2023 10th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE) | 979-8-3503-2272-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICITACEE58587.2023.10277172

Machine Learning Diabetes Diagnosis Literature Review

Muhammad Rafian Wijoseno
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
muhammadrafianwijoseno@mail.ugm.ac.id

Adhistya Erna Permanasari
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
adhistya@ugm.ac.id

Azkario Rizky Pratama
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
azkario@ugm.ac.id

Abstract*—This paper presents a systematic literature review on the use of machine learning in diagnosing Diabetes Mellitus (DM). The study examines the application of machine learning algorithms and datasets in diabetes research. The findings highlight the effectiveness of Random Forest and the prevalence of the PIMA Indian dataset in this field. Early detection of diabetes is crucial for effective management and prevention of complications. However, challenges such as limited healthcare access and undiagnosed cases exist. The analysis reveals challenges related to dataset quality, sensitivity-specificity trade-offs, outliers, and missing data. To overcome these challenges, future research should expand the training dataset, incorporate additional parameters, and address outlier handling techniques. Feature selection methods and careful consideration of sensitivity-specificity trade-offs are also recommended. Despite these challenges, machine learning has the potential to improve diabetes diagnosis and enhance medical care. This study provides valuable insights for future advancements in machine learning-based diabetes diagnosis.

Keywords*—Diabetes Melitus, Diagnosis, CDSS, Machine Learning, Limitations

I. INTRODUCTION

Diabetes is a major public health problem worldwide. The prevalence of diabetes in adults over the age of 18 increased from 4.7% in 1980 to 8.5% in 2014. Diabetes was the leading cause of death in the world in 2019, after a sizable 70% increase from 2000 [1]. Diabetes Mellitus (DM) is a metabolic condition characterized by chronic hyperglycaemia and abnormalities of carbohydrate, lipid, and protein metabolism caused by deficiency of insulin production, insulin action, or both. [2]. Diabetes increases the risk of heart disease, peripheral artery, and cerebrovascular disease, as well as cataracts, erectile dysfunction, and non-alcoholic fatty liver disease. They are also more susceptible to infectious diseases such as tuberculosis, and outcomes tend to be worse [3].

Diabetes Mellitus is characterized by blood glucose (blood sugar) levels that are higher than usual, such as more than 200 mg/dl at the same time [4]. High blood pressure, obesity, family history of diabetes mellitus, age, lifestyle, and unhealthy diet can all contribute to the cause of diabetes mellitus. Diabetes mellitus generally has no obvious symptoms at first. As a result, many people are not aware of diabetes mellitus until they experience the consequences of uncontrolled diabetes mellitus, which can lead to long-term complications, damage to blood vessels and nerves, and even death [5]. Early identification of diabetes mellitus and better diagnosis using predictive models have generally been recommended to reduce mortality rates and improve decision making for prevention and further treatment. Predictive models implemented in clinical decision support systems (CDSS) can be used to help doctors assess the risk of diabetes mellitus and provide appropriate treatment to further manage the risk.

A decision support system (DSS) is a computer-based information system that functions to combine various information from different resources and then connect it so that it can assist in decision making, especially in complex problems. A clinical decision support system (CDSS) is a meeting point for patient-specific information and medical knowledge-based information. CDSS functions by adding medical knowledge obtained from current patient information and linking it so that it ultimately helps in deciding medical problems [6].

CDSS based on machine learning has recently been widely used in the health care field. Machine learning algorithms (MLA) such as backpropagation neural networks (BPNN), multilayer perceptron (MLP), logistic regression (LR), support vector machine (SVM), and random forest (RF) have previously been shown to be effective decision-making tools for prediction various diseases based on individual data in previous studies. Several studies have also demonstrated the benefits of hybrid models, such as nave Bayes (NB), Bayes net (BN), RF, and MLP majority voting, two-stacked SVM, and RF with linear models, in predicting various diseases.

Many types of research on diagnosing diabetes mellitus are published differently and complexly. Thus, there are many ways that can be used to diagnose diabetes. This literature review aims to identify and analyse the latest research on machine learning diagnosing diabetes mellitus by conducting a Systematic Literature Review (SLR).

The remainder of this literature review is organized as follows. We describe the research methods in part II. We offer the SLR results and discussion in Section III. In the final phase, we summarize our work.

II. METHODOLOGY

A. Review Method

The purpose of this study was to gather useful data from the most recent five years of research articles on machine learning analysis for diabetes diagnosis.

The SLR includes the planning stage, conduction stage, and reporting stage, this analysis evaluates various literature studies that discuss machine learning diagnosing diabetes.

An evaluation of the requirement for a literature review is done during the planning phase. The first step is to locate thorough information on machine learning for diabetes prediction. By defining the research questions, literature sources, search tactics, selection standards, data extraction techniques, and synthesis procedures, the review protocol was developed to constrict the scope of this study. This review was created using the PICOC technique, which stands for Population, Intervention, Comparison, Outcome, and Context.

The literature review is conducted using the review process at the Conduction step. In this phase, the literature is found, gathered, picked, extracted from, and synthesized. We then create a report for this SLR in the IEEE Xplore standard A4 double-column format.

B. Research Question

The research questions must be addressed during the critical evaluation of the most pertinent extracted articles because they reflect the SLR objectives. Below are the research questions for this SLR.

RQ1. What kind of dataset is used for diagnosing diabetes?

RQ2. What are the current trends in machine learning approaches for diabetes diagnosis?

RQ3. What are the challenges and limitations associated with using machine learning for diabetes diagnosis?

C. Search Strategy & Selection Criteria

Using a combination of the keywords we chose to extract the research articles from the relevant libraries, we conduct the SLR by searching the online literature. Words from the following research questions are included below:

Machine Learning, Diabetes, Diabetic, Diagnosis, Analysis, Detection, Prediction, Limitations

The following key words are used to complete the following search.

((“Machine Learning”) AND (“Diabetes” OR “Diabetic”) AND (“Diagnosis” OR “Analysis” OR “Detection” OR “Prediction))

IEEE and ScienceDirect are two well-known search libraries that were chosen for the extraction of material. Both libraries have unique features and methods for searching the catalogue. As a result, a few minor changes are made to the query string to produce more pertinent and acceptable material. It was necessary to conduct many searches using various combinations of the chosen keywords to find the Query. Table I displays the search query's results along with a few key parameters.

TABLE I
SELECTION CRITERIA

Inclusion Criteria	1. Research papers published in scientific peer-reviewed journals. 2. Research papers published from year 2019 till 2023. 3. Studies that implemented or proposed machine learning models to diagnose diabetes. 4. Research papers that provide limitations on their work.
Exclusion Criteria	1. Research papers that are unrelated with machine learning models to diagnose diabetes. 2. Research papers published before year 2019 or after 2023. 3. Research papers that do not contain any results. 4. Research papers that do not provide limitations on their work.

III. RESULT

Table II shows the digital library, the number of papers based on the search results, and the manual selection process.

TABLE II
NUMBER OF SELECTED PAPERS THROUGH THE SELECTION PROCESS

Search Result	Filtered Result	Selected Papers
1,841	217	21

The review approach, as described in the part above, was followed in order to get the results. Out of the initial search results, a total of 217 articles were identified. Following the application of specific inclusion and exclusion criteria, a refined selection of 21 papers was obtained for further analysis.

A. Dataset

To evaluate a machine learning model of diabetic diagnosis, it is important to use standard datasets that have been used by other researchers. The following are some of the datasets used in making machine learning models for diagnosing diabetes.

- PIMA Indian diabetes from the National Institute of Diabetes and Digestive and Kidney [7]–[19]. The dataset contains 768 instances and 9 attributes.
- Early-Stage Diabetes Risk Prediction Database (ESDRPD) [7], [20]. The ESDRPD contains 520 samples, each with 15 features and one output.
- UCI Machine Learning Repository [21] that has 17 attributes.

The following are some datasets obtained from electronic medical records.

- Electronic medical records in Hospital Selayang (HS) and Hospital Sultanah Bahiyah (HSB) [22].
- Electronic medical records in Smart Digital Clinic, METEDA [23]. The dataset consists of 147,664 patients seen for 15 years.

The following are some datasets obtained from datasets collected individually.

- PPG signal from the handle Empatica E4 wristband and class labels for 217 patients in a hospital in Cuenca, Ecuador [24].
- Khulna Diabetes Center, Khulna, Bangladesh [25] dataset has 340 instances and 26 features, [26] has 289 instances and 13 features, and [27] has 464 instances and 22 features.

B. Current Trends Machine Learning in Diagnosing Diabetes

Based on the review results obtained as many as 21 research papers. From the reviewed papers, the most widely used machine learning method is using Random Forest, as shown in Fig. 1.

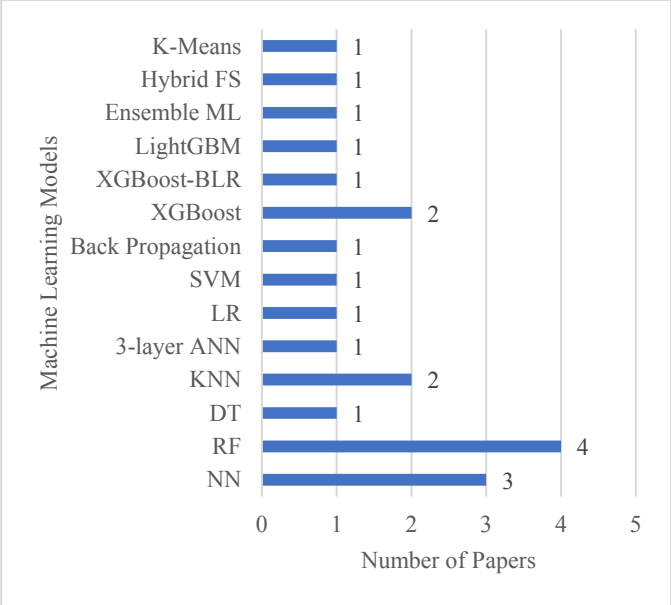

Fig. 1. Distribution of Selected Papers by Machine Learning Models

There are several different model validation techniques, but in this research article there are two types of validation methods used. Fig 2. shows the validation model used in machine learning research papers diagnosing diabetes.

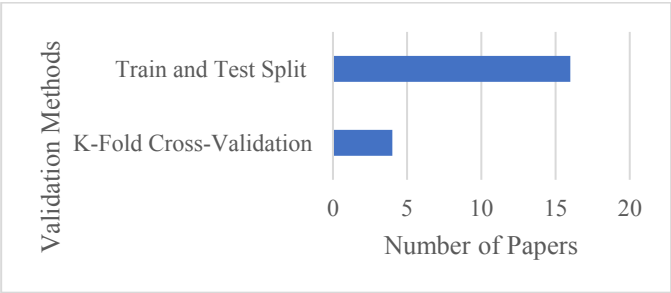

Fig. 2. Validation methods and number of papers

The train and test split validation [7], [9], [10], [12]–[22], [25]–[27] technique is done by dividing the data into a training group and a test group. The test data is retained and does not expose the machine learning model to it, until it is time to test the model. Similar to a test split validation, a k-fold cross-validation [8], [11], [23], [24] splits your data into more than two groups.

In Table III, it can be seen the research paper years, the machine learning model used, and the comparison model used in the research paper.

TABLE III
A RESUME OF SELECTED PAPERS FROM THE DIABETES DIAGNOSIS LITERATURE.

Ref.	Year	Dataset	Models	Baseline
[7]	2022	PIDD and ESDRPD	XGBoost-BLR	Previous work on the same dataset
[8]	2022	PIDD	LightGBM	KNN, LR, SVM, RF, and XGBoost
[9]	2020	PIDD	Three-Layered ANN	N/A
[10]	2021	PIDD	Neural Network	Discriminant, RT, CHAID, and SVM
[11]	2022	PIDD	RF	SVM, RF and KNN
[12]	2023	PIDD	Neural Network	SVM and MLP
[13]	2019	PIDD	K-Means	N/A
[14]	2022	PIDD	Innovative Back Propagation	DT Classifier
[15]	2023	PIDD	RF	SVM, DT, and XGBoost
[16]	2020	PIDD	KNN	N/A
[17]	2021	PIDD	KNN	N/A
[18]	2022	PIDD	RF	N/A
[19]	2022	PIDD	DT	N/A
[20]	2020	ESDRPD	Neural Network	Logistic, SVM, DT, RF, and Boosting
[21]	2021	UCI Repository	XGBoost	DT, RF, SVM, MLP, and LR
[22]	2022	EHR	SVM	LR, ANN, and CHAID
[23]	2022	EHR	XGBoost	Previous work on the same dataset
[24]	2021	Hospital	Hybrid FS	KNN, SVM, RF, and XGBoost
[25]	2020	Hospital	Ensemble Machine Learning	N/A
[26]	2021	Hospital	LR	XGBoost and RF
[27]	2020	Hospital	RF	AdaBoost

C. Challenges and Limitations with using Machine Learning for Diabetes Diagnosis

Several papers mention the challenges and limitations in using machine learning to diagnose diabetes. In papers [7], [9]–[11], [13], [20]–[22], [24], [26], [27] there are limitations in the dataset used, future work should focus on improving the prediction accuracy by expanding the training dataset and incorporating additional parameters. The performance of the application can be further enhanced by identifying and integrating other relevant variables and increasing the size of the training dataset.

In paper [23] states that there is Sensitivity vs. Specificity Trade-off. The prediction models in the study were developed to maximize sensitivity (true positive rate), potentially resulting in a trade-off with specificity (true negative rate). In paper [8] it is stated that the removal of outliers from the dataset is mentioned as a factor that affected the accuracy. Outliers can have a significant impact on model performance, and their handling can introduce variability in the results.

Paper [14] discusses the limitations of accuracy evaluation on larger datasets and the absence of feature selection methods in the analysis. Similarly, in paper [15], the study acknowledges the limitation of not utilizing feature selection methods to improve performance. In paper [12], it is noted that while the implemented models or algorithms are not significantly affected by outliers, their presence during the training process may slightly reduce accuracy. Additionally, the choice of data preprocessing method can lead to variations in outcome values.

In paper [25], challenges related to the collection of accurate patient information and instances of missing data in the dataset are mentioned. The authors of paper [16] identify limitations in the model, including the computational laziness of the KNN classifier and its inability to effectively handle outliers during the Min-Max normalization process. The KNN algorithm also requires substantial memory and processing time as it utilizes the entire dataset for prediction. Moreover, the presence of a few outliers has a significant impact on feature scaling during the Min-Max normalization procedure.

Paper [17] observes that increasing a specific value improves classification reliability but leads to less distinct boundaries between classes. In paper [18], it is concluded that the model fails to eliminate the time delay between actual glucose levels and continuous glucose monitoring. Additionally, the complete justification of the correlation between blood pressure and diabetes remains unachieved.

IV. Conclusions

In conclusion, this study contributes to the existing knowledge on the application of machine learning in diabetes research by providing a comprehensive analysis of the current state of the field. The findings highlight the significance of the PIMA Indian dataset and the effectiveness of Random Forest as a machine learning algorithm in addressing the complexities of diabetes diagnosis. This knowledge is valuable for researchers, healthcare professionals, and policymakers involved in diabetes management.

The significance of the findings extends to practical implications for healthcare practice and policy. Early detection of diabetes is crucial for effective management and prevention of complications. The study emphasizes the importance of leveraging machine learning technology to identify diabetes at its early stages or predict diabetes risk, especially considering the challenges in healthcare access, particularly in rural areas. By utilizing machine learning algorithms, healthcare providers can enhance medical care and treatment for individuals with diabetes, leading to improved outcomes and quality of life.

The study also highlights several challenges and limitations in machine learning-based diabetes diagnosis, such as dataset quality issues, sensitivity-specificity trade-offs, outlier impact, and missing data handling. These challenges provide a roadmap for further research and action. Future studies should focus on expanding the training dataset, incorporating additional parameters, and addressing outlier handling techniques to improve prediction accuracy. Additionally, the use of feature selection methods and careful consideration of the sensitivity-specificity trade-off will contribute to more accurate outcomes. These recommendations guide future research endeavors to overcome the challenges and enhance the reliability and effectiveness of machine learning-based diabetes diagnosis.

Overall, this study's comprehensive analysis and insights into machine learning-based diabetes diagnosis contribute to the body of knowledge in the field. By addressing the identified challenges and leveraging the potential of machine learning, there is a promising opportunity to significantly improve care and treatment for individuals with diabetes. The findings of this study provide valuable guidance for future research and pave the way for advancements in machine learning-based diabetes diagnosis, ultimately benefiting individuals with diabetes and the healthcare community as a whole.

ACKNOWLEDGMENT

The authors express their gratitude to UGM for their support and assistance in conducting this research. The provision of facilities and financial support through Final Project Recognition Grant Universitas Gadjah Mada Number 5075/UN1.P.II/Dit-Lit/PT.01.01/2023 is greatly appreciated by the researchers.

References

[1] W. H. Organization, “Improving diabetes outcomes for all, a hundred years on from the discovery of insulin: report of the Global Diabetes Summit,” 2021.

[2] I. B. Wayan Kardika, S. Herawati, and I. W. P. Surtiya Yasa, “PREANALITIC AND INTERPRETATION BLOOD GLUCOSE FOR DIAGNOSE DIABETIC MELITUS,” *E-Jurnal Medika Udayana*, vol. 2, no. 10, pp. 1707–1721, 2013.

[3] W. H. Organization, “HEARTS D: diagnosis and management of type 2 diabetes,” World Health Organization, 2020.

[4] D. Hestiana, “FAKTOR-FAKTOR YANG BERHUBUNGAN DENGAN KEPATUHAN DALAM PENGELOLAAN DIET PADA PASIEN RAWAT JALAN DIABETES MELLITUS TIPE 2 DI

KOTA SEMARANG," *JHE (Journal of Health Education)*, vol. 2, no. 2, pp. 137–145, 2017.

[5] American Diabetes Association, *American Diabetes Association Standards of Medical Care In Diabetes*, vol. 41. USA: ADA, 2018.

[6] S. Aghazadeh, A. Q. Aliyev, and M. Ebrahimnejad, "The role of computerizing physician orders entry (CPOE) and implementing decision support system (CDSS) for decreasing medical errors," in *2011 5th International Conference on Application of Information and Communication Technologies (AICT)*, IEEE, Oct. 2011, pp. 1–3. doi: 10.1109/ICAICT.2011.6110916.

[7] Y. Wu *et al.*, "Novel binary logistic regression model based on feature transformation of XGBoost for type 2 Diabetes Mellitus prediction in healthcare systems," *Future Generation Computer Systems*, vol. 129, pp. 1–12, Apr. 2022, doi: 10.1016/j.future.2021.11.003.

[8] C. Charitha, A. D. Chaitrasree, P. C. Varma, and C. Lakshmi, "Type-II Diabetes Prediction Using Machine Learning Algorithms," in *2022 International Conference on Computer Communication and Informatics, ICCCI 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICCCI54379.2022.9740844.

[9] K. Lakhwani, S. Bhargava, K. K. Hiran, M. M. Bundele, and D. Somwanshi, "Prediction of the Onset of Diabetes Using Artificial Neural Network and Pima Indians Diabetes Dataset," in *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering, ICRAIE 2020 - Proceeding*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/ICRAIE51050.2020.9358308.

[10] B. Jain, N. Ranawat, P. Chittora, P. Chakrabarti, and S. Poddar, "A machine learning perspective: To analyze diabetes," *Mater Today Proc*, Feb. 2021, doi: 10.1016/j.matpr.2020.12.445.

[11] S. S. Bhat, V. Selvam, G. A. Ansari, and M. D. Ansari, "Analysis of Diabetes mellitus using Machine Learning Techniques," in *2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)*, IEEE, Nov. 2022, pp. 1–5. doi: 10.1109/IMPACT55510.2022.10029058.

[12] H. Song and S. Lee, "Implementation of Diabetes Incidence Prediction Using a Multilayer Perceptron Neural Network," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, Dec. 2021, pp. 3089–3091. doi: 10.1109/BIBM52615.2021.9669583.

[13] M. Raihan, Md. T. Islam, F. Farzana, Md. G. M. Raju, and H. S. Mondal, "An Empirical Study to Predict Diabetes Mellitus using K-Means and Hierarchical Clustering Techniques," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, Jul. 2019, pp. 1–6. doi: 10.1109/ICCCNT45670.2019.8944552.

[14] A. Pradeepik and R. Sabitha, "Analysis of Anomaly Detection of Diabetes Using Decision Tree Classifier and an Innovative Back Propagation Algorithm using Fit as a Parameter," in *2022 International Conference on Business Analytics for Technology and Security (ICBATS)*, IEEE, Feb. 2022, pp. 1–6. doi: 10.1109/ICBATS54253.2022.9759012.

[15] P. Rani, R. Lamba, R. K. Sachdeva, P. Bathla, and A. N. Aledaily, "Diabetes Prediction Using Machine Learning Classification Algorithms," in *2023 International Conference on Smart Computing and Application (ICSCA)*, IEEE, Feb. 2023, pp. 1–5. doi: 10.1109/ICSCA57840.2023.10087827.

[16] S. C. Gupta and N. Goel, "Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, Aug. 2020, pp. 980–986. doi: 10.1109/ICSSIT48917.2020.9214129.

[17] V. Lopatka, I. Meniailov, and K. Bazilevych, "Classification and Prediction of Diabetes Disease Using Modified k-neighbors Method," in *2021 IEEE 12th International Conference on Electronics and Information Technologies (ELIT)*, IEEE, May 2021, pp. 46–50. doi: 10.1109/ELIT53502.2021.9501151.

[18] A. K. Shrivastava, K. V, K. S, and S. M, "Early Diabetes Prediction using Random Forest," in *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, Aug. 2022, pp. 1154–1159. doi: 10.1109/ICESC54411.2022.9885683.

[19] P. Thotad, G. R. Bharamagoudar, and B. S. Anami, "Predictive Analysis of Diabetes Mellitus Using Decision Tree Approach," in *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, IEEE, Aug. 2022, pp. 1–7. doi: 10.1109/ASIANCON55314.2022.9909122.

[20] J. Ma, "Machine Learning in Predicting Diabetes in the Early Stage," in *Proceedings - 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence, MLBDBI 2020*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 167–172. doi: 10.1109/MLBDBI51377.2020.00037.

[21] M. A. R. Refat, Md. Al Amin, C. Kaushal, M. N. Yeasmin, and M. K. Islam, "A Comparative Analysis of Early Stage Diabetes Prediction using Machine Learning and Deep Learning Approach," in *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, IEEE, Oct. 2021, pp. 654–659. doi: 10.1109/ISPCC53510.2021.9609364.

[22] N. A. Azit, S. Sahran, V. M. Leow, M. Subramaniam, S. Mokhtar, and A. M. Nawi, "Prediction of hepatocellular carcinoma risk in patients with type-2 diabetes using supervised machine learning classification model," *Heliyon*, vol. 8, no. 10, p. e10772, Oct. 2022, doi: 10.1016/j.heliyon.2022.e10772.

[23] A. Nicolucci *et al.*, "Prediction of complications of type 2 Diabetes: A Machine learning approach," *Diabetes Res Clin Pract*, vol. 190, Aug. 2022, doi: 10.1016/j.diabres.2022.110013.

[24] A. Prabha, J. Yadav, A. Rani, and V. Singh, "Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier," *Comput Biol Med*, vol. 136, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104664.

[25] Md. T. Islam, M. Raihan, F. Farzana, N. Aktar, P. Ghosh, and S. Kabiraj, "Typical and Non-Typical Diabetes Disease Prediction using Random Forest Algorithm," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, Jul. 2020, pp. 1–6. doi: 10.1109/ICCCNT49239.2020.9225430.

[26] Md. M. Hassan, Z. J. Peya, S. Mollick, Md. A.-M. Billah, Md. M. Hasan Shakil, and A. U. Dulla, "Diabetes Prediction in Healthcare at Early Stage Using Machine Learning Approach," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, Jul. 2021, pp. 01–05. doi: 10.1109/ICCCNT51525.2021.9579869.

[27] Md. T. Islam, M. Raihan, N. Aktar, Md. S. Alam, R. R. Ema, and T. Islam, "Diabetes Mellitus Prediction using Different Ensemble Machine Learning Approaches," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, Jul. 2020, pp. 1–7. doi: 10.1109/ICCCNT49239.2020.9225551.