Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 10
October 2023

Machine Learning Methods for Diabetes Prediction: A Literature Review Paper

[1] Poonam Sengar, [2] Dr. Sandipkumar R. Panchal
[1] Research Scholar, Gujarat Technological University, Ahmedabad,
[2] Vice-Principal, Dr. Subhash Technical Campus, Junagadh

***Abstract*—** Diabetes mellitus is a long-lasting medical condition stemming from either insulin resistance or insufficient insulin production. Insulin is responsible to control level of glucose in blood. Hyperglycemia, alternatively known as raised blood glucose or raised blood sugar, frequently results from uncontrolled diabetes and can lead to significant harm to various bodily systems, particularly the nerves and blood vessels, when left untreated over time. Early diagnosis and prediction of diabetes can significantly improve patient health. Machine learning methods have become valuable tool for predicting diabetes, and their capacity to analyze extensive datasets and uncover complex patterns. The exploration of machine learning methods aims to uncover suitable strategies for efficiently classifying the diabetes dataset and extracting valuable patterns. Over the years, numerous researchers have endeavored to develop precise diabetes prediction models. This comprehensive review aims to encapsulate the diverse array of methodologies, advancements, and findings in the field, shedding light on the intricacies associated with predicting diabetes through the lens of modern methods of machine learning

Index Terms— diabetes, classification, machine learning, clustering.

Introduction

Diabetes develops when your body struggles to efficiently absorb sugar (glucose) into its cells for energy utilization, resulting in an accumulation of excess sugar in your bloodstream. Poorly managed diabetes can have severe consequences, potentially causing harm to various organs and tissues in your body, such as the heart, kidneys, eyes, and nerves. The prevalence of diabetes diagnoses is on the rise globally, encompassing India as well. Most of the cases are type-2 diabetes caused by pancreas slowing losing ability to make insulin.

There are four diabetes types, with Diabetes Mellitus (DM) being categorized as Type-1, also known as Insulin-Dependent Diabetes Mellitus (IDDM). This form of DM arises from the body's inability to produce adequate insulin, necessitating insulin injections for patients [1]. Type-2 diabetes, also referred to as Non-Insulin-Dependent Diabetes Mellitus (NIDDM), occurs when the body's cells struggle to utilize insulin effectively. Type-3, known as Gestational Diabetes, arises from heightened blood sugar levels during pregnancy, particularly when diabetes hasn't been previously diagnosed [1]. Diabetes carries the burden of enduring complications, and individuals with diabetes face elevated risks of various health issues [1]. In India, around 77 million adults are believed to have type 2 diabetes, and approximately 25 million are in a prediabetic state, putting them at greater risk of developing diabetes in the near future reported by WHO [2]. Over half of the population remains unaware of their diabetic condition, which can result in health complications if left undetected and untreated [2]. Individuals who have diabetes face a two- to three-fold higher likelihood of experiencing heart attacks and strokes. [2]. When neuropathy, which is nerve damage in the feet, occurs in conjunction with reduced blood flow, it raises the chances of developing foot ulcers, infections, and, ultimately, requiring limb amputation [2]. Diabetic retinopathy, which is a significant reason for vision loss, happens when the tiny blood vessels in the retina get damaged over a long time due to diabetes. Additionally, diabetes is one of the main reasons for kidney failure [2].

The primary risk factors contributing to diabetes are obesity and overweight. A substantial portion of the diabetes burden can be mitigated or delayed through behavioral modifications, including

adopting a healthy diet and engaging in regular physical activity.

I. MACHINE LEARNING METHODS

ML (Machine Learning) is a specific approach within the realm of artificial intelligence that gathers information from training data. Recent advancements in machine learning will simplify and make diabetes detection more affordable. There are plenty of available datasets related to diabetes. Machine learning methods are primarily classified into three categories: supervised machine learning, unsupervised machine learning, and reinforcement learning.

Supervised learning

In supervised learning, the machine learns through provided examples. A supervised learning algorithm works with a known dataset of input data and corresponding known outcomes (output) to teach a model how to make accurate predictions for new data. This iterative cycle continues until the algorithm achieves a significant level of accuracy and performance. Supervised machine learning encompasses a variety of algorithms, such as linear and logistic regression, multiclass classification, and SVM (support vector machine)

Unsupervised learning

In unsupervised learning machine learns independently studying data to uncover underlying patterns, without relying on labeled output for guidance. By analyzing available data, the algorithm deduces correlations and relationships, interpreting large datasets and organizing them to reveal their inherent structure. Unsupervised learning methods such as clustering, fuzzy clustering, hierarchical clustering, K-means clustering, and association rule mining, to derive insights and patterns from data without explicit labels or guidance.

Reinforcement learning

In reinforcement learning, an autonomous agent interacts with an environment in discrete time steps. The agent's objective is to learn a policy, denoted as a mapping from states to actions, which enables it to make decisions that optimize a numerical scalar signal, referred to as a 'reward.' These decisions are based on the agent's observations of the environment's current state and are intended to maximize the expected cumulative reward over an extended period. Reinforcement learning use a dynamic feedback loop, where in the agent continuously explores different actions, receives feedback in the form of rewards or penalties, updates its policy through a learning process, and refines its decision-making capabilities over time [3].

Literature Review

This literature review shows that most of the researchers have employed diverse classifier to achieve highest accuracy to detect persistent illness. In the domain of incurable disease prognosis and diagnosis, researchers have utilized a range of classifiers, including SVM, neural networks, decision trees, naïve Bayes, and others.

Mujumdar et al. [1] author implemented various classification algorithms, such as the SVM classifier, RF (Random Forest Classifier), DT (Decision Tree Classifier), Extra Tree Classifier, boosting Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm, LR (Logistic Regression), K-NN, Gaussian Naïve Bayes, and Bagging algorithm, along with the Gradient Boost Classifier, to analyze two distinct datasets. These datasets comprised the PIMA (Pima Indians Diabetes Database) dataset and another diabetes dataset for evaluating different models. It was observed that LR achieved the highest accuracy, reaching 96% on the diabetes dataset and 76% on the PIMA dataset.

Mercaldo et al. [4] use feature selection strategy in data preprocessing represents a pivotal step in enhancing accuracy. The evaluation was conducted on real-world data derived from the Pima Indian population near Phoenix, Arizona. Six distinct classification algorithms were employed to train the model. Initial results yielded a precision of 0.757 and a recall of 0.762 without applying feature selection. Subsequently, precision improved to 0.770, and recall reached 0.775 after the applying feature selection method. The author opted the best-first and greedy stepwise feature selection algorithm. WEKA platform used for the classification analysis.

Messan et al. [8] explored the early prediction of diabetes in their study, employing five distinct algorithms: GMM (Gaussian mixture model), ANN (Artificial Neural Network), SVM, EM (Extreme Learning Machine), and LR. Among these algorithms, the ANN classifier achieved the highest accuracy of 89% using the MATLAB tool. The study was carried out on a limited dataset for diabetes

prediction. Ultimately, the researchers concluded that ANN demonstrated exceptional accuracy in predicting diabetes.

Nilashi et al. [9] In this paper, the authors introduced a novel knowledge-based system designed for the prediction of diseases. The system incorporated various techniques, including clustering, noise removal, and prediction methodologies. Specifically, Classification and Regression Trees (CART) are employed to generate fuzzy rules that form the basis of this knowledge-based system. The proposed approach is evaluated using several publicly available medical datasets, encompassing Pima Indian Diabetes, Mesothelioma, WDBC, StatLog, Cleveland, and Parkinson's telemonitoring datasets. The key findings of this research indicate that the proposed method substantially enhanced the accuracy of disease prediction. The experimentation results demonstrated notable improvements in prediction accuracy across the mentioned datasets. This suggests that the fusion of fuzzy rule-based techniques with CART, along with the incorporation of noise removal and clustering methodologies, holds significant promise for the effective prediction of diseases from real world medical datasets.

Jingyu Xue et al. [10] employed supervised machine-learning algorithms, including SVM, NB (Naive Bayes classifier), and LightGBM (Light Gradient Boosting Machine), to train on real data sourced from UCI repository, encompassing both diabetic patients and potential diabetic patients aged 16 to 90. Through a comparative analysis of performance based on accuracy, the SVM exhibited the highest performance which was 96.54%. The NB classifier model achieved an accuracy of 93.27%, while the LightGBM model achieved an accuracy of 88.46%.

Hashi et al. [11] presented an expert healthcare predictive decision support system engineered for diabetes prediction. The author implemented the decision Tree (C 4.5) and K-Nearest Neighbor (KNN) algorithms. This model was trained using PIMA diabetes dataset. Among two algorithms, the C4.5 algorithm achieved greatest accuracy of 90.43%.

Joshi et al. [12] introduced a system designed for early diabetes prediction in patients. The dataset was not mentioned. The author employed two distinct machine learning algorithms, namely SVM and LR. The experiments revealed that SVM outperformed, with the highest achieved accuracy reaching approximately 79%. Python used to implement the algorithms.

Kandhasamy, J. P et al. [13] conducted a comparative assessment of various machine learning algorithms to assess their effectiveness in diabetes prediction. The study compared DT (Decision tree J48), KNN (K-Nearest Neighbors), RF, and SVM, and the findings revealed that the DT achieved the lowest accuracy among the considered classifiers. The performance analysis of these classifiers was conducted through a series of experiments using UCI data repository for diabetes mellitus. WEKA tool was used for the implementation.

Syed et al. [14] proposed a comprehensive approach to diabetes prediction and risk assessment. It includes dataset selection, data visualization, dimensionality reduction, a thorough evaluation of machine learning and probabilistic modeling techniques, assessment of diabetes risk factors, and the application of fuzzy logic and rule-based systems for predicting different levels of diabetes. Various machine learning algorithms and probabilistic modeling techniques were applied to the preprocessed dataset. These include LDA (Linear Discriminant Analysis), Logistic Regression, GLMNET (Generalized Linear Model), SVM Radial (Support Vector Machine with Radial Kernel), k-Nearest Neighbors (kNN), NB, Regressive Partitioning (rpart), Boosted Tree (C5.0), Bagged CART (treebag), RF, and Generalized Boosted Modeling. The research involved evaluating the Diabetes Risk Factor (DRF) using selected features. Membership functions were designed, and fuzzy rules were constructed. This suggests the application of fuzzy logic and rule-based systems to obtain different levels of diabetes. Fuzzy logic allows for the representation of uncertainty and imprecision in medical data, enabling more nuanced predictions and recommendations.

Kadhm et. al. [15] introduced a classification-based approach that assigns each data sample to its appropriate class using k-means clustering. This approach consists of two primary stages: data preparation and classification. In first stage, ten clusters are formed, each containing samples

categorized as either diabetic or healthy. Groups of a reduced size within each cluster are excluded. The DT algorithm is then applied to each subset, and the final result was determined through majority voting among all classifier outcomes. The experimental results illustrated that the proposed system outperforms existing systems in terms of achieving the highest level of accuracy.

Zhu et al. [16] proposed a method consisting of three sub-stages. In the initial stage, PCA (principal component analysis) performed on dataset to reduce dimensionality. In the subsequent substage, K-means clustering used to identify the outliers and removing inaccurately classified data. In the final substage, four separate classification algorithms were applied to the dataset. The experimental findings demonstrated that utilizing PCA+K-means followed by LR produced the highest level of accuracy.

Talha Mahboob Alam et al. [18] concentrated on early diabetes prediction to enhance treatment outcomes. The authors began with data preprocessing and subsequently employed classifier techniques. Notably, they utilized principal component analysis for significant attribute selection. In their diabetes prediction efforts, they implemented ANN, random forest (RF), and K-means clustering techniques. The experimental outcomes showcased the potential of the ANN, achieving the highest accuracy at 75.7%.

Subashree et al. [19] utilized the bootstrapping resampling technique to enhance accuracy as data preprocessing step and subsequently employed Naïve Bayes, Decision Trees, and KNN. The author used the WEKA as software tool and PIMA diabetes dataset for analysis to predict diabetes.

Md. Kamrul Hasan et al. [20] employed an ensemble approach, combining different machine learning classifiers such as kNN, DT, Adaboost,RF,NB,XGboost and MLP (multilayer perceptron) to enhance predictive performance. Experimental results revealed that the ensembling classifier (Adaboost+XGboost) exhibited superior performance to prediction of diabetes compared to all other employed algorithms. PIMA Diabetes dataset was used.

Lakshmi Priya et al. [21] introduced early stage diabetes prediction system. This system used the Naïve Bayes Classifier on preprocessed dataset. The dataset consists of diabetic patient data collected from a clinic repository, comprising 1865 instances with various attributes.

Priyanka Rajendra et al. [22] proposed ensemble methods to enhance prediction accuracy compared to single models. The experiments encompassed two distinct datasets: the PIMA Indians Diabetes dataset and another dataset from Vanderbilt. This study highlighted that, in addition to algorithm selection, various factors played pivotal roles in improving model accuracy and runtimes. These factors included data preprocessing, addressing redundant and null values, normalization, cross-validation, feature selection, and the incorporation of ensemble techniques. The highest accuracy achieved was approximately 78% for Dataset 1, achieved through the ensemble technique of Max Voting; while for Dataset 2, it reached around 93%. The analysis was conducted using the Python IDE.

Jobeda Jamal Khanam et al. [23] utilized seven different machine learning algorithms on a preprocessed dataset for diabetes prediction. The experimental outcomes highlighted that LR and SVM displayed robust performance in predicting diabetes.

Satish Kumar Kalagotla et al. [24], the study encompassed three key stages: (1) developing a correlation technique for feature selection, (2) implementing the Ada-Boost technique on the selected features for classification, and (3) introducing a novel stacking technique that combined MLP (multi-layer perceptron), SVM, and LR on the selected features. To showcase their versatility, the suggested models were also tested on other datasets, including the Cleveland heart disease and Wisconsin breast cancer diagnostic datasets. The experimental findings showcased that the stacking approach exhibited superior performance in comparison to other models, particularly when contrasted with methods applied to the PIMA dataset.

Tigga et al. [25] the author compared the performance of various six classification algorithms using statistical measures on a diabetic dataset collected through questionnaires administered both online and offline. The algorithms included LR (Logistic Regression), k-nearest neighbor, SVM, NB, DT, and RF. Furthermore, the same algorithms were

employed on the PIMA database for comparative analysis. According to the experimental results, Random Forest exhibited the highest accuracy of 94.10% for the author's dataset, surpassing all other methods. RF also achieved the highest accuracy of 75% when applied to the PIMA dataset.

Deepti et al. [26] focused on pregnant women with diabetes. The author implemented Naive Bayes, SVM, and Decision Tree, to predict early-stage diabetes in patients using the PIDD (Pima Indians Diabetes Database). The performance of these algorithms was evaluated based on various metrics, including precision, accuracy, F-Measure, and recall. The findings indicated that Naïve Bayes outperformed the other algorithms, achieving the highest accuracy at 76.30%. The WEKA tool was utilized for conducting the experiments.

Conclusion

Machine learning methods have exhibited substantial promise in diabetes prediction by utilizing feature selection, diverse classification algorithms and rigorous model evaluation. As technological advancements and data availability continue to expand, the application of machine learning in diabetes prediction is poised to yield more precise and clinically relevant results, thereby aiding in the early diagnosis and management of diabetes. The references cited in this literature review offer comprehensive insights into the current state of machine learning in diabetes prediction and serve as valuable resources for further exploration and research in this evolving field. In conclusion, this paper serves as a valuable resource for researchers looking to enhance the predictive capabilities of automatic diabetic detection systems. By leveraging the existing dataset and focusing on data preprocessing and algorithmic improvements, we aim to support researchers in their pursuit of more effective clinical prediction models for diabetes.

References

[1] A. Mujumdar, V. Vaidehi, Diabetes prediction using machine learning algorithms, International Conference on Recent Trends in Advanced Computing", 2019, ICRTAC, 2019.

[2] World Health Organization. https://www.who.int/india/health-topics/mobile-technology- for-preventing-ncds.

[3] LP. Kaelbling and Aw. Moore, "Reinforcement Learning A Survey", Journal of Artificial Intelligence Research, vol-4, pp.237-285, 1996.

[4] Mercaldo, F., Nardone, V., & Santone, A. (2017). Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. International Conference on Knowledge Based and Intelligent Information and Engineering. Procedia Computer Science, 112(C), 2519-2528.

[5] Asma A. Al Jarullah. Decision tree discovery for the diagnosis of type II diabetes. International conference in innovations in information technology. New York: IEEE 2011

[6] Repalli P (2011). Prediction on diabetes using data mining approach. Stillwater: Oklahoma State University.

[7] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal.

[8] Komi, M., Li, J., Zhai, Y., & Zhang, X. (2017, June). Application of data mining methods in diabetes prediction. In Image, Vision and Computing (ICIVC), 2017 2nd International Conference on (pp. 1006-1010). IEEE.

[9] Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An Analytical Method for Diseases Prediction Using Machine Learning Techniques. Computers & Chemical Engineering. Volume 106, 2 November 2017, Pages 212-223

[10] Jingyu Xue, Fanchao Min and Fengying Ma(2020). Research on Diabetes Prediction Method Based on Machine Learning. Journal of Physics: Conference Series, Volume 1684, The 2020 International Seminar on Artificial Intelligence, Networking and Information Technology 18-20 September 2020, Shanghai, China.

[11] Emrana Kabir Hashi, Md. Shahid Uz Zaman and Md. Rokibul Hasan. An expert clinical decision support system to predict disease using classifcation techniques. In: International conference on electrical, computer and communication engineering (ECCE), 2017 IEEE, February 16–18,2017, Cox's Bazar, Bangladesh.

[12] Tejas N. Joshi, Prof. Pramila M. Chawan, Logistic Regression and SVM Based Diabetes Prediction System, International Journal for Technological Research in Engineering, Volume 5, Issue 11, July-2018.

[13] Kandhasamy, J. P., & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 45-51.

[14] Syed Muzamil, H. Balaji, N. Ch. S. N. Iyengar and Rennie D. Catiles (2017). Soft Computing approach to provide recommendation on PIMA diabetes. International Journal of Adavance Science and Technology. Vol. 106, pp. 19-32.

[15] Mustafa S. Kadhm, Ikhlas Watan Ghindawi, Duaa Enteesha Mhawi(2018). An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 6 (2018) pp. 4038-4041

[16] Changsheng Zhua, Christian Uwa Idemudiaa, Wenfang Fengb (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Informatics in Medicine Unlocked.Published by Elsevier Ltd.

[17] Priyanka Sonar, Prof. K. JayaMalini (2019). Diabetes prediction using different Machine learning approaches. Proceedings of the Third International Conference on Computing Methodologies and Communication.

[18] Talha Mahboob Alama, Muhammad Atif Iqbala, Yasir Alia, Abdul Wahabb, Safdar Ijazb, Talha Imtiaz Baigb, Ayaz Hussainc, Muhammad Awais Malikb, Muhammad Mehdi Razab , Salman Ibrarb , Zunish Abbasd(2019). A model for early prediction of diabetes. Informatics in Medicine Unlocked published by Elsevier Ltd.

[19] S. Saru, S.Subashree (2019). Analysis and prediction of diabetes using machine learning. International Journal of Emerging Technology and Innovative Engineering Volume 5, Issue 4, April 2019 (ISSN: 2394 – 6598)

[20] Md. Kamrul Hasan, Md. Ashraful Alam, Dola Das, Eklas Hossain 3 and Mahmudul Hasan (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE Access.

[21] K. Lakshmi Priya, Mourya Sai Charan Reddy Kypa(2020). A Novel Approach to Predict Diabetes by Using Naive Bayes Classifier. Proceedings of the Fourth International Conference on Trends in Electronics and Informatics (ICOEI 2020).

[22] Priyanka Rajendra, Shahram Latifi(2021). Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update 1 (2021) 100032.

[23] Jobeda Jamal Khanam, Simon Y. Foo(2021). A comparison of machine learning algorithms for diabetes prediction. The Korean Institute of Communications and Information Sciences (KICS). Publishing services by ElsevierB.V.

[24] Satish Kumar Kalagotla , Suryakanth V. Gangashetty, Kanuri Giridhar (2021). A novel stacking technique for prediction of diabetes. Computers in Biology and Medicine 135 (2021) 104554

[25] Neha Prerna Tigga, Shruti Garg, Prediction of type 2 diabetes using machine learning classification methods, Procedia Computer Science 167 (2020) 706–716.

[26] Deepti Sisodiaa, Dilip Singh Sisodia. Prediction of Diabetes using Classification Algorithms. International Conference on Computational Intelligence and Data Science (ICCIDS 2018). Published in Elsevier.

[27] Susmita Ray. A Quick Review of Machine Learning Algorithms. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon).

[28] J. U. Duncombe, “Infrared navigation—Part I: An assessment of feasibility,” IEEE Trans. Electron Devices, vol. ED-11, pp. 34-39, Jan. 1959.

[29] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, “Rotation, scale, and translation resilient public watermarking for images,” IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.