Different Techniques Used For Predicting Diabetes -Literature Survey

KRISHNA PRIYA A S, SHEMITHA P A, Dr G KIRUTHIGA

[1] *STUDENT, DEPARTMENT OF COMPUTER SCIENCE, IES COLLEGE OF ENGINEERING, KERALA, INDIA*
[2] *ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE, IES COLLEGE OF ENGINEERING, KERALA, INDIA*
[3] *ASSOCIATE PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE, IES COLLEGE OF ENGINEERING, KERALA, INDIA*

ABSTRACT

Diabetes is a life-threatening disease that has no cure. If this disease strikes you once, it will be with you for the rest of your life. At the same time, having too much glucose in your blood might cause health problems. Kidney illness, heart disease, stroke, vision, dental, foot, and nerve damage are just a few examples. As a result, you can take steps to manage your Diabetes and avoid complications. Type 1 and type 2 diabetes are the two most common types of Diabetes. This type of Diabetes causes issues such as the body's inability to manufacture or utilize insulin. However, other types of Diabetes, such as gestational diabetes, can develop during pregnancy. High blood sugar levels caused by gestational Diabetes can harm your pregnancy and your baby's health. To diagnose and administer Diabetes, a variety of machine learning and data mining technologies are applied. This research focuses on recent advances in machine learning that have had a substantial influence on diabetes detection and diagnosis. Machine learning techniques are employed to classify diabetic patients in this study. The accuracy of categorization is attained by classifying diabetic patients. Diabetes is one of India's significant health issues nowadays. It's a set of syndromes that cause blood sugar levels to rise too high. It's a long-term disorder that disrupts the body's blood sugar control mechanisms. Diabetes mellitus prevention and prediction are becoming more popular in medical research. The purpose of this research is to conduct a survey of the various diabetes prediction strategies currently in use.

Keywords: - *Diabetes, Machine learning, data mining, Multiperceptron, K-Nearest Neighbours, Logistic Regression, Random forest, Random tree, Naive Bayes, prediction and classification*

1. INTRODUCTION

Diabetes Mellitus (DM), also known as Diabetes, is a widespread chronic condition that poses a significant risk to human health. Diabetes is a long-term illness that affects people all around the world. When the body is unable to produce enough insulin, this occurs. Insulin, one of the most important hormones in the body, is secreted by the pancreas and is required to maintain glucose levels. It can cause symptoms such as frequent urination, thirst, and hunger. Insulin shots, a nutritious diet, and regular exercise can all assist in managing Diabetes. Diabetes can cause blindness, high blood pressure, heart disease, kidney disease, and other complications[1]. Diabetes is divided into four types.

Type 1 Diabetes: When the pancreas fails to deliver enough insulin, the body develops type 1 diabetes. It can occur at any point in one's life. e.g., children and adolescents [2].

Type 2 Diabetes: When the amount of insulin produced is insufficient to meet the body's needs, type 2 diabetes develops. Obesity increases the risk of developing type 2 diabetes due to parental inheritance, seniority, and corpulence. The majority of it occurs between the ages of 40 and 50. Gestational Diabetes: -It is the third most common kind of Diabetes, and it most commonly affects pregnant women due to an excess of glucose in the blood. Pregestational Diabetes: This is a different type of Diabetes that occurs when insulin-dependent Diabetes develops before becoming pregnant [2]

2. LITERATURE REVIEW

This section of the paper discusses some of the studies on medical diagnosis utilizing machine learning and data mining approaches.

Sonu Kumari and Archana Singh developed [3] a smart and effective Neural Network-based methodology for the automated detection of Diabetes Mellitus. The research [4] used ANNs to approach the goal of diagnosing and highlighted the need for preprocessing and replacing missing values in the dataset under consideration. With the Modified training set, higher accuracy was gained while training the set took less time. Sajida[8] used the CPCSSN (Canadian primary care sentinel surveillance Network) dataset and three machine learning approaches to predict diabetes diseases (DD) at an early stage, allowing people to live longer and avoid death. In this study, Bagging, Adaboost, and decision tree (J48) were used to predict Diabetes, and the results were compared, with the researcher concluding that the Adaboost approach was more successful and accurate than the other methods in the Weka data mining tools. Sadri [20] employed data mining algorithms such as Naive Bayes, RBF Networks, and J48 to diagnose type 2 diabetes. They made use of the WEKA tool. Finally, they discovered Naive Bayes, which outperformed other algorithms by 76.96 per cent. In this paper[27], Diabetes is predicted using ensemble voting classifiers for the Pima Indian diabetes dataset. When compared to other classification algorithms, the highest accuracy of 80% and 81 per cent is achieved for the data set by using 10-fold cross-validation and spitting data into 30% testing and 70% training.

This research study by J. Pradeep Kandhasamy and S. Balamurali [58] compares the performance of algorithms used to predict Diabetes using data mining approaches. J48 Decision Tree, KNearest Neighbors, and Random Forest, as well as Support Vector Machines, were used to classify patients with diabetes mellitus. The authors compared four diabetes mellitus prediction models utilizing eight key features in two separate scenarios. One is before the dataset is preprocessed. According to the findings, the decision tree J48 classifier has a greater accuracy of 73.82 per cent than the other three classifiers.

When compared to earlier studies, the dataset provided more accurate results after preprocessing. Both KNN (k=1) and Random Forest outperform the other three classifiers in this scenario, and they yield 100 per cent accuracy. As a result, we can deduce that deleting the noisy data from our dataset will yield a positive solution to our problem.

This paper[5] demonstrates how data mining classification algorithms such as Naive Bayes, Logistic Regression, C5.0, SVM, and ANN are used to model actual Diabetes Mellitus Prediction, and a comparison is made using their Metric Measures such as Accuracy, Precision, Sensitivity, Specificity, and F1 Score.

. Based on their Accuracy measurements, the C5.0 and Logistic Regression are equally good as a consequence of the research work. Rahul and Minyechil Alehegn [7] looked into different data mining techniques and how they may be used. Machine learning algorithms were used to analyze various medical data sets. The accuracy of a single algorithm was lower than that of an ensemble. The decision tree provided great accuracy in the majority of studies. The technologies used to forecast diabetes datasets in this study are a hybrid system called Weka and Java. This paper[18] discusses multiple supervised classifier machine learning techniques that were applied to the training set, which was created by removing features that were unrelated to diabetes prediction. The chi-squared test was used, and only the traits that were ranked highest were given additional weightage. It was thought to be more likely to predict the onset of Diabetes. On this training set, it was discovered that the Neural Networks method produced the best accurate results. The paper[44] examines the three forms of Diabetes and the factors that contribute to its development. It also employs categorization and prediction techniques. This results in a better level of disease prediction accuracy. The study paper[45] investigates the various data mining algorithm approaches that have been used to forecast diabetes illness. Classification and Naive Bayes is one of the most widely used algorithms for disease prediction in this publication.

In this study, Pradeep and Dr Naveen [10] compared and measured the performance of machine learning approaches based on their accuracy. The technique's accuracy varies depending on whether it was preprocessed or not.

Following preprocessing, as identified in this study, implies that in the prediction of diseases, data preprocessing has an effect on the prediction's performance and accuracy. Song [6]describes and explains many categorization algorithms based on variables such as blood sugar, blood pressure, skin thickness, insulin, BMI, diabetes pedigree,

and age. Pregnancy parameters were not included in the studies to predict diabetes illness (DD). The researchers in this study used only a tiny sample of data to predict Diabetes. This work employed five different algorithms: GMM, ANN, SVM, EM, and Logistic regression. Finally. The researchers came to the conclusion that ANN (Artificial Neural Network) provided high accuracy for diabetes prediction. P. Chen[19] performed statistical testing on medical measurement index data of both diabetic and non-diabetic patients in their research. They also employed boosting algorithms to provide superior diabetes model categorization based on the medical data provided. A medical bioinformatics analysis was performed in this work [39] to predict Diabetes. The WEKA programme was used as a data mining tool for diabetes diagnosis. The Pima Indian diabetes database was obtained from the University of California, Irvine, and was used for analysis. The data was examined and processed in order to develop a model that accurately predicts and diagnoses Diabetes. In this study, we will use the bootstrapping resampling technique to improve accuracy and then compare the performance of Nave Bayes, Decision Trees, and (KNN). Yunsheng[9] proposed a new method for removing outliers/OOBs (out of a bag) using the KNN algorithm and DISKR (reduce the size of the training set for K-nearest neighbour). In addition, storage space was decreased in this investigation. As a result, after deleting parameters or instances that have little influence or factor, the space complexity has decreased, and the researchers have improved their accuracy. V. Kumar and L. Velide[21] used a data mining approach for diabetic disease prediction and treatment. They employed Naive Bayes, JRip, J48 (4.5), DT, and NN methods. For implementation, they employed the WEKA tool. For the J48 algorithm, they achieved a level of accuracy of 68.5 per cent. The research study [46] delves into a thorough examination of existing data mining methods for the prediction of diabetics. The research publication [46] goes into great depth regarding present data mining algorithms for diabetes prediction. It also explains the three forms of Diabetes: Type 1, Type 2, and Type 3. The goal of diabetes prediction is to employ data mining approaches such as the K-Nearest Neighbor Algorithm, Bayesian Classifier, Naive Bayesian Classifier, Bayesian Network, and Bayesian Network to forecast Diabetes. The impacts of Diabetes on patients are also discussed in this paper. The proposed methodology in this paper[54] attempts to provide an effective hybrid classification framework for predicting and monitoring Diabetes illness. The major goal of this study is to find and build models that can help medical practitioners work more efficiently while also benefiting patients. Monika and Pooja [11] have talked about the most recent advancements in medical science research as well as historical data retrieval methods. Additionally, we have clarified the language and learning methods used in data mining and machine learning. To increase accuracy by applying clustering and classification algorithms, D. Jeevanandhini, E. Gokul Raj, V. Dinesh Kumar, and N. Sasipriyaa [12] did a performance analysis for the type 2 diabetes mellitus dataset. Here, they used eight key attributes to compare the four prediction models. The Support Vector Machine (SVM) classifier, which outperforms the other three classifiers, obtains a greater accuracy of 77.82 per cent, according to this study's findings. Different data mining algorithms were employed and applied to Pima datasets by Dr K. Thangadurai and N. Nandhin[13]. It has been discovered that the genetic algorithm performs better than the five data mining algorithms. In this paper[49], we compared a few different classification algorithms using the Matlab tool for analysis. Following a comparison, we found that neural network methods are more precise and have a lower error rate. Additionally, our interface gives users the option to choose a suitable prediction algorithm. We get to the conclusion that ANN is more accurate than other models.

This evaluation paper[2] focuses on several predictive analysis techniques and methods and makes use of an early estimate of the number of diabetes diagnoses from patient records. In the realm of health records, several analytical techniques are used to anticipate diabetes cases and identify efficient treatments. The proposed Logistic Regression model's AROC is 84.0 per cent with a sensitivity of 73.4 per cent in this study [37], while the proposed GBM model's AROC is 84.7 per cent with a sensitivity of 71.6 per cent. Compared to the Random Forest and Decision Tree models, the GBM and Logistic Regression models perform better. In this study[38], we used machine learning approaches to examine early diabetes prediction by taking into account several risk variables associated with this condition. Predicting diabetic patients may be possible by extracting knowledge from real health care datasets. On adult population data, we conducted trials employing four well-known machine learning methods, including Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), and C4.5 decision tree, to accurately predict diabetes mellitus. The most crucial problem in a real-world scenario is the detection and analysis of clinical activity because these operations are extremely difficult due to a lack of training samples and data. Diabetes mellitus can be diagnosed and prognoses in a number of ways. This survey[16] offers several data mining strategies to address the challenge of diagnosing diabetic illness. From the analysis, we learn about a number of issues and findings in the processing of clinical datasets. In his article, Sowjanya [17] devised an android application-based approach to address the lack of awareness regarding DM. The DT classifier was utilized by the application to forecast users' degrees of Diabetes. The system also offered advice and information on Diabetes. In this paper[14], the "Vote" methodology, which combines three ensembling methods, The prediction accuracy was increased by the

ensembling technique to AUC = 0.922. According to the study[15], ensembling and SMOTE techniques can be used to predict incident diabetes using the information on cardiorespiratory fitness. Devi describes the creation of a composite classification model for the Pima Indian diabetic database (PIDD). K-means and K Nearest Neighbor are combined in this model (KNN). For the identical k-values, they contrast the outcomes of simple KNN, cascaded K-means, and KNN. The outcomes are then compared using WEKA tool-calculated statistical parameters like accuracy, sensitivity, and specificity. For k=5, the accuracy of K-means and KNN is 97 per cent, compared to 73.17 per cent with basic KNN and 97.4 per cent for amalgam KNN. Amalgam KNN has a basic accuracy of 96.87 per cent for k=3.accuracy for simple KNN is 72.65%. The author came to the conclusion that when K's value rises, the algorithm's performance does as well. For the classification of Diabetes illness, V. Anuja Kumari and R. Chitra[22] employed SVM with Radial Basis Function Kernal. For implementation, they made use of MATLAB R2010a. They discovered a 78 per cent accuracy rate.

The performance of this method is assessed using a 10-fold cross-validation accuracy and confusion matrix in Preeti Verma, Inderpreet Kaur, and Jaspreet Kaur's paper[36]. In comparison to previous spline SSVM techniques, the classification accuracy obtained using 10-fold cross-validation is 96.58 per cent. The study's findings demonstrated that the modified spline SSVM was useful for diagnosing Diabetes, which is a highly encouraging finding when compared to earlier findings. In the research of Drs. B.L. Shivkumar and S. Thiyagarajan[23], For the classification of type DM patients, a powerful machine learning method is suggested. The ideal hyper-plane that splits the various classes will be found by this machine learning technique used for classification. An approach that focuses on choosing the qualities that fail in the early identification of Diabetes Miletus using Predictive analysis was proposed by Sneha and Tarun[24]. The examination of diabetic data reveals that the decision tree algorithm and random forest have the highest specificities, with 98.20 per cent and 98.00 per cent, respectively. The best accuracy according to naive Bayesian results is 82.30 per cent. In order to increase the classification accuracy, the research also generalizes the selection of the best characteristics from the dataset. This paper[47] emphasizes how data mining methods can be quite useful in the early stages of forecasting and, as a result, taking preventative measures before a disease is diagnosed. The primary objective of this study is to compare various algorithms and recommend the best one for pattern recognition or prediction in the healthcare industry. Following the deployment of these methods, it can be said that the Decision Tree provides the best accuracy for the PID dataset (75.6%). Rapid Miner is the tool used for testing and validation, and all algorithms were tested and trained in a ratio of 70:30.

In this paper[33], Diabetes is predicted using ensemble voting classifiers for the Pima Indian diabetes dataset, and when compared to other classification methods, the highest accuracy of 80% and 81 per cent are attained for the data set by applying 10-fold cross-validation and spitting data into 30%. Testing and seventy per cent instruction. The paper [32] used ANNs to approach the goal of diagnosing and showed the necessity of preprocessing and replacing missing values in the dataset under consideration. With the Modified training set, higher accuracy was gained while training the set took less time. In this study [28], classification methods for diabetes data were classified using Binary Logistic Regression, Multilayer Perceptron, and K-Nearest Neighbor, and classification accuracy was compared. A Class wise KNN(CKNN) methodology[34] for classifying diabetic datasets was proposed, in which the dataset is first preprocessed using normalization and then classified using an improvised model of the KNN algorithm, or

class-wise KNN algorithm. The accuracy of this strategy is 78.16 per cent. Bansal and Rohan [35] employed a KNN classifier to diagnose Diabetes, using Particle Swarm Optimization (PSO) methods to choose the attributes. It has been demonstrated that this strategy has a 77 per cent prediction accuracy.

Oracle Data Miner and Oracle Database 10g were utilized for analysis and storage, respectively, in Abdullah's [40] work.

In this investigation, the parameters or factors were found. Based on their percentage, the target variables were found. The patient's care was the main focus of this investigation. In order to forecast their treatment, the patient was divided into two groups based on their ages: elderly and young. Dietary control reveals a high percentage in this study for both young and old people. The treatment predicted the percentage produced by an SVM. In this study, the researchers applied several data mining approaches, according to Xue-Hui Meng [41].

Employing real-world data sets and distributed questioners to gather information, and forecast diabetic illnesses, In this study, three techniques—ANN, Logistic Regression, and J48—were compared using SPSS and Weka tools for data analysis and prediction, respectively. In the end, it was determined that the J48 machine learning technique offers effective and improved accuracy.

Using machine learning and an EMR database, Byoung Geol Choi, Seung-Woon Rha, Sung Wook Kim, Jun Hyuk Kang, Ji Young Park, and Yung-Kyun Noh[26] created and effectively validated a T2DM prediction system that accurately predicted the 5-year prevalence of T2DM. Discuss how Least Square Support Vector Machine (LS-SVM) and Generalized Discriminant Analysis (GDA) is employed for the diagnosis of diabetes disease. Additionally, a

brand-new cascade learning system built on least square support vector machines and generalized discriminant analysis was proposed. Two stages make up the suggested system. In the first stage, they used generalized discriminant analysis as a preprocessing step to distinguish between feature variables in healthy and patient (Diabetes) data. In the subsequent stage, they classified the diabetes dataset using LS-SVM. The proposed system dubbed GDA-LS-SVM obtained 82.05 per cent classification accuracy using 10-fold cross-validation, which is extremely promising when compared to the previously published classification methodologies, whereas LS-SVM obtained 78.21 per cent classification accuracy using 10-fold cross-validation. To pick the subset of attributes from the original data for our suggested method[57], we employed an attribute selection filter. Next, we applied the J48 and decision stump data mining classification approaches, which are used to forecast Diabetes. The confusion matrix is used to assess classifier performance in terms of accuracy and execution time. The Random Tree Algorithm provides an accuracy of 86.59 per cent, which is higher than that of other classifiers, and it also classifies data sets more quickly than other classifiers. The proposed system in the study paper[48] uses well-known and widely-used machine learning algorithms. J48, KNN, NB, and Random Forest are the algorithms employed in this study. In the case of PIDD employing stacking meta, the proposed method offers superior accuracy of 93.62 per cent. An ensemble method offers greater accuracy than a single prediction algorithm in the instance of the huge dataset 130-us hospital. The most popular Data Mining tools and methodologies, as well as the kinds of datasets and features that have been utilized to forecast diabetes disease in patients, are summarised in this work [50]. The effectiveness and accuracy of each algorithm employed in this study are also included. The research showed that preprocessing the data will improve the performance of the classifiers.

For the diagnosis of Diabetes mellitus in this paper[53], we employed the K-nearest neighbour approach. For K=3, 5, we have determined accuracy and error rates. The outcome indicates that the accuracy rate and error rate would cause both rises as the value of k rises.KNN is a well-known artificial intelligence method that is extremely useful for diagnosing problems. KNN allows for the production of outcomes that are more precise and effective. Different categorization strategies were used in the data mining procedure in this paper[55]. These methods have been used to separate the data into various sets, making it simple to identify relationships between various attributes. Health care practitioners have employed a variety of data mining approaches to aid in the identification of diabetes conditions. Additionally, data mining methods including support vector machines, automatically created groups, bagging algorithms, and kernel densities are employed. We employed three distinct classification algorithms in this experiment, namely Xero, OneR, and Naive Bayes, in this paper[56]. This study demonstrates that the Naive Bayes model is the fastest and zero is the slowest.

The same data set, identically split into training and forecasting sets, is being tested in this paper[59] using both a logistic regression model and a linear perceptron model. Data will be available to compare the ROC curves produced by the ADAP, the logistic regression, and the linear perceptron for this prediction. As opposed to linear regression, which produced a confusion matrix and led to the compact, precise formation of weights based on characteristics and attributes, LS-SVM classifies the estimation of probabilistic models, where the scheme can define the range in which the prediction can be made more accurately based on the type of categorization in Diabetes and precautions, respectively. As a result, the resultant values and derived with greater accuracy in this paper [62]. This research [64]focuses on several data mining approaches and strategies that are employed for the early diagnosis of various Diabetes. Data mining is a process used to extract valuable information from a large number of existing data, allowing you to learn more. Therefore, using data mining techniques and approaches will assist in anticipating Diabetes and also lower the cost of treatment. In order to forecast Diabetes and identify effective treatments for it, data mining techniques are used in the medical data sector.

3. CONCLUSIONS

This document lists the most popular methods, datasets, and features that have been applied to predict the presence of Diabetes in individuals. The effectiveness and accuracy of each algorithm employed in this study are also included. Some studies also concentrate on minimizing correlation or misclassification of datasets. There is still a significant gap between accuracy and computing speed that needs to be closed. The Diabetes dataset has numerous opportunities for research.

4. REFERENCES

[1] Monika, Pooja Sharma.2018. Survey on Prediction and Analysis of Diabetic Data using Machine Learning Techniques from International Journal of Computer Sciences and Engineering, E-ISSN: 2347-2693

[2] D. Jeevanandhini , E. Gokul Raj ,V. Dinesh Kumar, N. Sasipriyaa.2018. Prediction of Type2 Diabetes Mellitus Based on Data Mining from International Journal of Engineering Research & Technology (IJERT), technology (IJERT) ISSN: 2278-0181

[3] Dr. K. Thangadurai, N.Nandhini.2016. Comparison of data mining algorithms for prediction and diagnosis of Diabetes mellitus from International Journal of Scientific & Engineering Research, Volume 7, Issue 5, ISSN 2229-5518

[4] Manal Alghamdi, Mouaz Al-Mallah, Steven Keteyian, Clinton Brawner, Jonathan Ehrman and Sherif Sakr.2017. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project

[5] M Nirmala Devi, Balamurugan.S Appavu alias, U.V Swathi.2013.An amalgam KNN to predict Diabetes Mellitus, IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology, Madurai, Tamil Nadu, India

[6] B. Senthil Kumar, Dr R. Gunavathi.2016. A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis from International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified, ISSN (Online) 2278-1021

[7] Ms K Sowjanya.2015. MobDBTest: A machine learning-based system for predicting diabetes risk using mobile devices. IEEE International Advance Computing Conference (IACC).

[8] Ambika Rani Subhash, Ashwin Kumar UM.2019. Accuracy of Classification Algorithms for Diabetes Prediction from International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958

[9] K. Saravananathan, T. Velmurugan, "Analyzing Diabetic Datausing Classification Algorithms in Data Mining", International Journal of Science and Technology, Vol. 9 (43), November 2016

[10] S. sa'di, A. Maleki, R. Hashemi, Z. Panbechi, and K. Chalabi.2015. Comparison of Datamining Algorithms in the Diagnosis of Type II Diabetes, IJCSA, vol. 5, no. 5.