International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

Prediction of Diabetes Using Machine Learning and Deep Learning Approaches: A Survey

Preetha Rajagopalam.M [1*], Dr.Anuratha.V[2] and Dr.Elamparithi.M[3]
Department Of Computer Science, Kamalam College of Arts and Science, Anthiyur,
Bharathiar University, Coimbatore, Tamil Nadu, India.
E-Mail: [1*]preethamuttikkal@gmail.com, [2]profanuratha@gmail.com, [3]profelamparithi@gmail.com

Abstract: One of the most prevalent and severe illnesses in the world today is diabetes. In addition to being bad for the blood, it also leads to several illnesses that kill many people each year, including blindness, kidney problems, and heart problems. Therefore, a system that can precisely identify people with diabetes utilizing their medical information needs to be developed. Numerous traditional methods exist for monitoring the health of people with diabetes. Patients must attend a diagnostic facility, speak with their doctor, and wait for some time to get the findings of the typical screening process. While several techniques have been developed over the last few years to identify diabetes, approaches such as machine learning (ML) and deep learning (DL) provide more informative outcomes. This paper reviewed all diabetes predictions based on ML and DL approaches. Furthermore, to create optimal solutions for diabetes detection and prediction, this study emphasizes the difficulties and potential avenues for future research in this field.

***Keywords**: Diabetes detection, Pima Indians Diabetes, Machine Learning, Deep Learning, fine-tuned algorithm, hybrid model.*

1. INTRODUCTION

In both developed and emerging nations, diabetes is a disease that is getting worse and more morbid [1]. *Diabetes* is a long-term medical condition that directly harms the pancreas, thereby making insulin production impossible for the body. Numerous variables, including overweight, sedentary behaviour, high blood pressure, and abnormal cholesterol levels, can contribute to diabetes [2]. Diabetes can raise the chance of dying young and lead to problems in many different areas of the body. Diabetes affects more people than only those who are ill. In addition, the illness has an impact on the ill person's family as well as the entire community. Diabetes has become a widespread issue. Based on WHO estimates, 422 million people worldwide have diabetes. The majority of these people reside in nations with low and moderate incomes. Diabetes is a cause of death for 1.6 million people annually [3]. The following are the major types of diabetes found in humans: Type 1 DM, Type 2 DM, gestational, and presentational diabetes. Medical diagnosis is one of medical science's most challenging and essential jobs. The patient's plasma glucose quantity, diastolic pressure, triceps folds of skin thickness, blood insulin, body mass, age, and other parameters are measured to forecast diabetes disease. The patient then sees a specialist physician. The decision-making process is entirely drawn out and can occasionally take many months or even weeks, making the physician's work extremely challenging.

A vast number of medical datasets are readily available these days. As a result, handling enormous amounts of data by humans might be challenging or even impossible [4]. The majority of diabetic people must live with their condition for the rest of their lives because they usually find out about it too late to receive a complete cure. As a result, efficient computer-based methods are preferred to conventional methods. Technologies based on computers improve accuracy while saving both time and money. The goal is to create a prediction model that can reliably identify diabetes at an early stage and shield people from developing the disease by combining DL and ML algorithms [5]. ML has emerged as a promising approach to diagnose and prevent diabetes. To identify diabetes early on, however, ML approaches like Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), etc. are employed. Unfortunately, when the dataset is large, ML models perform poorly, and they also require human interaction to carry out the feature learning process for diabetes early detection. Erroneous forecasts can be caused by unanticipated occurrences or biased training, among other things.

DL is a rapidly expanding concept that functions similarly to the human mind. It can resolve the selectivity-invariance conundrum effectively and show the facts on several levels. There are several ways that DL approaches are applied in the area of medical prognosis. Numerous studies demonstrate that DL techniques yield superior results, reduce errors in classification rates, and are more resilient to noise than

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

alternative approaches. It can quickly decipher a problematic problem and manage enormous volumes of data. This encourages a survey of current ML and DL-related algorithms for diabetes early detection to be presented.

2. BACKGROUND INFORMATION

Diabetes can raise the chance of dying young and lead to problems in many different areas of the body. Heart attacks and strokes are also two to three times more common in adults with diabetes. Diabetes that is not well managed might raise the risk of death and other severe problems through pregnancy. The phases of early diabetes detection generally include feature selection, classification, data pre-processing, and data collecting. Figure 1 displays the diabetes prediction flow chart.

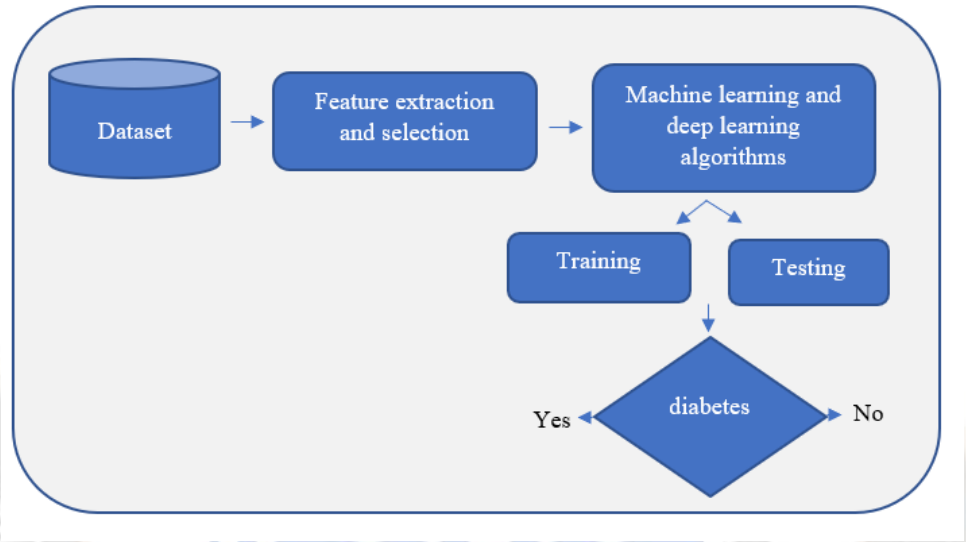

Figure 1: Diabetes detection system

2.1 Dataset Collection

Since we need to use the gathered data to train our classification algorithms, data gathering is the first stage in the analysis process. The National Institute of Diabetes's Pima Indians Diabetes (PID) database and Kaggle can be used by numerous researchers to identify diabetes. Its objective is to enable the discovery of knowledge, organization, and deconstruction from unstructured data sources. It offers a place for handling and organizing data so that information may be found more efficiently.

2.2 Data Pre-processing

Preparing the data is the next stage. The obtained dataset will then go to the pre-processing phase. Data preprocessing is the procedure of converting obtained data into a comprehensible format. However, the dataset cannot include zero values, given our current understanding of medicine and sciences. These 0 values result from either missing measurements or mistakes in data entry. We employed data imputation using the median values of the corresponding columns to solve this problem.

2.3 Feature Selection

Following pre-processing, feature selection narrows down a feature set by keeping the most sensible traits and eliminating drossy ones. Feature selection was applied to obtain valuable characteristics from the collection of data.

2.4 Classification

The final stage in the detection of diabetes is categorization. The goal is to categorize the patients as either having diabetes or not. Two methods for autonomously detecting and preventing signs of diabetes are ML and DL. ML is a potent tool for diagnosing diabetes. LR, SVM, RF, and DT are ML techniques. DL applies a multilayer approach to modify the hidden layers of the neural network. The most popular DL models, Deep Neural Network (DNN), Long-Short Term Memory (LSTM), and Convolution Neural Network (CNN), yield better results.

3. REVIEW ON METHODS USED FOR DIABETES PREDICTION

Some investigators are now more concerned with early diabetes identification due to the rise in diabetes cases. A substantial amount of research has been done on the

identification of diabetes. This review thoroughly examines the approaches used, the results gained, and the restrictions found in each study. Through this, we hope to give a thorough and objective overview of the advancements and difficulties in the field of diabetes research. A summary of recent studies on ML for the detection of diabetes is provided in Table 1.

Table 1: Summary of methods for Diabetes detection

Author & Ref No.	Methodology used	Dataset used	Accuracy (%)	Benefits	Drawbacks
Önder Yakut [3]	RF, Extra Tree and Gaussian process.	PID dataset	RF=81.17, Extra tree=78.57, Gaussian process classifier=77.49	Through the use of computer-aided diagnosis tools, this approach may be helpful in the decision-support process.	This approach works poorly and has a lower precision rate since it might be challenging to obtain possession of more extensive information, which takes longer and costs more.
Aaditi Ranganath Satam *et al.* [6]	LR, DT, RF.	PID dataset	LR=79, RF=78, DT=75	It produces better accuracy since it is used to express a binary outcome and makes the link between the dependent variable and one or more independent factors easier to understand.	RF and DT do quite poorly since it takes longer and demands a lot of computer power and resources.
Rashi Rastogi and Mamta Bansal [7]	LR, SVM, NB, and RF.	Kaggle	LR=82.46, SVM=79.22, NB=79.22, RF=81.	It achieves higher accuracy since it is significantly more straightforward to set up and train than other ML models. It is also one of the most effective algorithms if all the results that the data reflects are separated by linearity.	Due to inadequate standard information, the model SVM and NB accuracy is significantly lower than that of LR, and it also lengthens the time that has passed.
Muhammad Exell Febriana *et al.* [8]	KNN and NB.	PID Dataset	KNN= 73, NB=76.	Because this model is quick, highly adaptable, and designed to balance linear concerning	Its drawbacks are its poor speed, high computing cost, and memory and

				the number of predictions and columns, it works well.	storage problems with big datasets.
Zaigham Mushtaq *et al.* [9]	LR, SVM, KNN, Gradient Boosting (GB), NB, and RF.	PID Dataset	LR=79.2, SVM=74.5, KNN=79.4, GB=79.7, NB=79.7, RF=81.7.	Because the RF model is faster and smoother than other linear models for classification, it performs well.	The primary issue is that because the SVM model contains noise and outliers, it is inappropriate for massive datasets. Hence, this model has low performance.
Monalisa Panda1 *et al.*[10]	LR, GB, KNN and SVM.	Kaggle	LR=80, GB=81, SVM=80, KNN=78	The proposed ML model can capture complex data trends, enabling it to work well in classification if there is noise.	Due to its large memory requirements and sluggish learning process, KNN performs poorly compared to other algorithms.

For diabetes prediction, **Aga Maulana *et al.*** [11] created and improved the XGBoost method using ML models such as KNN, DT, RF, NB, and SVM. When it comes to reducing the number of people who are misclassified as diabetic and non-diabetic but are non-diabetic, the XGBoost approach works effectively. XGBoost employs an ensemble learning approach, which builds a robust and accurate overall model by combining multiple weak prediction models. Based on the PID dataset, the optimized XGBoost model performed best, as seen by 82.68% precision, 84% precision, and 71.76% specificity. This algorithm's primary drawback is that it makes use of the data imputation technique, which might induce biases or inaccuracies when employed to impute the dataset's missing values. To detect diabetes, **Isfafuzzaman Tasin *et al***. [2] introduced ML classifiers such as LR, KNN, RF, XG Boosting, Adaboost, and SVM in conjunction with an explainable AI (XAI) interface. With fewer prediction stages and a quicker convergence to the minimal error, the XG Boost model offers a more straightforward path. With an accuracy of 81%, the XGBoost model on the Kaggle dataset performed the best out of all the ML models. It may nevertheless overfit, mainly when learned on smaller amounts of data.

In imbalanced datasets, **Akash Ashok Choutele** [5] created models such as DNN, XGBoost, and RF for diabetes prediction. An imbalanced class can exclude important information and produce less reliable predictions. Changing the threshold value as a primary tactic, cost-sensitive learning, and sampling—which need more searches to find the best algorithm—are used to get around this. The performance of the algorithms used to improve the class imbalance solutions has been compared. With an accuracy of 87%, the results demonstrated that the model XGBoost employing the Kaggle dataset performed the best. However, because it needs a lot of data for training, this model is prone to overfitting. Utilizing five- and ten-fold cross-validation, **Safial Islam Ayon and Md. Milon Islam [4]** trained the DNN's features to build their diabetes diagnostic technique. We check every dataset into training and evaluation sets, utilizing five- and ten-fold cross-validation. It provides the precise outcome for the entire dataset. The results demonstrated that the PID dataset, with five-fold cross-validation, performed the best, with a 98.80% sensitivity, 98% accuracy, 96% specificity, and 0.99% F1-score. Using five-fold cross-validation has several drawbacks; one is that it needs much storage to be processed.

Bala Manoj Kumar P et al. developed the DNN classification algorithm for precise forecasting on the PID dataset [12]. The author used extra trees and RF to select relevant features. Following every epoch, each node's weight will gradually be adjusted to reduce the input error with learning rate alpha. The method attained more remarkable outcomes than similar techniques. The primary constraint of the approach is computational time because of the substantial volume of data needed for training. To predict diabetes, **Suja A. Alex *et al.*** [13] developed DL models such as CNN, CNN-Long short-term memory, ConvLSTM, DCNN, and SMOTE-based Deep LSTM. The SMOTE-based deep LSTM employs methods such as data pre-processing, class-imbalance processing, reshaping, and deep LSTM modelling to address and make predictions regarding class imbalance. Using the Kaggle dataset, the results demonstrated that the SMOTE-based deep LSTM performed the best, with an accuracy of 99.64%. This model's drawback is that it does not address multiclass classification issues or prioritize treating intraclass imbalance.

In order to forecast diabetes using the PID database, **P. Bharath Kumar Chowdary and Dr. R. Udaya Kumar** [14] developed convolutional LSTM networks. The model convLSTM carries out timing analysis as it extracts abstract features. The result showed that this model works well, as seen by its 96.8% diabetic detection rate. Although this model improves accuracy, it is not accurate since it takes a long time and expensive machines to process the volume of data. Using the PID database, **Namrata Nerkar *et al*. [15]** introduced the neural network model for early diabetes prediction. These networks are capable of producing results with insufficient data. Even if one of the nodes malfunctions, the output remains unaffected. As a result, the network can tolerate errors more readily. The result demonstrated how well the model works, with an accuracy of 85% indicating the presence or absence of diabetes—however, potential overfitting without appropriate regularization results in resource- and computational-intensive operations.

4. DISCUSSION

Diabetes can be identified in its early stages by reviewing medical records, and appropriate treatment can educate the public about the condition. By collecting and analysing data from the PID dataset and Kaggle, the literature review has produced diabetes detection and prevention strategies implemented using ML and DL models. The ML model is frequently employed to detect diabetes early on. The most popular machine-learning algorithms include SVM, DT, RF, KNN, and LR. These are common as they combine multiple weak models of prediction to produce a strong and meaningful accuracy, and they can capture complicated patterns in noisy data. However, ML algorithms increase the elapsed time and come with computational costs, sluggish performance, memory, and storage concerns for massive datasets. DL models provide several benefits when it comes to diabetes detection. The literature mentioned above employed CNN, DNN, and LSTM models. As a backpropagation algorithm, it gathers features and improves the performance of diabetes diagnosis while executing multiple complex operations at once. This eliminates the need for manual feature extraction. However, a large amount of data is needed for training; therefore, the multiclass classification problem differs from its intended use.

5. CONCLUSION

Diabetes is a chronic illness that needs to be avoided before it causes problems for people. Globally, diabetes is a significant cause of death each year. Therefore, early diabetes detection is crucial for effective treatment. This study's initial addition is to the overall flow chart of the diabetes detection system. Subsequently, the literature on diabetes prediction was presented in a tabular format for easy comprehension and reference, along with recommendations for ML and DL approaches. The study then evaluates previous research on the subject to highlight the advantages and disadvantages of that body of knowledge. In order to improve outcomes, the scope of this study's future scope includes gathering more private data from a larger patient group.

COMPLIANCE WITH ETHICAL STANDARDS

Conflict of Interest: I, Preetha Rajagopalam.M, declares no conflicts of Interest to disclose.

Ethical Approval: This article does not contain any studies with human participants or animals performed by any of the authors.

REFERENCES

[1] Dutta, Aishwariya, Md Kamrul Hasan, Mohiuddin Ahmad, Md Abdul Awal, Md Akhtarul Islam, Mehedi Masud, and Hossam Meshref. "Early prediction of diabetes using an ensemble of machine learning models." *International Journal of Environmental Research and Public Health* 19, no. 19 (2022): 12378.

[2] Tasin, Isfafuzzaman, Tansin Ullah Nabil, Sanjida Islam, and Riasat Khan. "Diabetes prediction using machine learning and explainable AI techniques." *Healthcare Technology Letters* 10, no. 1-2 (2023): 1-10.

[3] YAKUT, Önder. "Diabetes prediction using colab notebook-based machine learning methods." *International Journal of Computational and*

Experimental Science and Engineering 9, no. 1 (2023): 36-41.

[4] Ayon, Safial Islam, and Md Milon Islam. "Diabetes prediction: a deep learning approach." *International Journal of Information Engineering and Electronic Business* 12, no. 2 (2019): 21.

[5] Choutele, Akash Ashok. "Diabetes Prediction with the Help of Machine Learning." *Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN) ISSN: 2799-1172* 3, no. 03 (2023): 27-34.

[6] Aaditi Ranganath Satam ``Diabetes Prediction using Machine Learning" International Journal of Modern Developments in Engineering and Science (IJMDES) ISSN (Online): 2583-3138 no. 02 (2023)

[7] Rastogi, Rashi, and Mamta Bansal. "Diabetes prediction model using data mining techniques." *Measurement: Sensors* 25 (2023): 100605.

[8] Febrian, Muhammad Exell, Fransiskus Xaverius Ferdinan, Gustian Paul Sendani, Kristien Margi Suryanigrum, and Rezki Yunanda. "Diabetes prediction using supervised machine learning." *Procedia Computer Science* 216 (2023): 21-30.

[9] Mushtaq, Zaigham, Muhammad Farhan Ramzan, Sikandar Ali, Samad Baseer, Ali Samad, and Mujtaba Husnain. "Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques." *Mobile Information Systems* 2022 (2022): 1-16.

[10] Panda, Monalisa, Debani Prashad Mishra, Sopa Mousumi Patro, and Surender Reddy Salkuti. "Prediction of diabetes disease using machine learning algorithms." *IAES International Journal of Artificial Intelligence* 11, no. 1 (2022): 284.

[11] Maulana, Aga, Farassa Rani Faisal, Teuku Rizky Noviandy, Tatsa Rizkia, Ghazi Mauer Idroes, Trina Ekawati Tallei, Mohamed El-Shazly, and Rinaldi Idroes. "Machine Learning Approach for Diabetes Detection Using a Fine-Tuned XGBoost Algorithm." *Infolitika Journal of Data Science* 1, no. 1 (2023): 1-7.

[12] Nadesh, R. K., and K. Arivuselvan. "Type 2: diabetes mellitus prediction using deep neural networks classifier." *International Journal of Cognitive Computing in Engineering* 1 (2020): 55-61.

[13] Alex, Suja A., N. Z. Jhanjhi, Mamoona Humayun, Ashraf Osman Ibrahim, and Anas W. Abulfaraj. "Deep LSTM Model for Diabetes Prediction with Class Balancing by SMOTE." *Electronics* 11, no. 17 (2022): 2737.

[14] Chowdary, P. Bharath Kumar, and R. Udaya Kumar. "An effective approach for detecting diabetes using deep learning techniques based on convolutional LSTM networks." *International Journal of Advanced Computer Science and Applications* 12, no. 4 (2021).

[15] Nerkar, Namrata, Vaishnavi Inamdar, Likhit Kajrolkar, and Rohit Barve. "Diabetes Prediction using a Neural Network." *International Research Journal of Engineering and Technology (IRJET)* 8, no. 2 (2021): 330-333.