ISSN No:-2456-2165

Diabetes Prediction using Machine Learning

[1]Sahil Kumar Suman
Department of Computer Science and Engineering
Chandigarh University, Punjab, India

[2]Natasha Sharma
Department of Computer Science and Engineering
Chandigarh University, Punjab, India

[3]Udeshna Saikia
Department of Computer Science and Engineering
Chandigarh University, Punjab, India

[4]Dhiti
Department of Computer Science and Engineering
Chandigarh University, Punjab, India

[5]Rahul Chauhan
Department of Computer Science and Engineering
Chandigarh University, Punjab, India

[6]Nandini Singh
Department of Computer Science and Engineering
Chandigarh University, Punjab, India

Abstract:- Diabetes has been recorded as a serious global health issue today. It's a long-term metabolic disease that takes place when blood glucose levels elevate in the human body. Early and accurate diabetes diagnosis is essential for managing the condition precisely and will prevent complications quickly. This count proposes a comprehensive and effective machine-learning method for detecting and treating diabetes. The dataset that was used contains many clinical and demographic variables such as age, BMI, family history and various blood test results. To identify the most relevant variables, the technique prioritizes the data to control for missing values and to normalize features. The next stepis to go through a strict feature selection process. For the training and validation of the model, SVM, RFM, Logistic Regression, and Support Vector Machines (SVM) are just a few of the machine learning algorithms that are employed. The performance of each of these algorithms is checked using metrics like accuracy, redundancy, uniqueness, and receiver operating characteristic (ROC) curve area. An ensemble perspective is also explored to combine the benefits of multiple models and increase overall predicting power. The recommended model is tested on various test datasets for assessment purposes of its generalizability. The main purpose of the project is to create a robust and trustworthy diabetes detection tool that can be used in clinical settings to aid medical professionals with advanced diagnosis and individualized treatment planning. The results demonstrate growing performance and the potential for machine learning to increase diabetes detection accuracy. The importance of the proposed model to subtle patterns in different patient data sets suggests that it could apply to a large range of demographics. This work lays the root level for future analysis into enhancing and expanding the capabilities of diabetes detection models, which will advance ongoing efforts to apply machine learning to healthcare applications.

Keywords:- *Diabetes, Early Diagnosis, Machine Learning, Ac-Curacy, Healthcare Applications.*

I. INTRODUCTION

Millions of people worldwide are adversely affected by diabetes mellitus, a chronic metabolic disease marked by persistent hyperglycemia and has become a major public health concern. Early intervention and specified treatment plans are made possible by punctual and accurate diabetes prediction, which isneeded in proactive healthcare management [1]. The artificial intelligence field of machine learning, which is evolving quickly, has shown great promising results in the area of healthcare, especially in terms of diagnosing and predicting complicated medical conditions. To make a more proactive and preventive approach to healthcare, this study focuses on utilizing machine learning techniques to forecast when diabetes will manifest.

Age-long diabetes analysis techniques often depend on clinical risk factors and statistical models. These methods might not have the sensitivity and specificity required for accurately and precisely identifying people who are at risk, though [2]. A favourable substitute is given by machine learning, which can identify patterns and relationships in large and varied datasets. Ma- chine learning models can analyze a wide variety of patient-related features, like clinical history, laboratory results, and demo-graphic data, RTML by utilizing highly accessible and developedalgorithms.

The purpose of this research is to develop a trustworthy and precise diabetes prediction model. That will assist in recognizing the disease early and enlightening on its main contributing factors. By permitting healthcare professionals to apply interventions and preventive decisions customized to each patient's unique profile, these models have the potential to completely bring a change in the way that healthcare is provided**.**

This research is an attempt to mark the limitations of current prediction models and contribute to the ongoing efforts to upgrade and update the accuracy and reliability of early diabetes detection [3]. The final objective is to make a better future for every patient in humankind.

ISSN No:-2456-2165

II. MACHINE LEARNING TECHNIQUES

Utilizing machine learning methodologies has been severelybenefiting diabetes detection in a never-seen way before. It can analyze large and great numbers of datasets of various types such as insulin and non-insulin, family history etc. Below are the ML techniques that are being highly used for prediction of diseases worldwide:

A. Logistic Regression:

This binary classification algorithmsimulates the probability that an instance belongs to a particular class. Well, it is frequently used in diabetes detection to calculate a person's risk of developing diabetes based on a variety of input features.

B. Decision Trees:

Decision trees are data structures that are ensemble trees, with each node representing a choice made in response to a specific feature [4]. These are employed in the detection of diabetes to establish classification rules, which facilitate easy interpretation and comprehension of the design- decision-making process.

C. Random Forests:

During training, the Random Forest learning technique builds multiple decision trees and outputs the class mode. It is renowned for being flexible and able to manage sizable datasets with a variety of properties, which qualifies it for diabetes prediction.

D. Support Vector Machines (SVM):

Strong classification performance is achieved by SVM, an algorithm that handles both linear and non-linear data. SVM looks for the hyper-planethat best divides data points into distinct classes according to their features to detect diabetes.

E. Neural Networks:

The use of profound learning, especially neural systems, has grown in popularity for the diagnosis of diabetes. Deep neural networks and multi-layer perceptrons (MLPs) are capable of identifying complex patterns in data thatmay be linked to an increased risk of diabetes by capturing intricate relationships.

F. K-Nearest Neighbors (KNN):

KNN is a non-parametriclearning algorithm that is instance-based a fresh occurrence classified by to feature space's majority class of its k-nearest neighbors. KNN predicts an individual's diabetes status by identifying and grouping similar individuals.

G. Naive Bayes:

The algorithm based on probability The Bayes theorem is the foundation of naive Bayes. Well, it is computationally efficient because it assumes that features are conditionally independent [5]. It is used in the detection of diabetes to calculate the probability that an individual possesses the illness based on characteristics that are seen.

H. Ensemble Learning:

Ensembled approaches improveperformance by joining predictions from several models. By applying strategies like bagging and boosting to different base models, the diabetes detection system's accuracy and generalizability can be improved.

I. Feature Selection Techniques:

Recursive feature elimination (RFE) and feature importance from tree-based models aretwo feature selection techniques that assist in determining the most pertinent features for diabetes prediction, thereby lowering dimensionality and possibly enhancing model performance.

Based on the objectives of the diabetes detection task and the characteristics of the dataset, these machine-learning techniques can be used singly or in combination. The selection of the algorithm is influenced by variables such as the type of data, requirements for interpretability, and the intended ratio of sensitivity to specificity in the prediction of diabetes.

III. IMPORTANCE OF FEATURE SPECIFICATION

Ensuring effective creativity in machine learning for diabetes detection is very crucial and for that, we need Feature Specification. Specific characterization related to the relevant and targeted disease is necessary for differentiating between healthy and diseased states [6]. The model's capacity for accurate prediction is directly impacted by the discriminative strength of its features, and robustness, interpretability, and dimensionality reduction are all enhanced by meticulous feature selection. Additionally, while making healthcare applications ethics and prejudice of an individual must always be kept in mind being an example of careful feature definition and to lessen possible inequalities in prediction [7]. For achieving all critical aspects of properties of a good model, the machine learning techniques also must justify all qualities. Following are the machine learning techniques, we use for diabetes detection in our study:

A. Random Forest:

Using an ensemble learning approach called Random Forest, one may increase overall forecast curacy and resilience by merging numerous decision trees' predictions. The technique generates a large number of decision trees during the training phase. In the context of feature definition within a Random Forest model [8]. To provide variation among the different trees, a randomly chosen subset of the dataset's characters is used in the construction of each tree. This unpredictability promotes generalization, reduces overfitting, and captures different facets of the data [9]. In a Random Forest model, determining each input feature's relevance to the prediction task is a necessary step in the feature specification process. Furthermore, Random Forest is a more optimal option for a variety of datasets due to its versatility in handling both numerical and categorical features without needing a lot of preprocessing.

B. Linear Regression:

Regardless of being a fundamental statistical technique, linear regression has particular advantages when it comes to diabetes prediction. Given that interpretability is frequently just as important as predictive accuracy in healthcare scenarios, healthcare professionals can easily understand how each feature affects the predicted outcome thanks to the simple nature of linear regression [10]. Each feature's coefficient from linear regression can be readily understood and shows the direction and strength of each feature's influence on diabetes prediction. Linear regression is a useful tool in healthcare settings where transparent decision-making is essential, especially if the goal is to establish a clear understanding ofthe impact of individual features on diabetes risk.

IV. USING RTML DATASET

As of January 2022, to the best of mine, no well-defined or well-recognized dataset referred to as "RTML" has been created specifically to diagnose diabetes. However, since then, other data sets could have been produced, or RTML might be related to a particular dataset within a particular context or organization. A variety of popular databases are often employed in research and development for machine learning-based assessments of diabetes. Some of these include:

A. Pima Indians Diabetes Database:

This dataset, which contains details about age, blood pressure, BMI, and other medical conditions, is frequently used to predict diabetes. It started with research done on Pima Indian women.

B. UCI Diabetes Dataset:

A set of data for diabetes identification is available from the UCI Machine Learning Repository and includes information on sex, BMI, average blood pressure, BMI, age, and six blood serum parameters. values.

C. Diabetes Dataset from Kaggle:

Kaggle, a platform that hosts data science contests, maintains datasets for the identification of hyperglycemia [11]. These datasets may vary in terms ofthe quantity and qualities they cover.

D. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Dataset:

NIDDK provides access to datasetson hypertension research, including clinical information suitablefor machine learning applications.

I suggest looking through the most recent sources, papers onthe topic, or the particular platform or organization related to "RTML" for more up-to-date and trustworthy data if "RTML" refers to a specific collection of data added after my previous update, or inside a certain context or organization [12]. Furthermore, a range of datasets are frequently available on websites like Kaggle, the UCI Deep Learning Repository, and others; hence, searching these sources may provide datasets appropriatefor diabetes diagnosis [13].

V. RESULT AND ANALYSIS

Decision trees, SVM, random forests, and KNN were amongthe machine learning techniques we employed for our diabetes forecasting model. The secret to creating an accurate model is machine learning with superior features that will help health professionals in their work and make the world a more livable place free of disease or with more healthful ways to cure disease.

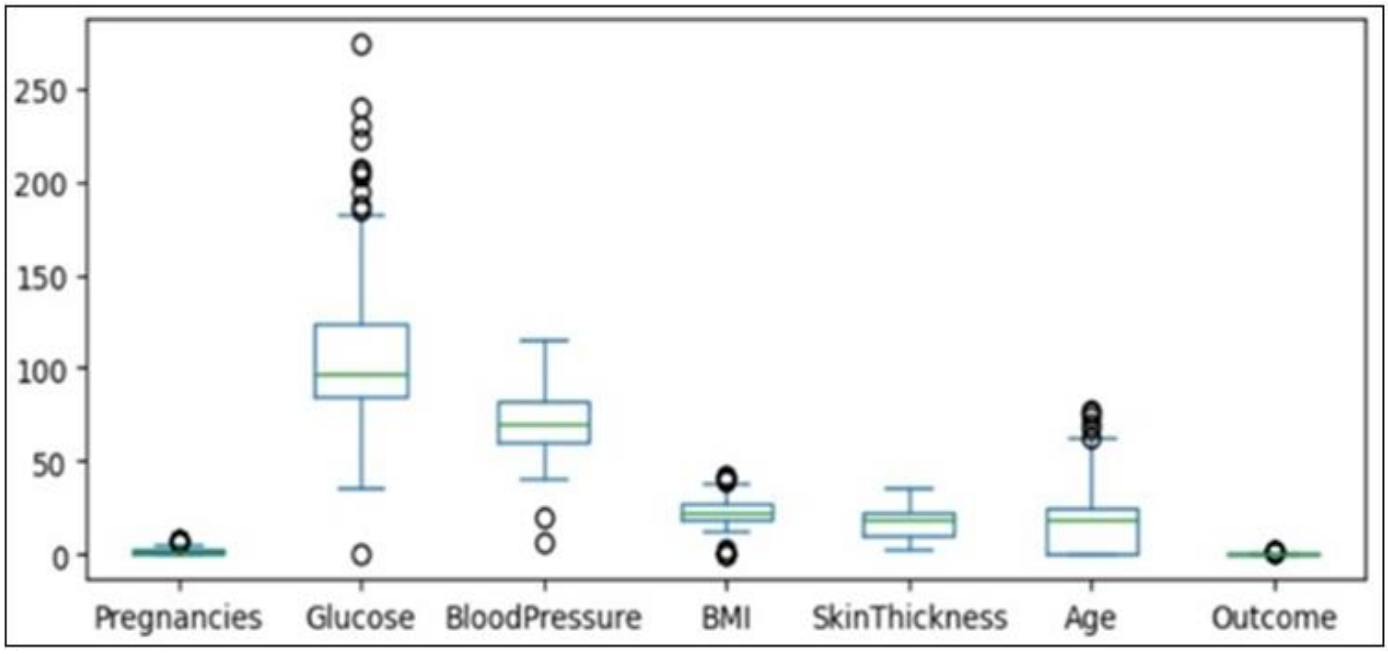

Fig 1 Dataset Graph using RTML

ISSN No:-2456-2165

In the above fig, we have an output of a dataset from various kinds of diabetes-infected people who are tested with RTML. We have no. of people distinguished in each category of predictive diabetes inflation for our future analysis and work uponit to find out the accuracy of our research.

VI. WITHOUT USING RTML DATASET

Using previous clinical information to create a prediction model is the method of detecting diabetes without real-time monitoring using machine learning. At first, a collection of datais selected, including conventional attributes like age, BMI, blood pressure, and cholesterol, along with labels identifying whether or not diabetes is present. Following a series of steps, the dataset is divided into training and testing sets. data preparation operations, like managing lost values and generalizing numerical properties [14]. To identify patterns and correlationsbetween diabetes outcomes and feature sets, machine learning methods, such as logistic regression or decision trees are used for training datasets.

After the model is trained, it is evaluated to check its accuracy, precision and recall. To improve the model's predictive power and enhance its parameters, hyperparameter modifying may be carried out. It may be used to predict diabetes in new instances once the model has been verified, which makes it a beneficial tool in the healthcare field. The model's truthfulness is assured by routine maintenance and monitoring, which makesit possible to alter it in response to changing population health features or evolving datasets [15]. With its roots in historical data, this machine learning technique effectively detects diabetes without requiring real-time monitoring, al- allowing for prompt and precise adjustments in healthcare procedures.

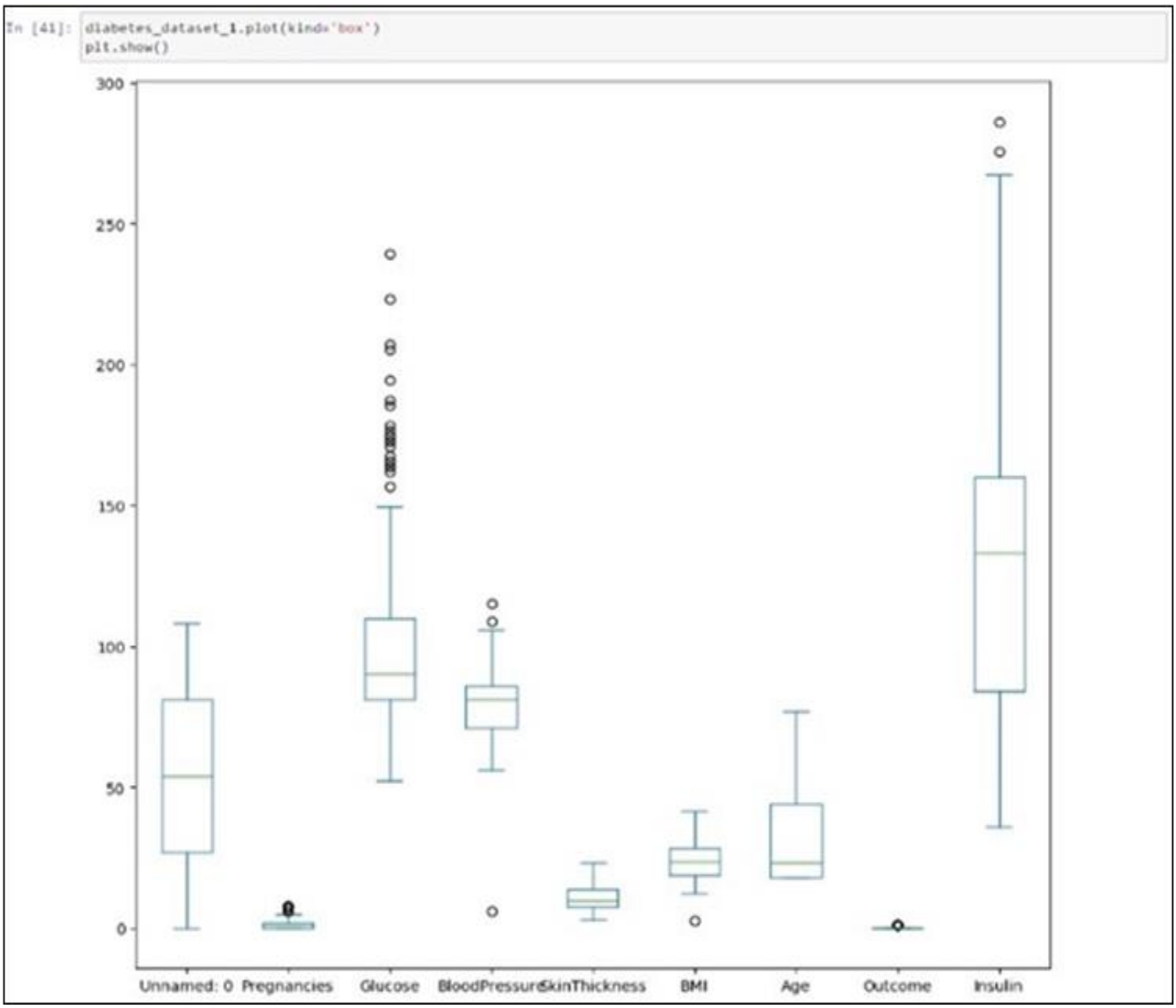

Fig 2 Dataset Graph without RTML.

In the above fig, we have an output of dataset from variouskinds of diabetes-infected people who are tested without RTML. We have no. of people distinguished in each category of predictive diabetes inflation for our future analysis and workupon it to find out the accuracy of our research.

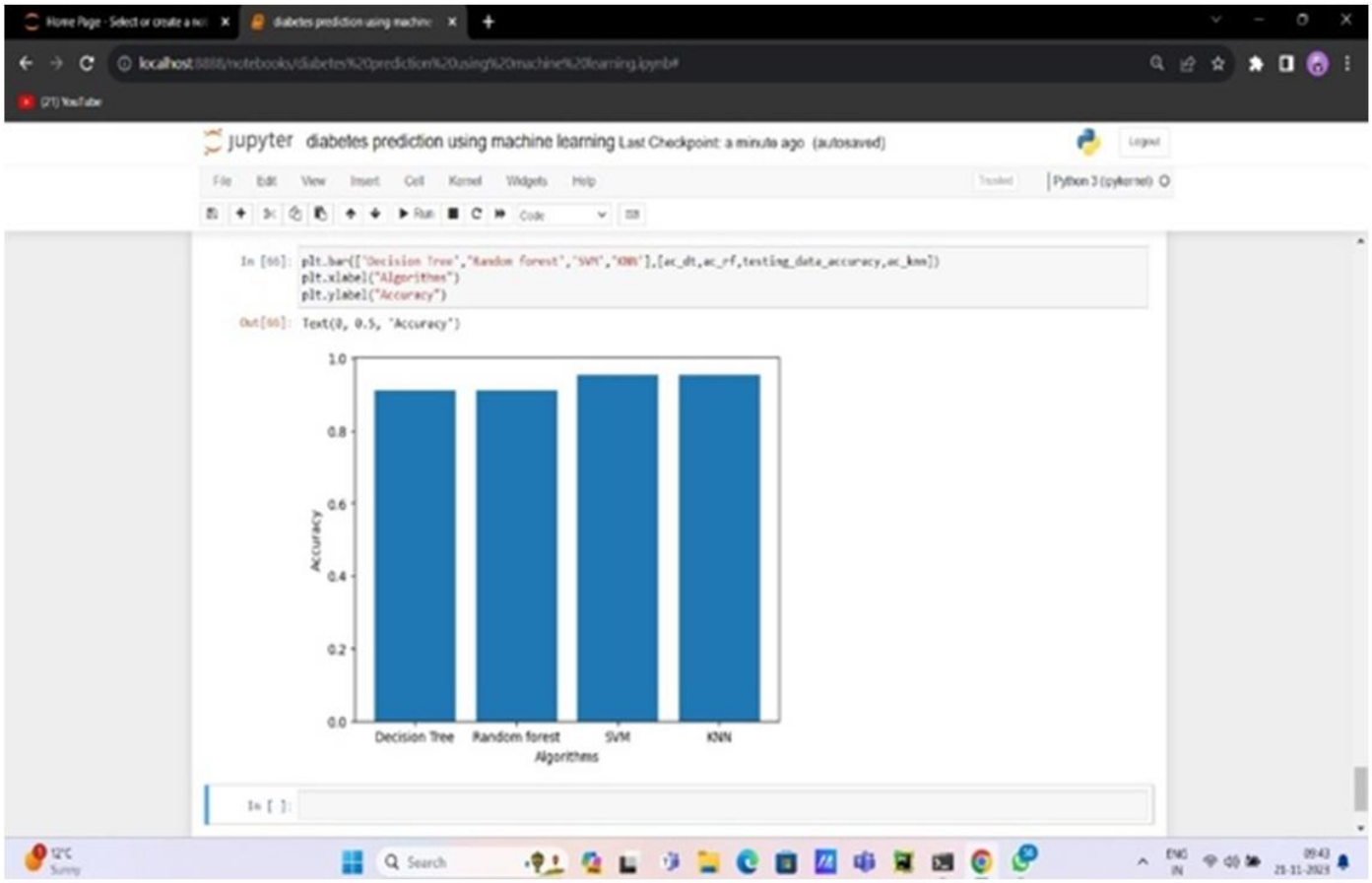

Fig 3 Accuracy Levels using these ML Techniques.

Here in this figure, we record the accuracies that we have found for predicting diabetes using various ML techniques such as Decision Tree, Random Forest, SVM and KNN.

VII. CONCLUSION

To conclude, we have made a research paper with an effort tomake a better future for humankind with early detection of diseases like diabetes which happens on inflation of blood sugar levels in the human body. This disease varies on a wide range of factors including BMI, age, family history, eating habits etc. However, the harsh outcomes of this disease can be prevented through early detection and its on-time curation. With the developing and evolving technologies, this is possible that the fatal rates or disease rates of humans go down in the coming ages. For such a thing to happen we have machine learning techniques that aid in creating this kind of model which can be considered a boon in the field of healthcare. To bring a better change for humankind and build a disease-free environment for the world, this search paper is written.

REFERENCES

[1]. Y. Dubey, P. Wankhede, T. Borkar, A. Borkar and K. Mitra, "Diabetes Prediction and Classification using Machine Learning Algorithms," 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), Dhaka, Bangladesh, 2021, pp. 60-63, doi: 10.1109/BECITHCON54710.2021.9893653.

[2]. S. A. Shampa, M. S. Islam and A. Nesa, "Machine Learning-based Diabe- tes Prediction: A Cross-Country Perspective," 2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM), Gazipur, Bangladesh,2023, pp.1-6,doi: 10.1109/NCIM59001.2023.10212596.

[3]. E. Daniel, J. Johnson, U. A. Victor, G. V. Aditya and S. A. Sibby, "An Efficient Diabetes Prediction Model using Machine Learning," 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2023, pp. 1202-1208, doi: 10.1109/ICESC57686.2023.10193277.

[4]. S. S et al., "A Comparative Analysis of Diabetes Prediction Models using Machine Learning Algorithms," 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2022, pp. 261-265, doi: 10.1109/ICACCS54159.2022.9785280.

[5]. C. Charitha, A. Devi Chaitrasree, P. C. Varma and C. Lakshmi, "Type-II Diabetes Prediction Using Machine Learning Algorithms," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-5, doi: 10.1109/ICCCI54379.2022.9740844.
[6]. S. Samet, M. R. Laouar and I. Bendib, "Use of Machine Learning Techniques to Predict Diabetes at an Early Stage," 2021 International Conference on Networking and Advanced Systems (ICNAS), Annaba, Algeria, 2021, pp. 1-6, doi: 10.1109/ICNAS53565.2021.9628903.
[7]. S. Mahajan, P. K. Sarangi, A. K. Sahoo and M. Rohra, "Diabetes Mellitus Prediction using Supervised Machine Learning Techniques," 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 2023, pp. 587-592, doi: 10.1109/InCACCT57535.2023.10141734.
[8]. M. Pal, S. Parija and G. Panda, "Improved Prediction of Diabetes Mellitus using Machine Learning Based Approach," 2021 2nd International Conference on Range Technology (ICORT), Chandipur, Balasore, India, 2021, pp. 1-6, doi: 10.1109/ICORT52730.2021.9581774.
[9]. P. Tumuluru, L. R. Burra, K. K. Sushanth, S. N. Vali, C. H. M. H. SaiBaba and P. Yellamma, "DPMLT: Diabetes Prediction Using Machine Learn- ing Techniques," 2022 International Conference on Electronics and Re- newable Systems (ICEARS), Tuticorin, India, 2022, pp. 1127-1133, doi: 10.1109/ICEARS53579.2022.9751944.
[10]. L. H.N., A. S. Reddy and K. Naidu, "Analysis of Diabetic Prediction Using Machine Learning Algorithms on BRFSS Dataset," 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), Ti-runelveli, India, 2023, pp. 1024-1028, doi: 10.1109/ICOEI56765.2023. 10125804.
[11]. S. Samet, M. R. Laouar and I. Bendib, "Diabetes mellitus early-stage risk prediction using machine learning algorithms," 2021 International Conference on Networking and Advanced Systems (ICNAS), Annaba, Alge- ria, 2021, pp. 1-6, doi: 10.1109/ IC-NAS53565.2021.9628955.
[12]. M. Paliwal and P. Saraswat, "Research on Diabetes Prediction Method Based on Machine Learning," 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2022, pp. 415-419, doi: 10.1109/IC- TACS56270.2022.9988050.
[13]. K. Sidana, "Prediction of Diabetes using Machine Learning Algorithms," 2023 11th International Conference on Internet of Everything, Micro-wave Engineering, Communication and Networks (IEMECON), Jaipur, India, 2023, pp. 1-6, doi: 10.1109/IEMECON56962.2023.10092335.
[14]. B. Rathi and F. Madeira, "Early Prediction of Diabetes Using Machine Learning Techniques," 2023 Global Conference on Wireless and Optical Technologies (GCWOT), Malaga, Spain, 2023, pp. 1-7, doi: 10.1109/GCWOT57803.2023.10064682.
[15]. P. Dalve, D. Bobby, A. Marathe, A. Dusane and S. Daga, "Comparison of Performance of Machine Learning Algorithms for Diabetes Detection," 2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2023, pp. 1-7, doi: 10.1109/ICAECT57570.2023. 10118315.