Unraveling Diabetes Detection: A Comparative Study of Machine Learning Approaches

Jiten Agarwal
School of Computer Engineering
KIIT Deemed to be University,
Bhubaneswar, India 751024
Email: jitenagarwal05@gmail.com

Hardik Chauhan
School of Computer Engineering
KIIT Deemed to be University,
Bhubaneswar, India 751024
Email: hardik1057@gmail.com

Sneha Gupta
School of Computer Engineering
KIIT Deemed to be University,
Bhubaneswar, India 751024
Email: sneha9777gupta@gmail.com

Shreya Mondal
School of Computer Engineering
KIIT Deemed to be University,
Bhubaneswar, India 751024
Email: shreyamondal2362@gmail.com

Saumyadeep Mahanta
School of Computer Engineering
KIIT Deemed to be University,
Bhubaneswar, India 751024
Email: saumyadeepmahanta@gmail.com

Santos Kumar Baliarsingh
School of Computer Engineering
KIIT Deemed to be University,
Bhubaneswar, India 751024
Email: santos.baliarsinghfcs@kiit.ac.in

***Abstract*—The core focus of our work has been the utilization of Machine Learning (ML) to examine healthcare, with the main aim of recognizing early diabetes cases. Giving an early diagnosis of diabetes to prevent the growth of serious complications is the main task to achieve for a thorough treatment of the problem. Diabetes, being a chronic and enduring condition, presents a significant global challenge, being a primary contributor to premature mortality. In this work, we employed ML and statistical methodologies namely Decision Trees (DT), k-Nearest Neighbours (kNN), Naive Bayes (NB), Random Forest Classifier (RFC), and Support Vector Machine (SVM) to conduct a comparative evaluation of these modeling forecasts as an essential tool for diabetes diagnosis. While keeping the experimentation and evaluation rigorous, we study how successful the models are and their advantages and disadvantages within the healthcare field. We aim to disseminate findings that represent the practical applications within a real-world healthcare system. Moreover, it forms the foundation for individualized blood sugar diagnosis and therapy, ultimately enhancing mental well-being. By accomplishing these steps, personalized medicine could potentially offer a solution for diabetes treatment in the future.**

***keywords*—Diabetes, Machine learning, Comparative evaluation.**

I. Introduction

A. Background

Diabetes, a chronic metabolic disorder characterized by elevated blood sugar levels brought on by inadequate insulin synthesis or utilization poses significant health risks worldwide. If left untreated, it can increase the likelihood of conditions such as cardiovascular disease, kidney damage, neuropathy, and vision impairment. While Type 1 diabetes typically develops in childhood or in adolescence and necessitates lifelong insulin therapy. Type 2 diabetes, more prevalent among adults and often linked to obesity and lifestyle factors, can be managed through dietary modifications, exercise, medication, and insulin therapy as needed. Effective diabetes management entails vigilant monitoring of blood sugar levels, adherence to a strict diet, regular physical activity, and the timely administration of insulin or oral medications to prevent complications and maintain overall health. In the contemporary era, Diabetes mellitus has emerged as a rapidly escalating health concern, particularly affecting various demographic groups. Early and timely detection of diabetes is essential for initiating effective interventions and preventing complications from arising. Conventional diagnostic methods often fall short in accurately identifying individuals at risk of developing the disease, highlighting the necessity to explore innovative techniques. ML has gained traction as an approach that leverages data-driven predictive patterns to enhance diabetes diagnostics.

B. Motivation

The motivation driving the research is the pressing need to overcome obstacles that obstruct the diagnosis of diabetes. Despite advancements in health technology, the persistent issue of missed diagnoses and delayed detection continues to incur significant health and economic costs. Against this backdrop, our research embarked on a mission to analyze and compare various ML algorithms to identify the most effective approach for diabetes detection. We aimed to develop methods that not only provided higher accuracy but also ensured interventions were delivered without delay.

Our study sought to bridge the gap between technology and healthcare. By leveraging ML, our goal is to equip physicians with advanced tools that facilitate early disease identification. This entails harnessing data and uncovering correlations that might otherwise go unnoticed, leveraging the strengths of algorithms. Ultimately, our initiative aimed to provide impactful insights that empower healthcare providers to make informed decisions.

In short, our journey was one of empowerment–empowering clinicians to take proactive measures, empowering patients to pursue wellness with determination, and empowering humanity as a whole in the battle against diabetes.

C. Contributions

This report contributed to the healthcare analytics field with a comprehensive assessment of the performance of different ML algorithms in diagnosing diabetes. From a theoretical standpoint, the study not only explored the basics of the predictive algorithms that have been used but also validated their implementation using empirical data, thus capturing the performance of diverse models such as DT [1], SVM [2], NB [3], kNN [4], and RFC [5]. We brought together findings obtained by analyzing real-world data and provided practice-centered advice, which may additionally be used to create evidence-informed guidelines. Moreover, in addition to the current study, our research is a basement for yet-to-be-investigated areas in predictive analytics, as this will lead to the establishment of innovations and advancements in diabetes management strategies. This approach, therefore, ensures that the findings we get are both credible and sensible to make regarding the current developments in healthcare analytics. The proposed approach contributes to individualized blood sugar diagnosis and therapy by:

1) ***Enhanced Diagnostic Accuracy****:*

a) Machine Learning Algorithms: Using the NB, SVM, DT, kNN, and RFC ML algorithms, higher values are obtained in diagnosing diabetes accurately. Each algorithm analyzes patients data differently, comparing similarities and disparities that may signal diabetes.

b) Model Comparison: The comparison of these algorithms helps the research identify which of the algorithms should be considered the most effective for diabetes prediction. For instance, the Random Forest Classifier abbreviated as RFC is most preferred because its high accuracy and insensitivity to sample size are helpful in accurate diabetes diagnosis.

2) ***Early Detection and Prevention****:*

a) Early Signs Identification: It is recommended to diagnose it earlier to avoid or delay the worst consequences which can be a diabetic complication such as heart diseases, kidney disorders, and neuropathy.

b) Timely Intervention: Those patients who receive the diagnosis in the initial phase, can start their management immediately, thereby improving their long-term prognosis and minimizing additional expenses.

3) ***Personalized Medicine****:*

a) Tailored Treatment Plans: The use of Machine Learning allows certain data on the particular patient genetic profile, life habits, past health issues to form the individual approach to treatment. These are also preferred over other treatments as they are more personal and consider treatment factors of individual patients.

b) Risk Factor Analysis: It assists in identifying specific patterns and potential threats in the patients data and individualizing therapy. This means that two patients with the same disease may get different treatments due to their unlike profiles.

II. Literature review

Sisodia *et al.* [6] deployed three ML algorithms involving SVM, NB, and DTs to forecast the outbreak of diabetes. The authors achieved 76.30% accuracy by using the NB classifier with the Pima Indians Diabetes Dataset (PID)[7].

Rastogi *et al.* [8] utilized different ML algorithms like RFC, K- means clustering, Linear regression and Logistic Regression, DTs, SVM and NB for the forecasting of diabetes. The scholars found that logistic regression was a more reliable tool than the other ML techniques, with an accuracy of 82.46% with the diabetes dataset from Kaggle.

Daanouni *et al.* [9] implemented an ML algorithm on the dataset of the PID for the prediction of diabetes. This was achieved by experimenting with artificial neural networks, kNN, deep neural networks, and DTs on the data set. They concluded that within Deep Neural Networks, the feed-forward network was the most effective, achieving the highest accuracy rate of 90%.

Xue *et al.* [10] SVM, NB, and Light Gradient-Boosting Machine (LightGBM) ML algorithms. The authors highlighted the importance of considering various approaches when distinguishing successful prediction strategies. The Sylhet Diabetes Hospital in Bangladesh gave this dataset.

Dutta *et al.* [11] pointed out that the classifiers used include NB, RFC, Decision Tree (DT), Extreme Gradient Boosting (XGBoost), and LightGBM.The most definitive finding was achieved by employing logistic regression, which has a 96% record of accuracy acquired from a just-released DDC dataset in Bangladesh.

Zou *et al.* [12] believed that early detection of the disease could assist physicians in managing patient conditions and improving clinical decision-making. In this case, he provided diabetes prediction in the model through RF, DTs, and deep learning networks. The information furnished was gathered from the physical examination records of patients in a Chinese hospital called Luzhou. The authors studies in combination pointed to an accuracy rate of RFC standing at 80.84%.

Soni *et al.* [13] involved six ML models: kNN, SVM, DT, Logistic Regression, RFC, and Gradient Boosting, which were employed to forecast diabetes using the PID. The best-performing method was RFC, whose accuracy function was 77%.

Muhammad *et al.* [14] presented in the year 2020 explored classifiers such as RFC, gradient boosting, kNN, SVM, logistic regression, and NB. Based on the Mohamed Specialist Hospitals data from Nigeria, the writer coined this name. Unlike algorithms such as Gradient Boosting (82%), kNN(75%), NB (86%), Logistic Regression (90%), and RFC (90%), we concluded that RFC was still the one with the highest percentage of accuracy (87%).

III. Proposed Work

The prevalence of diabetes was the issue at hand, as evidenced by the loss of numerous lives, yet it remained a condition for which a complete cure had yet to be found. Sometimes diabetes is discovered too late after many years,

and the patient may live with its symptoms without even knowing about it. A lot of progress has been made in building the system to diagnose diabetes, and indeed, more efforts have become relevant with the ever-evolving Artificial Intelligence (AI). Specifically, our study focused on the implementation of five mainstream Machine-learning algorithms on three repositories of datasets and the inspection of accuracy performance indicators for all the ML algorithms on all three datasets.

A. Preprocessing

1) Data Collection: The primary and most important step in constructing any ML model is data collection. The accuracy score of the algorithms used are derived from the sample used. The data provided by the three datasets [15], [16], [17] on diabetes at Kaggle has been collected, and we trained our model on this data to predict whether someone has diabetes or not.

2) Data Standardization: From sklearn, we employed the StandardScaler object from the preprocessing module. For this, it cuts down on standardization by taking the mean out and dividing it through to get a unit variance as shown in Equation 1.

$$Standardization = \frac{y - x}{z} \tag{1}$$

Where y, x, and z denote the data point, the mean of the dataset, and the Standard deviation of the dataset.

A fit method of the sklearn library is used to obtain the mean and standard deviation of the data. The process then involves the use of the transformation method, with the mean and standard deviation from previously calculated values deducted from it.

3) Data Splitting: The 80-20 rule divided the general data into train and test sets. This means that 20% was used for testing, and the remaining 80% was used for training.

B. Models Used

1) Naive Bayes: In the present study, the Gaussian-Naive Bayes (GNB) variation of the NB classifier is utilized, as it fits the continuous data well and uses the idea that all features have a normal distribution. In light of the strict conditionality of independent features, the probabilistic classification classifier NB applies the Bayes theorem. The algorithm hypothesizes the presence of one feature of a certain group as unbiased and independent of the other features; that is why it is named NB. When we want an immediate, responsive action, this is when the algorithm tends to be used more often. The model's performance is considered by determining its accuracy. The mathematical formula for Bayes Theorem is shown in Equation (2)

$$P(M \mid N) = \frac{P(N \mid M) \cdot P(N)}{P(M)} \tag{2}$$

Where:
$P(M \mid N)$ denotes the posterior probability of class (M) given predictor (N).
$P(M)$ and $P(N)$ are the probabilities of class and predictor respectively.
$P(N \mid M)$ signifies the likelihood, which is the probability of the predictor given the class.

2) Support Vector Machine: They are employed for linear and non-linear classification, regression, and outlier detection. The primary goal of this approach is to create a hyperplane that separates an N-dimensional space into distinct classes. The algorithm aims to find a hyperplane that maximizes the separation margin between two classes. It selects a hyperplane with the maximum distance from the nearest data points on either side, known as the maximum margin hyperplane if it exists. SVMs are versatile and can perform various tasks such as image detection, spam detection, and face detection.

3) Decision Trees: This algorithm decides choices based on questions and a criteria framework. The approach is just like a flowchart, where a node means a feature, a branch means a decision and a leaf points to the result. Components of a DT:-

a) Root Node: This node extends from the top or starting point of the decision tree, where all the features are divided into different branches based on various conditions.

b) Decision Nodes: These nodes, which are also called decision nodes, are constructed by splitting the root node. They represent intermediate decisions.

c) Leaf Nodes: The last nodes of the tree from where even further splitting does not happen. Furthermore, they give concrete results.

The algorithm is composed of recursion in such a way that it rives the data set into minute subsets. Each node is evaluated, and the algorithm determines which feature to split the data. In our algorithm, the maximum depth is 2 means we will use a maximum of 2 levels in our tree. Moreover, this reduces the model's tendency to pick up only specific data patterns, known as overfitting, ensuring that the resultant tree is not overly complex. The criterion used is entropy and it is calculated as shown in Equation (3)

$$E(S) = -\sum_{i=1}^{n} p_{C_i} \log_2(p_{C_i}) \tag{3}$$

Where $E(S)$ represents the entropy of dataset S, p_{C_i} is the proportion of instances in S that belong to class C_i and n is the number of classes. Information gain is the reduction in entropy or surprise by splitting a feature and it is calculated as shown in Equation (4)

$$Gain(S, A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} E(S_v) \tag{4}$$

Where S_v denotes the subset of S for which attribute A has value v, $Values(A)$ is the set of all possible values for an attribute A, and S is the whole sample space.

4) k-Nearest Neighbours: The approach is fairly easy, straightforward, and accurate based on the samples that are as close or the closest one to the reference as possible, known as the neighbors, where the closeness is calculated using

Euclidean Distance, Manhattan Distance, etc. This specific model will assess the five nearest neighbors for its predictions (k = 5) and will use Euclidean distance as shown in Equation (5).

$$d(d,e) = \sqrt{(d_1 - e_1)^2 + (d_2 - e_2)^2 + \ldots + (d_n - e_n)^2} \quad (5)$$

Where d and e are the two points in the data collection. d_i and e_i are the i^{th} coordinates respectively, and n is the number of dimensions or features in the dataset. Since it makes no assumptions about the data, the algorithms are better able to adapt to various patterns and generate predictions based on the data's local structure. The choice of 'k' significantly influences the performance of the algorithm.

5) Random Forest Classifier: It is a powerful machine-learning algorithm that works by creating several DTs during the training phase. While predicting, the algorithm totals the results of all trees; for classification tasks, the preferred method of decision-making is through voting, while for regression tasks, the preferred method is through averaging. It predicts the output with high accuracy for large datasets. RFCs are widely used for classification, regression, and reduced overfitting. In our model, we configured a Random Forest Classifier (RFC) with the following parameters: estimators=100 to create a maximum of 100 trees, max_depth=100 to ensure that the maximum depth of any tree does not exceed 100, and min_samples_split=20 to specify that at least 20 samples are required to split an internal node. The min_samples_leaf is set to 1, indicating that a node must have at least one sample to be considered a leaf. The max_features is set to the square root of the total number of features to determine the number of features considered for the best split. Finally, the random_state is set to 5 to ensure reproducibility of the results.

IV. Experiment

A. Dataset

A large dataset is needed to undertake a comparative study of various features and models. Here, we have used three datasets from different sources and with different numbers of features in the data set. We sourced the data sets from Kaggle, making them more resilient and versatile.

1) Diabetes Dataset: [15] This Dataset was sourced from Kaggle and has 769 entries and 9 features namely Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, and Outcome. It also focuses on Pima Indian women, making it specific to this population group only.

2) Diabetes prediction dataset: [16] This dataset was sourced from Kaggle and has 100001 entries and 9 features namely Age, hypertension, Gender, heart_disease, BMI, smoking_history.Blood_glucose_level, HbA1c_level, and diabetes. It includes general health indicators without specifying a population, suggesting broader applicability.

3) Diabetes Health Indicators Dataset: [17] This Dataset was sourced from Kaggle and has 253681 entries and 22 features.It contains diverse demographic and health data, making it suitable for general applicability across various populations. This has been summarized in the table I.

Table I
DATASET

Name of dataset	Features	Target Class
Diabetes Dataset	9	2
Diabetes prediction dataset	9	2
Diabetes Health Indicators Dataset	22	2

B. Experiment setting

In this study, we have compared machine-learning algorithms on various datasets to predict diabetes.
The initial step involved the Standardisation and Label encoding of the datasets, this is achieved by using StandardScaler() and LabelEncoder() function. Following this, the datasets were split using an 80-20 split i.e. 80% of the data was used to train the model, and the rest was used to test the model to calculate the accuracy.
Subsequently, kNN, RFC, DT, NB, and SVM were applied to the datasets, and their performance was assessed using specificity, accuracy scores, sensitivity, and CM that are calculated.

Subsequently, the accuracy scores derived from these models are represented in a barplot, providing a visual interpretation of the model's performance. This graphical representation facilitates a more intuitive understanding of the distribution and variance of accuracy scores across different models.

V. Results and Discussion

A. Analysis of Algorithms on the Datasets

In this section, we compared the accuracy of each algorithm using five models: NB, SVM, DT, kNN, and RFC on three different datasets.

1) Diabetes Dataset: For this dataset, RFC was the best fit for predicting Diabetes with its accuracy score of 81.2%, specificity as 86%, and sensitivity as 71%, and the CM is shown in Table II.

Table II
CM FOR DIABETES DATASET

	Predicted Diabetes	Predicted Not Diabetes
Actually Diabetes	86	13
Actually Not Diabetes	16	39

2) Diabetes prediction dataset: For this dataset, DT was the best fit for predicting Diabetes with its accuracy score of 97.2%, specificity as 100%, and sensitivity as 67%, and the CM is shown in Table III.

Table III
CM FOR DIABETES PREDICTION DATASET

	Predicted Diabetes	Predicted Not Diabetes
Actually Diabetes	18299	0
Actually Not Diabetes	551	1150

3) Diabetes Health Indicators Dataset: For this dataset, DT and SVM were the best fits for predicting Diabetes with an accuracy score of 86.1%, specificity as 100%, and sensitivity as 0%, and their CM is shown in Table IV.

Table IV
CM FOR DIABETES HEALTH INDICATORS DATASET

	Predicted Diabetes	Predicted Not Diabetes
Actually Diabetes	43671	0
Actually Not Diabetes	7065	0

4) Analysis of implemented algorithms on the datasets: As observed the performance of five different ML models: kNN, SVM, NB, RFC, and DT on three different datasets (Diabetes Dataset, Diabetes prediction dataset, Diabetes Health Indicators Dataset.

kNN performed consistently across all datasets with the highest performance on the Diabetes prediction dataset.

SVMs also showed consistent performance across all datasets, slightly outperforming kNN on the Diabetes prediction dataset.

NB had the lowest performance on the Diabetes prediction dataset and Diabetes Health Indicators Dataset compared to the other models. However, its performance on the Diabetes Dataset was comparable to that of SVM.

RFC performed well on all datasets, with its best performance on the Diabetes prediction dataset.

DT had the highest performance on the Diabetes prediction dataset and Diabetes Health Indicators Dataset, but its accuracy dropped on the Diabetes Dataset.

These observations could help in choosing the right model based on the dataset at hand. However, it is important to consider other factors such as the nature of the data, the interpretability of the model, and the computational resources available. These results are summarised in Table V.

VI. GRAPHICAL COMPARISON OF CLASSIFICATION MODELS ON THE DATASETS

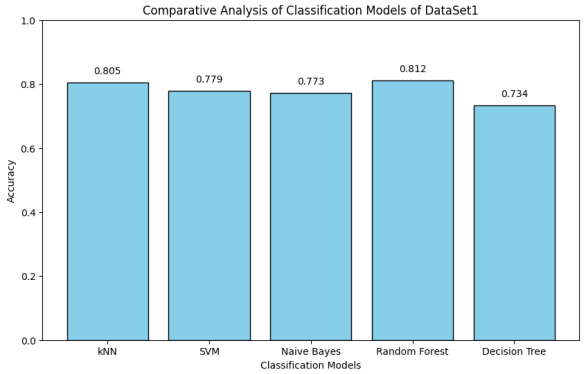

Figure 1. Analysis of all algorithms on Diabetes Dataset

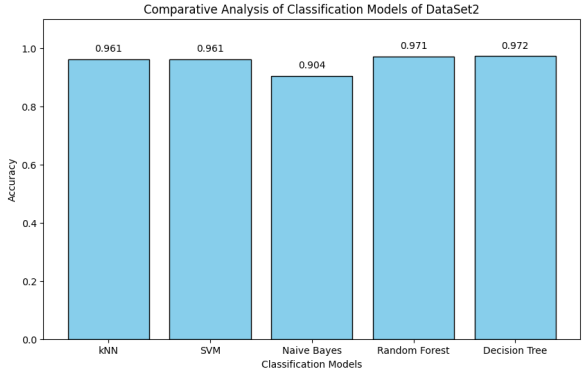

Figure 2. Analysis of all algorithms on Diabetes prediction dataset

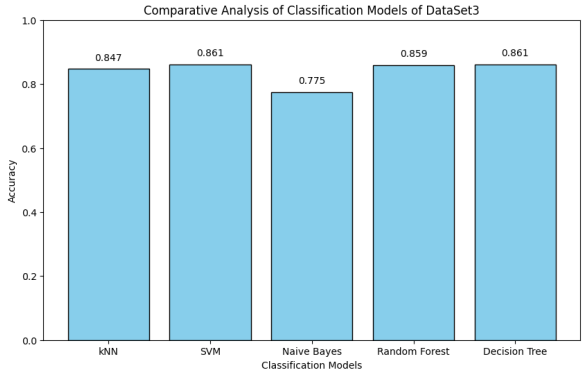

Figure 3. Analysis of all algorithms on Diabetes Health Indicators Dataset

VII. CONCLUSION AND FUTURE WORK

It has been identified that this study critically analyzed five ML techniques used in predicting diabetes: kNN, DT, RFC, SVM, and NB. Hence, we deduced the effectiveness as well as the performance of these techniques through their application of them to the diagnosis of diabetes. Lastly, our analysis concluded that from Table V the RFC was the most probable model for diabetes prediction since it showed better predictive accuracy and robustness compared to the other models. However, the outcomes of each model were different when the overall scores were measured which is shown in Figure 1, Figure 2, and Figure 3, which means that diabetic predictors should be chosen based on various parameters. In addition to this, this research also contributed to the whole knowledge base regarding diabetes prediction by showing how the ML approaches could be applied to improve and enhance early detection and management of this disease. Similarly, these models could assist healthcare workers with better choices and interventions to reduce the likelihood of diabetes complications.

Soon, the use of ML algorithms will enable us to envision an era of diabetes prediction. To boost forecast capability as well as resilience, one can start with the ensemble learning approach, which involves a manifold of models with the best features. Boosting model performances and generalizations

Table V
COMBINED TABLE FOR ALL DATASETS AND METRICS

Dataset	Algorithm	Accuracy	Specificity	Sensitivity
Diabetes Dataset	KNN	0.805195	0.909091	0.618182
	SVM	0.779221	0.898989	0.563636
	NB	0.772727	0.858586	0.618182
	RFC	0.811688	0.868687	0.709091
	DT	0.733766	0.777778	0.654545
Diabetes Prediction Dataset	KNN	0.96110	0.991802	0.630805
	SVM	0.96130	0.995409	0.594944
	NB	0.90380	0.927263	0.651381
	RFC	0.97055	0.996557	0.690770
	DT	0.97245	1.000000	0.676072
Diabetes Health Indicators Dataset	KNN	0.847406	0.951020	0.206794
	SVM	0.860750	1.000000	0.000000
	NB	0.774539	0.807561	0.570417
	RFC	0.859094	0.859094	0.169002
	DT	0.860750	0.970735	0.000000

through ensemble methods such as stacking, boosting, and bagging is very successful in several instances. By including and combining development-specific attributes and data sources like environmental factors, lifestyle habits, or genetic biomarkers, the pathophysiology of diabetes can be elucidated. ML algorithms can process complex relationships and patterns thanks to finding meaning and interconnecting different data sets. It enables advanced forecasting and tailored measures to be used.

REFERENCES

[1] S. B. Kotsiantis, "Decision trees: a recent overview," *Artificial Intelligence Review*, vol. 39, pp. 261–283, 2013.

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[3] I. Rish, "An empirical study of the nave bayes classifier," *IJCAI 2001 Work Empir Methods Artif Intell*, vol. 3, 01 2001.

[4] J. Sun, W. Du, and N. Shi, "A survey of knn algorithm," *Information Engineering and Applied Computing*, vol. 1, 05 2018.

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018.

[7] "Pima indians diabetes database," 2016, Pima Indians Diabetes Database.

[8] R. Rastogi and M. Bansal, "Diabetes prediction model using data mining techniques," *Measurement: Sensors*, vol. 25, p. 100605, 12 2022.

[9] O. Daanouni, B. Cherradi, and A. Tmiri, "Diabetes diseases prediction using supervised machine learning and neighbourhood components analysis," 03 2020, pp. 1–5.

[10] J. Xue, F. Min, and F. Ma, "Research on diabetes prediction method based on machine learning," *Journal of Physics: Conference Series*, vol. 1684, p. 012062, nov 2020.

[11] A. Dutta, M. K. Hasan, M. Ahmad, M. A. Awal, M. A. Islam, M. Masud, and H. Meshref, "Early prediction of diabetes using an ensemble of machine learning models," *International Journal of Environmental Research and Public Health*, vol. 19, no. 19, 2022.

[12] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in Genetics*, vol. 9, 2018.

[13] M. Soni and S. Varma, "Diabetes prediction using machine learning techniques," *International journal of engineering research and technology*, vol. 9, 2020.

[14] M. Jibril, E. Algehyne, . Sani, and S. Usman, "Predictive supervised machine learning models for diabetes mellitus," *SN Computer Science*, vol. 1, 07 2020.

[15] A. D. Khare, "Diabetes detection," 2022, Diabetes Detection.

[16] M. Mustafa, "Diabetes prediction dataset," 2023, Diabetes Prediction Dataset.

[17] A. Teboul, "Diabetes health indicators dataset," 2022, Diabetes Health Indicators Dataset.