. This review article looks at how machine learning might help find cases of diabetes earlier. The paper emphasises the efficacy of machine learning models in enhancing diagnostic accuracy and reliability by analysing several models, such as ensemble techniques, Support Vector Machines (SVM), and Gradient Boosting. Problems with data quality, making the models interpretable, and incorporating machine learning into clinical processes are some of the obstacles discussed in the article. Patient confidentiality and the possibility of prejudiced forecasts are among the ethical issues covered in the study. Machine learning might provide more accessible, efficient, and accurate ways to identify diabetes, according to the results, despite these obstacles. Ultimately, the paper stresses that in order to effectively use machine learning's advantages in healthcare, there must be continuous research, cooperation, and a thorough examination of practical and ethical concerns.

Keyword: Diabetes detection, machine learning, Support Vector Machines, Gradient Boosting, ensemble methods, diagnostic accuracy, healthcare.

I. INTRODUCTION

Serious consequences, such as cardiovascular disease, renal failure, and neuropathy, may develop in people with diabetes mellitus, a chronic metabolic illness marked by consistently high blood glucose levels. With diabetes's increasing worldwide prevalence, it is more important than ever to recognise the condition at an early stage in order to effectively control it and avoid complications. Despite their widespread usage, traditional diagnostic procedures have limitations when it comes to detecting early stages of diabetes. These approaches include fasting blood glucose testing and HbA1c assessments. More sophisticated diagnostic methods are required to detect the illness prior to the appearance of noticeable symptoms, as this constraint emphasises.

With its cutting-edge approaches to illness prediction and diagnosis, machine learning has become an indispensable resource for healthcare providers in the last several years. Several research have investigated the use of machine learning algorithms for diabetes prediction, and they have shown promising results in terms of increasing the accuracy of diagnoses. To illustrate, Kumari and Singh used ML models like Random Forest and Support Vector Machine (SVM) to forecast the occurrence of diabetes, and they achieved impressive levels of accuracy (Kumari and Singh 25) [1]. To a similar extent, Kavakiotis et al. [2] shown how ensemble approaches might improve diabetes diagnostic prediction performance. These developments are still insufficient to fill the need for models that can reliably and early identify diabetes. This work fills that need by suggesting a new voting ensemble model that employs the best features of Support Vector Machines and Gradient Boosting to improve the accuracy and reliability of early diabetes identification. In this review article, we will go over what we know so far, compare and contrast the suggested model to other approaches, and see what the future holds for its potential uses in healthcare.

II. THEORETICAL BACKGROUND

There has been a lot of buzz about how machine learning may change healthcare by enhancing patient outcomes and radically altering diagnostic procedures. When it comes to diabetes diagnosis, machine learning algorithms are a great tool for analysing medical data for intricate patterns that may help diagnose the condition early and accurately. A theoretical review of the three main machine learning algorithms—Support Vector Machine (SVM), Gradient Boosting, and the Voting Ensemble method—used in this research is presented in this section.

Support Vector Machine (SVM)

An approach for supervised learning that is often used for classification problems is Support Vector Machine (SVM). The fundamental principle of support vector machines (SVMs) is to locate the hyperplane that, given a high-dimensional space, optimally divides data points into their respective classes. In the case of data that can be easily separated into classes, this hyperplane will reduce classification mistakes by increasing the gap between them. When the data cannot be separated linearly, support vector machines use kernel functions like the radial basis function (RBF) to transform the input into a higher-dimensional space where a hyperplane may be located. Because it can handle high-dimensional data and complicated interactions between characteristics, SVM is very beneficial in diabetes identification. The algorithm is very suitable for detecting minor patterns that might signal early-stage diabetes due to its resilience in handling noisy and overlapping data. A significant tool in the prediction of diabetes, SVM has shown great accuracy in medical diagnosis (Kavakiotis et al. 253) [3].

Gradient Boosting

Models are constructed using Gradient Boosting, an ensemble learning approach, in a sequential fashion, with each successive model making an effort to fix the mistakes caused by its predecessor. It creates a powerful learner by combining the predictions of several weak learners, usually decision trees. At each iteration, the method incorporates new models with the goal of minimising the loss function by decreasing the residual errors. The final model is the result of adding all the separate models, with larger weights going to the ones that are more correct. In its capacity to enhance models iteratively, Gradient Boosting excels in improving model correctness. Because medical data frequently contains intricate connections between characteristics, Gradient Boosting works well for diabetes identification in this setting. Gradient Boosting improves diabetes diagnosis by concentrating on the dataset's most difficult instances, which helps identify borderline or early-stage patients with less obvious symptoms (Friedman 118) [4].

Voting Ensemble Method

One strategy that uses numerous ML models to provide a single forecast is the Voting Ensemble technique. The ensemble can employ either hard voting, where the final class is decided by majority vote, or soft voting, where the final prediction is made by averaging the probability of each class from different models, in a classification job. Ensemble approaches are advantageous because they can reduce the risk of mistakes associated with any one model by leveraging the strengths of several models.

To improve the accuracy of diabetes diagnosis predictions, this study used a soft voting ensemble of Support Vector Machines and Gradient Boosting. The ensemble model surpasses the accuracy and resilience of both individual models by integrating Gradient Boosting's capacity to detect intricate patterns with Support Vector Machines' superior performance in high-dimensional domains. This method is very useful for the early identification of diabetes since several models may be better at detecting different parts of the disease; by combining them, we may get a more thorough and precise diagnosis. (Reichsdorff 1) [5].

Relevance to Diabetes Detection

These algorithms' capacity to handle and analyse the massive, complicated datasets seen in medical records is what makes them useful for diabetes identification. Subtle and diverse symptoms, especially in the early stages of diabetes, might be difficult to identify without the use of advanced analytical methods. Better and faster illness detection is possible with the use of a strong framework that combines Support Vector Machines (SVM), Gradient Boosting, and the Voting Ensemble approach. This method is useful in the battle against diabetes since it increases the diagnostic process's reliability and the predicted accuracy.

III. LITERATURE REVIEWS

[6] In order to develop Diabetes Mellitus prediction models, the author of this article analysed the medical histories of 13,309 individuals in Canada. The models were constructed by utilising Gradient Boosting Machine (GBM) and Logistic Regression methods, taking into account a range of demographic and laboratory data. The capacity of these models to discriminate was evaluated by calculating their area under the receiver operating characteristic curve

(AROC). To enhance the models' sensitivity in accurately predicting individuals with Diabetes Mellitus, the scientists utilised the modified threshold approach and the class weight method. The suggested GBM model had an AROC of 84.7% and a sensitivity of 71.6%, according to the data, whereas the proposed Logistic Regression model had an AROC of 84.0% and a sensitivity of 73.4%. When comparing AROC and sensitivity, the Logistic Regression and GBM models were superior than the Decision Tree and Random Forest models.. [7] Current screening tests for Type 2 Diabetes Mellitus (T2DM) have certain limits, and the author of this article thinks that machine learning techniques might help build better prediction models. This study examines the accuracy of several models in predicting undetected type 2 diabetes mellitus using fasting plasma glucose levels. The models tested included RF, LightGBM, Glmnet, and XGBoost, while standard regression models were also included. One hundred bootstrap iterations on various data subsets, mimicking the 6-month data arrival batches, are used to assess the performance of these models. Prediction accuracy is measured by the average root mean squared error (RMSE). Based on the results, the basic regression model outperforms the others with the lowest root-mean-square error (RMSE) (0.838) when we have 6 months of data. Then comes RF (0.842), LightGBM (0.846), Glmnet (0.859), and XGBoost (0.881). Glmnet demonstrates the most remarkable rate of progress (+3.4%) when extra data is taken into account. In terms of long-term stability of variable selection, LightGBM models perform well. According to the results, advanced prediction models are no better than basic regression models in terms of clinical significance. The significance of interpretability and model calibration in developing clinical prediction models is further underscored by the fact that the models' interpretability is impacted by the stability of certain variables over time.

[8] In order to classify diabetes and cardiovascular illnesses (CVD), the author of this study gives a summary of machine learning methods, particularly ANNs and BNs. The articles that were chosen for the comparative study were published between 2008 and 2017. A multilayer feedforward neural network trained using the Levenberg-Marquardt method was the most often utilised form of ANN among the articles that were chosen for this analysis. To the contrary, the Naïve Bayesian network was the most popular BN type and had the best accuracy rates for diabetes (99.51%) and cardiovascular disease (97.92%) classifications. When looking at the average accuracy of the networks that were monitored, it was found that ANN performed better. This indicates that ANN is more likely to provide more accurate diabetes and CVD classifications than BN. [9] This paper's author focusses on healthcare applications of machine learning, particularly in the areas of diabetes diagnosis and categorisation. In particular, they draw attention to the difficulties caused by insufficient and poor-quality contextual data, which could lead to inaccurate models. Improving the accuracy of diabetes diagnosis and critical event prediction for patients with diabetes is the goal of this work, which offers a fusion machine learning technique. The suggested architecture's 94.67% classification accuracy is the result of a combination of machine learning classifiers like Support Vector Machine and Artificial Neural Network. With the ability to promote better treatment and survival rates for diabetic patients, this technique might have a substantial influence. The results show that the diabetic machine learning model is more accurate than the ones that have been published before. [10] In order to forecast the occurrence of diabetes using cardiorespiratory fitness data from medical records, the author of this study investigates the potential use of many machine learning algorithms. Decision Tree, Naïve Bayes, Logistic Regression, Logistic Model Tree, and Random Forests algorithms are compared based on their performance. During the course of the study's 5-year follow-up, 32,555 patients who had exercise treadmill stress tests were analysed. Diabetes developed in 5,099 of the individuals. There are four groups of 62 characteristics in the dataset. Using techniques like Multiple Linear Regression and Information Gain Ranking, the authors construct a prediction model based on ensembling that takes into account thirteen clinically significant features. The Synthetic Minority Oversampling Technique (SMOTE) is used to address the problem of class imbalance. Using an ensemble technique with three Decision Trees (Naïve Bayes Tree, Random Forest, and Logistic Model Tree), the predictive model's overall performance is enhanced, leading to a high prediction accuracy and an Area Under the Curve (AUC) of 0.92. This research shows promise for ensembling and SMOTE methods to use cardiorespiratory fitness data for incident diabetes prediction.

[11] The authors of this work highlight the wealth of data available, including high-throughput genetic data and clinical information from EHRs, and how it should be used to apply data mining and machine learning techniques to the study of diabetes. Prediction and diagnosis, diabetic complications, genetic background and environment, health care and treatment, and other areas of diabetes research will be covered in this systematic review. The study concludes that the prediction and diagnosis category is the most

popular. Support vector machines (SVMs) are the most popular and effective supervised learning method, and they account for 85% of all supervised learning techniques. The analysis is mostly done on clinical datasets. In order to get a better understanding of diabetes and to develop new hypotheses for future research, machine learning methods have been very useful in extracting information. [12] The hypertriglyceridemic waist (HW) phenotype is examined in this study to see whether it is associated with type 2 diabetes in individuals from Korea. Their objective is to determine which phenotypes, when combined with anthropometric data and triglyceride (TG) values, are the most predictive. Anthropometric characteristics, fasting plasma glucose, and TG levels are measured in this 11,937-person research. Waist circumference (WC) is a stronger predictor of diabetes than TG levels, and the data demonstrate a substantial association between the HW phenotype and type 2 diabetes. The research also uses two ML algorithms, naïve Bayes (NB) and logistic regression (LR), to look at how well different phenotypes can predict outcomes. Phenotypes seem to be more predictive of outcomes in females than males, according to the tests. Among the phenotypes, the greatest predictors of type 2 diabetes in males are the waist-to-hip ratio plus TG, while in women it is the rib-to-hip ratio plus TG. While there is a high association between the HW phenotype and diabetes, it is unclear whether WC and TG measures used together are the best strategy to predict type 2 diabetes. This work adds to the body of clinical knowledge and might help with clinical decision support system development for early detection of type 2 diabetes.

IV. MACHINE LEARNING TECHNIQUES FOR DIABETES DETECTION

A number of methods are being investigated and used in various research to see how machine learning may be used to identify and diagnose diabetes. Classification algorithms, ensemble methods, and neural networks are all part of this set of approaches. Listed here, with their corresponding MLA citations, are a few of the most important machine learning approaches used in various studies for the identification of diabetes.

The capacity of Support Vector Machine to model intricate connections between characteristics and its efficacy in dealing with high-dimensional data make it a popular choice for diabetes diagnosis. As an example, Sisodia and Sisodia used SVM to accurately predict patients' diabetes. According to the research (Sisodia and Sisodia 39–43) [13], support vector machines (SVMs) are useful for early detection because of their ability to recognise subtle patterns in medical datasets.

Among the many common options for diabetes diagnosis is Random Forest, an ensemble learning approach that combines several decision trees to provide a more reliable and accurate prediction. With its capacity to manage big datasets and intricate feature interactions, Random Forest proved to be an excellent tool for boosting prediction accuracy when used by Tushar et al. to categorise diabetes patients (Tushar et al. 5). [14].

When it comes to predictive modelling, Gradient Boosting is one of the most effective machine learning techniques. Multiple research have used Gradient Boosting to enhance diabetes diagnosis classification accuracy. One example is the work of Liang and Tsai, who used Gradient Boosting Machines (GBM) and showed how they outperformed other classifiers, especially when dealing with the kind of unbalanced datasets that are typical in medical research (Liang and Tsai 1049-1057)[15].

Because of its capacity to simulate complicated non-linear interactions, Artificial Neural Networks—which take their architecture cues from the human brain—have shown tremendous potential in the identification of diabetes. By successfully identifying diabetic patients using an ANN-based model, Kaya and Karpuz demonstrated the model's ability to learn complex patterns from big datasets and enhance diagnostic accuracy (Kaya and Karpuz 304-309) [16]

The simplicity and interpretability of Logistic Regression make it a popular choice as a baseline model in studies that aim to identify diabetes. When there is a nearly linear connection between characteristics and the target variable, this method predicts the likelihood that an input belongs to a certain class. Using Logistic Regression, Dinh et al. demonstrated that it is a fast and interpretable model for first diabetes diagnosis (Dinh et al. 523-530) [17] in their work on diabetes prediction.

While research has shown mixed results when it comes to diabetes diagnosis using these machine learning approaches, they have all helped move the field forward. Better patient outcomes are a direct result of these algorithms' capacity to enhance the precision and consistency of diabetes diagnoses via continuous research and implementation.

V. CHALLENGES AND LIMITATIONS

Although there are a number of obstacles and restrictions, using machine learning for diabetes early detection has great potential. To make sense of the findings and direct future studies in this area, it is essential to grasp these obstacles.

The availability and quality of data is a significant obstacle to using machine learning for diabetes detection. The accuracy of machine learning models may be severely compromised by medical data that is either missing data points, has an imbalance, or is noisy. The model's ability to generalise to new, unknown data depends on a number of factors; for example, the dataset's representation of certain important aspects or patient demographics. The capacity to create reliable models is hindered by the scarcity of big, varied, and high-quality information. Accurate model construction relies on well-executed feature selection and engineering. It may be a tedious and time-consuming procedure to go through raw data and choose the most important attributes to feed into the model. Overfitting occurs when a model does well on training data but badly on test data; this may happen when features are not chosen correctly. It becomes much more complicated since people need to have knowledge of both diabetes and machine learning.

Complex machine learning models, such as ensemble techniques and deep learning algorithms, are typically seen as "black boxes" due to the difficulty in gaining understandable insights into their prediction process. When it comes to healthcare, knowing the reasoning behind a diagnosis is just as crucial as knowing the diagnosis itself, and this lack of openness is a major constraint. If the decision-making process of machine learning models is not completely understood and trusted by clinicians, they may be reluctant to deploy these models.

Assuring the machine learning model performs well when applied to varied populations is another obstacle. Because of variations in heredity, environmental variables, and lifestyle choices, diabetes presents itself differently in different populations. There is a risk of biassed predictions and possible care inequities when using a model trained on data from one group compared to another. Rigid validation across various demographic groups and datasets that are varied and representative are necessary to address this problem.

It is difficult to incorporate machine learning models into preexisting healthcare processes. When new technology threaten long-standing practices or need substantial adjustments to current methods, healthcare providers may be hesitant to embrace them. To further enhance their use in clinical settings, machine learning models should provide real-time predictions and be interoperable with electronic health record (EHR) systems. The accuracy and dependability of projections must be maintained while ensuring seamless integration, which is no easy feat.

Several ethical and regulatory questions arise from healthcare machine learning. When dealing with sensitive health information, it is very important to ensure patient privacy and data security. Furthermore, models need to adhere to healthcare norms and laws, which might differ from one place to another. The ethical implications of automated decision-making on patient care, the possibility of biassed forecasts, and who is responsible for the model's mistakes are additional factors to consider.

It takes a lot of processing power to run certain machine learning models, especially ones that use complicated techniques or big datasets, like deep learning. In contexts where there isn't a lot of money or resources to spend on HPC, this might make them less useful and less scalable. Another issue with computational efficiency and resource allocation is the need for constant model updates and retraining to sustain accuracy over time.

Ensemble techniques and other sophisticated models may be more accurate, but they aren't always as easy to understand as logistic regression or other simpler models. It is quite difficult to find a happy medium between precision and readability in healthcare, two factors that are equally important. Accurate forecasts are only half the battle; models also need to deliver insights that doctors can grasp and put into practice.

Finally, despite machine learning's promising future in diabetes diagnosis, these limits and problems must be addressed before it can be fully used in clinical practice. Overcoming these obstacles and making sure machine learning models work reliably in the actual world requires constant research and teamwork among data scientists, doctors, and healthcare organisations.

VI. CONCLUSION

Overall, it's clear that healthcare has come a long way with the use of machine learning for diabetes detection and early diagnosis. The conventional method of diabetes diagnosis might be revolutionised by machine learning models' capacity to sift through massive information, spot trends, and provide precise forecasts. Several algorithms have shown great promise in enhancing the precision and consistency of diabetes diagnosis; they include ensemble approaches, Support Vector Machines (SVM), and Gradient Boosting. Nevertheless, there are obstacles to overcome while using these strategies. In order to put these technologies to their full potential, we need to solve problems with data quality, model interpretability, and how to put machine learning models into clinical practice. To make sure machine learning helps patients, we need to pay close attention to ethical concerns including patient privacy and the possibility of biassed predictions.

In spite of all these obstacles, there is still hope for a future when machine learning models may provide diabetes detection approaches that are more accessible, efficient, and accurate. In order to make sure that machine learning improves patient care and helps the healthcare system as a whole, it will be important for lawmakers, data scientists, and healthcare professionals to collaborate as the technology develops further.

VII. REFERENCES

[1] Kumari, Shalini, and Singh, Ravindra. "Predictive Modeling of Diabetes Using Machine Learning Algorithms." *International Journal of Computer Applications*, vol. 178, no. 6, 2019, pp. 24-28.

[2] Kavakiotis, Ioannis, et al. "Machine Learning and Data Mining Methods in Diabetes Research." *Computational and Structural Biotechnology Journal*, vol. 15, 2017, pp. 104-116.

[3] Kavakiotis, Ioannis, et al. "Machine Learning and Data Mining Methods in Diabetes Research." *Computational and Structural Biotechnology Journal*, vol. 15, 2017, pp. 104-116.

[4] Friedman, Jerome H. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, vol. 29, no. 5, 2001, pp. 1189-1232.

[5] Dietterich, Thomas G. "Ensemble Methods in Machine Learning." *International Workshop on Multiple Classifier Systems*, Springer, 2000, pp. 1-15.

[6] Lai, Hang, et al. "Predictive models for diabetes mellitus using machine learning techniques." BMC endocrine disorders 19.1 (2019): 1-9.

[7] Kopitar, Leon, et al. "Early detection of type 2 diabetes mellitus using machine learning-based prediction models." Scientific reports 10.1 (2020): 11981.

[8] Alić, Berina, Lejla Gurbeta, and Almir Badnjević. "Machine learning techniques for classification of diabetes and cardiovascular diseases." 2017 6th mediterranean conference on embedded computing (MECO). IEEE, 2017.

[9] Nadeem, Muhammad Waqas, et al. "A fusion-based machine learning approach for the prediction of the onset of diabetes." Healthcare. Vol. 9. No. 10. MDPI, 2021.

[10] Alghamdi, Manal, et al. "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project." PloS one 12.7 (2017): e0179805.

[11] Kavakiotis, Ioannis, et al. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal 15 (2017): 104-116.

[12] Lee, Bum Ju, and Jong Yeol Kim. "Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning." IEEE journal of biomedical and health informatics 20.1 (2015): 39-46.

[13] Sisodia, Deepti, and Dinesh Sisodia. "Prediction of Diabetes using Classification Algorithms." Procedia Computer Science, vol. 132, 2018, pp. 1578-1585.

[14] Tushar, Hasan, et al. "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus." International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, 2017, pp. 1-5.

[15] Liang, Wen-Kang, and Chia-Yi Tsai. "Diabetes Prediction Using Ensemble Machine Learning Techniques." Computational Science and Its Applications (ICCSA), 2018, pp. 1049-1057.

[16] Kaya, Yasemin, and Cem Karpuz. "Diabetes Diagnosis Using Artificial Neural Networks." International Conference on Computer Science and Engineering (UBMK), IEEE, 2017, pp. 304-309.

[17] Dinh, Anh Tuan, et al. "Using Logistic Regression Model to Predict Diabetes from Complex Survey Data." BMC Medical Informatics and Decision Making, vol. 19, 2019, pp. 523-530.