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: 21052976@kiit.ac.in

Hardik Chauhan
School of Computer Engineering
KIIT Deemed to be University,
Bhubaneswar, India 751024
Email: 21052904@kiit.ac.in

Sneha Gupta
School of Computer Engineering
KIIT Deemed to be University,
Bhubaneswar, India 751024
Email: 21052923@kiit.ac.in

Shreya Mondal
School of Computer Engineering
KIIT Deemed to be University,
Bhubaneswar, India 751024
Email: 21052362@kiit.ac.in

Saumyadeep Mahanta
School of Computer Engineering
KIIT Deemed to be University,
Bhubaneswar, India 751024
Email: 21052918@kiit.ac.in

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

***Abstract*—Looking at healthcare by utilizing machine learning will be at the heart of our job, with the main aim of recognizing early diabetes cases. Giving an early diagnosis of diabetes to prevent the growth of serious complications is clearly the main task to achieve for a thorough treatment of the problem. Diabetes is a chronic disease that is a global problem because it is a permanent disease. Therefore, diabetes is the leading cause of early death. We used machine learning and statistical methodologies 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 also what their advantages and disadvantages are within the field of health care. Our goal is to give out the outcomes that model the implementations in a real-life healthcare system. Further highlighting this, it serves as the basis for making blood sugar diagnoses and therapy on an individual basis, which then results in the improvement of mental health. In completing these steps, therefore, personalized medicine may well provide a solution for diabetes treatment in the future.**

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

I. Introduction

A. Background

It is present that in the modern era, diabetes mellitus has become a rapidly rising kind of health problem that is mainly growing among the demographic communities. In order to allow for the proper intervention and to keep complications at bay, early and timely detection of diabetes is of great significance. Conventional diagnostic methods can be significantly imprecise in the assessment of individuals who could perhaps develop a disease, to the extent that new techniques should be found. Machine learning (ML) is currently gaining in popularity as an ML approach that enables data-driven predictive patterns to generate diabetes diagnostics. Based on the latest figures from the WHO, it is estimated that more than 422 million people in the world live with diabetes. In fact, experts predict that the numbers will rise even more. One of the most essential tools for diabetes prevention and control is early diagnosis. This makes it possible to lower the effects of diabetes on both people and their healthcare systems. Nevertheless, the case is that some techniques of conventional diagnostics may not be precise enough to enable us to have a word about people who are already at risk. Machine learning provides a solution that enables, through an advanced process of data mining, the reduction of diagnostic time and the delivery of quality medical services. Algorithms of machine learning take into account a broad variety of factors, which include medical history, genetic propensity, style of life, and lab work, to determine who is likely to suffer from diabetes. With these algorithms at hand, they can detect multitudes of patterns and interconnections among huge resources of data, based on which they predict who may need proactive screening and intervention. What is more, machine learning empowers personalized medicine by providing doctors with patient-specific prognoses and a decision-making process that fits the specifics of the patients. Even though it may be seen as having multiple positive sides to it, machine learning, when it comes to the automated diagnosis of diabetes too, has its own set of challenges that come with it. Gaps in datasets, linkage of features, and interpretability of outcomes are among the main areas where the ML model should be improved in building robust models for diabetes detection. Researchers are dedicated to investigating different machine learning approaches, such as deep learning, which could be useful for successful diabetes detection problem-solving and the enhancement of efficiency and accuracy. Hence, it is time to move forward with a comparative study of machine learning models for the detection of diabetes in light of the above facts. This kind of study has been aimed at assessing the accuracy rates of different ML algorithms in effectively identifying diabetes-at-risk individuals. Sensitivity, specificity, and predictive value are the current evaluation criteria. This study endeavors to gain a comprehensive understanding of the advantages and drawbacks of the existing ML methods that aim at the detection of diabetes disease, on the grounds of which the latest techniques in disease detection are to be developed or pre-existing approaches are to be modified for the betterment of patient outcomes.

B. Motivation

Undertaking this research is motivated by the urgent need to overcome the barriers preventing diabetes diagnosis. Progress in health technology notwithstanding, the perennial concern of missed diagnosis and delayed detection continues to cost in terms of health effects and healthcare costs above the expected. Against this canvas, our research journey starts a mission to analyze and compare a variety of machine learning algorithms with the aim of discovering which one is the most successful in the search for diabetes. We aim to create a way to execute treatments that not only give higher precision but also are given at the proper time. In essence, our study looks to fill the gap between technology and health. Through the use of machine learning technology, our mission is to ensure that physicians are well-equipped with cutting-edge tools that make early disease identification and personalized care possible. This involves information hoarding and seeking correlations that might otherwise be overlooked, which is the strength of algorithms. In this regard, our initiative aims at delivering impactful knowledge, which will help the providers make appropriate choices. All in all, our journey is one of empowerment – empowerment for clinicians to take the initiative, empowerment for patients to explore the path to wellness with vigor, and empowerment for humanity as a whole in its battle for victory over diabetes.

C. Contributions

This signature report makes its contribution to the healthcare analytics realm by providing a comprehensive evaluation of machine learning models for diabetes diagnosis. From a theoretical standpoint, the study not only explores the basics of the predictive algorithms that have been used but also validates their implementation using empirical data, thus capturing the performance of diverse models such as Naive Bayes, Support Vector Machines, Decision Trees, K-Nearest Neighbours, and Random Forest. We bring together findings obtained by analyzing real-world data and provide 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. Along with the concreteness of certain types of our machine learning algorithms, it is important to highlight that they are both fully applicable and effective in the medical domain. This approach, therefore, ensures that the findings we get are both credible and sensible to make regarding the current developments in healthcare analytics.

II. Literature review

Sisodia *et al.* [1] used three machine learning methods to predict diabetes. The authors have reported 76.30% correctness using the Naive Bayes classifier with the Pima Indians Diabetes data set.

Rastogi *et al.* [2] used four different machine-learning methods in order to predict diabetes. The use of the logistic regression technique was found to be more precise than other Machine Learning techniques, at 82.46% with the Kaggle dataset, according to the scholars' findings.

Daanouni *et al.* [3] applied Machine learning to the Pima Indian Dataset for the prediction of diabetes. This involved testing various types of neural networks, such as Artificial Neural Network, k-Nearest Neighbours, Deep Neural Network, and decision trees, on the data set. Deep Neural Network is found to be the best algorithm, with an accuracy rate of 90%.

Xue *et al.* [4] use Support Vector Machine, Naive Bayes, and LightGBM machine learning Algorithms. The most significant element that must be considered by researchers in order to identify successful prediction strategies is multiple approaches for researching or contrasting each other's work according to authors. Bangladesh's Sylhet Diabetes Hospital provided this dataset, and it was also revealed that this method is better than other methods.

Dutta *et al.* [5] contend that the classifiers used include Naive Bayes, Random Forest, Decision Tree, XGBoost, and LightGBM. The most accurate way to present their findings is via Logistic Regression, which has an accuracy rate of 96% and was obtained from a recently classified dataset from Bangladesh.

Zou *et al.* [6] aimed at early disease detection and aiding this process. For instance, he predicted diabetes using random forest (RF), decision trees, and neural networks. The data set was collected from a Chinese hospital's Luzhou physical examination records. When combined, the results of the authors' studies show that the Random Forest accuracy rate is 80.84

Soni *et al.* [7] research involved six machine learning models: KNN, SVM, Decision Tree, Logistic Regression, Random Forest, and Gradient Boosting, for predicting diabetes in the Pima Indian Diabetes dataset. The best performance has been shown by Random Forest, with a score of 77% when testing the Machine Learning techniques.

Muhammad *et al.* [8], published in 2020, uses Random Forest, Gradient Boosting, KNN, SVM Logistic Regression, and Naive Bayes. The author names it this way according to Nigeria's Mohammed Specialist Hospital dataset. In comparison with such algorithms as Gradient Boosting (82%), KNN (75%), Naïve Bayes (86%), Logistic Regression (90%), and Random Forest (90%), we find out that random forest still maintains the highest accuracy (87%).

K. VijiyaKumar *et al.* [9] applied ML models to predict diabetes. The Random Forest ML model was implemented and its performance was tested for diabetes. The accuracy obtained is more than 90% (approximately). This is greater when compared to other machine learning algorithms for diabetes prediction.

L. V. R. Kumari *et al.* [10] apply several ML algorithms for diabetes prediction. The paper summarises the prediction accuracy of ML algorithms for diabetes. This work aims to make an early prediction of diabetes more precisely by using a variety of machine-learning algorithms.

P. Cıhan *et al.* [11] apply many ML models for diabetes prediction. The performance accuracy of different algorithms is found to be different. The performances of machine learning methods were calculated using precision, recall, ROC curve, and PRC criteria metrics. The Logistic Regression method is more successful than other methods in predicting diabetes disease accurately.

N. Fazakis *et al.* [12] applied the concept of ML for prediction of diabetes. The major contribution was to predict type-2 diabetes using ML. predict the type 2 diabetes using ML models. The paper successfully predicts type-2 diabetes using ML.

M. A. R. Refat *et al.* [13] applied deep learning models for the prediction of the same diseases. Deep learning applies the concept of Neural Network and searching deeper than the ML models for making predictions. It was found the XGBoost classifier performed best from the rest of the algorithms by approximately 100.0%

III. Proposed Work

Diabetes has been a problem that has caused enough casualties and still is a disease with no proper cure. In many cases, diabetes often goes undetected and people unknowingly suffer from it. Much work has been done to develop systems to detect diabetes and with the introduction of artificial intelligence, much progress has been made in this direction. In our study, we aim to implement five fundamental machine learning algorithms in three distinct datasets and compare the accuracy of all the algorithms across all three datasets.

A. Preprocessing

1) Data Collection: Data collection is the first and the most important step towards the construction of any machine learning model. The Accuracy scores of the algorithms used depend on the dataset used. We have collected three datasets on diabetes from Kaggle and our model is trained on these datasets to predict whether a person has diabetes or not.

2) Data Standardization: We have used the StandardScaler object from sklearn.preprocessing module. It helps in standardization by removing the mean and scaling to unit variance.

$$z = \frac{x - \mu}{\sigma}$$

Where x represents the data point, μ represents the mean of the dataset, and σ represents the standard deviation of the dataset.

The fit method calculates the mean and the standard deviation of the data. The transform method then uses the previously calculated values by subtracting the mean and dividing by the standard deviation to arrive at the final values.

3) Data Splitting: The standardized data has been split into train and test data using an 80-20 split i.e. 20% of the data is used for testing and 80% is used for training respectively.

B. Models Used

1) Naive Bayes: The Naive Bayes classifier in this study is the Gaussian Naive Bayes variant, which works well with continuous data and is based on the idea that the features have a normal distribution. Under strict independence requirements between features, the probabilistic classifier Naive Bayes applies the Bayes theorem. The algorithm assumes the occurrence of a certain feature is independent of the other features hence the name Naive Bayes. It is a simple tool to make quick predictions. It is more used in scenarios that require instantaneous responses. The performance of the model is evaluated by computing the accuracy and creating a confusion matrix. The mathematical formula for Bayes' Theorem is:

$$P(A \mid B) \ = \frac{P(B \mid A) \cdot P(B)}{P(A)}$$

Where:
$P(A \mid B)$ is the posterior probability of class (A) given predictor (B).
$P(A)$ and $P(B)$ are the probabilities of class and predictor respectively.
$P(B \mid A)$ is the likelihood which is the probability of the predictor given class.

2) Support Vector Machine (SVM): It can be used for linear as well as non-linear classification, regression, and even for outlier detection. Its main objective is to find an appropriate hyperplane that divides the data space into different classes in an N-dimensional space. The algorithm aims to find the hyperplane that represents the largest separation margin between two classes. It chooses that hyperplane whose distance from it to the nearest data point on each side is maximum. If such a hyperplane is present then it is known as the maximum-margin hyperplane. SVMs can be used to perform a variety of tasks from image detection to spam detection, face detection, etc.

3) Decision Trees: This algorithm renders choices based on an arrangement of questions and criteria. This strategy is comparable to a flowchart, where a node represents a feature, a branch represents a choice rule, and a leaf speaks to an outcome.
Components of a Decision Tree:-

a) Root Node: It is the first or the initial node of the decision tree from where all the features start dividing based on different conditions.

b) Decision Nodes: The nodes made by splitting the root node are called decision nodes. They represent intermediate decisions.

c) Leaf Nodes: The last nodes of the tree from where further splitting is not possible. They also represent the outcome.

The algorithm is based on recursion where it keeps dividing the data into smaller subsets. At each node, the algorithm chooses the best feature to split the data.

In our algorithm, we have kept the maximum depth of 2 which means that our decision tree will have at most 2 levels. This also helps to prevent overfitting make not allowing the tree to become too complex. The criterion used is 'entropy' and it is calculated as

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

Where $E(S)$ is 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. It is used for the information gain. Information gain is the reduction in entropy or surprise by splitting a feature and it is calculated as

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

where S is the total sample space, A is the attribute, $Values(A)$ represent the set of all possible values of attribute A, and S_v represents the subset of S for which attribute A has value v.

4) k-Nearest Neighbours (kNN): The approach is simple, straightforward, and effective, forecasting outcomes based on the closest or the nearest similar samples, known as 'neighbors', the closeness is calculated by using Euclidean Distance, Manhattan Distance, etc. Our specific model is set to consider the five nearest neighbors for its predictions(k=5) and it uses Euclidean Distance which is

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2}$$

Where p and q are two points in the dataset. p_i and q_i are the i^{th} coordinates respectively, and n is the number of dimensions or features in the dataset.

It does not require any assumptions about the data. This helps the algorithms to adapt to different patterns and make predictions based on the local structure of data. The choice of 'k' significantly influences the performance of the algorithm. A small value of k can make the model prone to overfitting. On the other hand, a large value of k can increase the bias resulting in high test error. Moreover, the value of k should also be an odd number as it ensures that there will always be a majority class hence preventing a tie.

5) Random Forest: It is a powerful machine-learning algorithm that works by creating several decision trees during the training phase. While predicting, the algorithm totals the results of all trees, either by voting (for classification tasks) or by averaging (for regression tasks), this takes less training time when compared to other algorithms. It predicts the output with high accuracy even for large datasets. Random Forests are widely used for classification, regression, and reduced overfitting. In our model we designed a Random Forest classifier with estimators=100 to create a maximum of 100 trees, max_depth=100 which makes sure that the maximum depth of any tree generated is not more than 100, min_samples_split=20 which specifies the minimum number of samples required to split an internal node which here is set to 20, the min_samples_leaf which defines the minimum number of samples required to be a leaf node is set to 1, max_features is set to square root which is the number of features required to considering when looking for the best split and random state =5 which is used to make the output deterministic.

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 didn't create the dataset we used for our research; we have sourced the data sets from Kaggle, making them more resilient and versatile.

1) Dataset_1 (Diabetes Dataset): This Dataset was sourced from Kaggle and has 769 entries and 9 features namely Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, and Outcome.

2) Dataset_2 (Diabetes Prediction Dataset): This Dataset was sourced from Kaggle and has 100001 entries and 9 features namely Gender, Age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level, and diabetes.

3) Dataset_3 (Diabetes Health Prediction Dataset): This Dataset was sourced from Kaggle and has 253681 entries and 22 features namely Diabetes_012, HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education and Income.

B. Experiment setting

In this study, we are performing a comparative analysis of existing machine-learning algorithms on different datasets for predicting diabetes.

The initial step involves the Standardisation and Label encoding of the datasets, this is achieved using StandardScaler() and LabelEncoder() function. Following this, the datasets are split using an 80-20 split i.e. 80% of the data is used to train the model, and the rest is used to test the model to calculate the accuracy.

Subsequently, the existing machine-learning models are applied to the datasets, and their performance is evaluated using the accuracy score and confusion matrix that are formulated using

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

and

	Predicted	**Predicted**
Actual	True Positive (TP)	False Negative (FN)
Actual	False Positive (FP)	True Negative (TN)

Where
True Positives (TP): These are cases in which we predicted yes (the event happened), and it did happen.
True Negatives (TN): We predicted no, and no event occurred.
False Positives (FP): We predicted yes, but the event didn't happen. Also known as "Type I error".
False Negatives (FN): We predicted no, but the event did happen. Also known as "Type II error".
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 conducted a comparative analysis of the accuracy of each algorithm using five models: Naive Bayes, Support Vector Machine, Decision Tree, k-Nearest Neighbours, and Random Forest on three different data sets.

1) Dataset_1(Diabetes Dataset): For this dataset, Random Forest Classifier was the best fit for predicting Diabetes with its accuracy score of 81.2% and confusion matrix as

86	13
16	39

2) Dataset_2 (Diabetes Prediction Dataset): For this dataset, Decision Tree was the best fit for predicting Diabetes with its accuracy score of 97.2% and confusion matrix as

18299	0
551	1150

3) Dataset_3 (Diabetes Health Prediction Dataset): For this dataset, Decision Tree and Support Vector Machine are the best fits for predicting Diabetes with their accuracy score of 86.1% and their confusion matrix as

43671	0
7065	0

4) Analysis of all algorithms on the datasets:

	KNN	**SVM**	**NB**	**RFC**	**DC**
Dataset_1	0.805195	0.779221	0.772727	0.811688	0.733766
Dataset_2	0.96110	0.96130	0.90380	0.97055	0.97245
Dataset_3	0.847406	0.860750	0.774539	0.859094	0.860750

We can observe the performance of five different machine learning models (k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest Classifier (RFC), and Decision Tree (DC)) on three different datasets (Dataset_1, Dataset_2, Dataset_3). kNN performs consistently across all datasets with the highest performance on Dataset_2. SVM also shows consistent performance across all datasets, slightly outperforming kNN on Dataset_2. Naive Bayes (NB) has the lowest performance on Dataset_2 and Dataset_3 compared to the other models. However, its performance on Dataset_1 is comparable to that of SVM. Random Forest Classifier (RFC) performs well on all datasets, with its best performance on Dataset_2. Decision Tree (DC) has the highest performance on Dataset_2 and Dataset_3, but its performance drops on Dataset_1. These observations can help in choosing the right model based on the dataset at hand. However, it's important to consider other factors such as the nature of the data, the interpretability of the model, and the computational resources available.

VI. Graphical Comparison of Classification Models on the Datasets

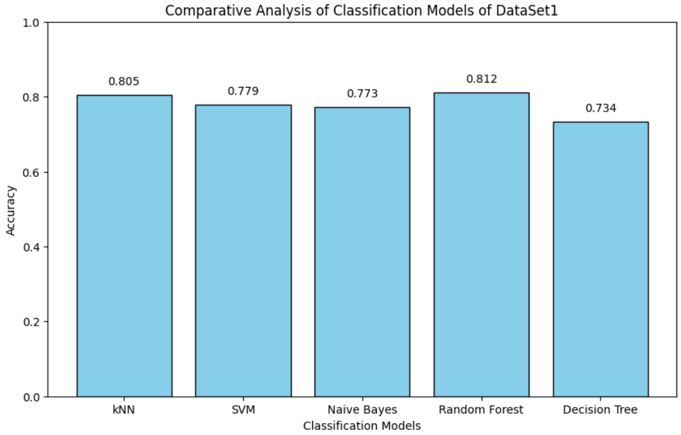

Figure 1. Analysis of all algorithms on Dataset_1

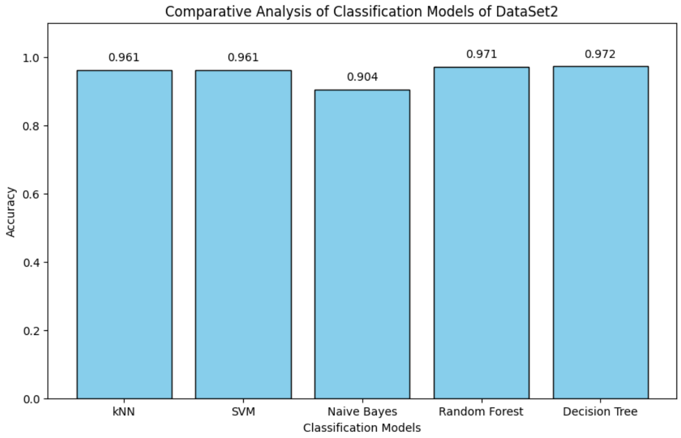

Figure 2. Analysis of all algorithms on Dataset_2

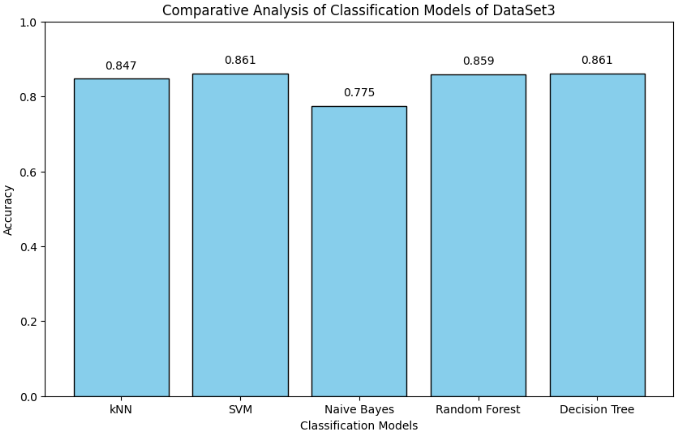

Figure 3. Analysis of all algorithms on Dataset_3

VII. Conclusion and Future Work

It has been identified that this study critically analyzes five machine learning techniques used in predicting diabetes: K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Machine (SVM), and Naive Bayes. Hence, we can deduce the effectiveness as well as the performance of these techniques through the application of them to the diagnosis of diabetes. Lastly, our analysis brings out the Random Forest as the most probable model for diabetes prediction since it displayed superior predictive accuracy and robustness than the other models. However, the outcomes of each model were different when the overall scores were measured, which means that diabetic predictors should be chosen based on various parameters.

In addition to this, this research also contributes to the whole knowledge base regarding diabetes prediction by showing how the Machine Learning approaches could be applied to improve early detection and management of this disease. Similarly, these models will assist healthcare workers with better choices and interventions to reduce the likelihood of diabetes complications.

In the near future, the use of machine learning 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 through ensemble methods such as stacking, boosting, and bagging has been shown to be 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. Machine learning 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] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018.

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

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

[4] 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.

[5] 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.

[6] 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.

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

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

[9] K. VijiyaKumar, B. Lavanya, I. Nirmala, and S. S. Caroline, "Random forest algorithm for the prediction of diabetes," in *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2019, pp. 1–5.

[10] L. R. Kumari, P. Shreya, M. Begum, T. P. Krishna, and M. Prathibha, "Machine learning based diabetes detection," in *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, 2021, pp. 1–5.

[11] P. Cıhan and H. Coşkun, "Performance comparison of machine learning models for diabetes prediction," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*, 2021, pp. 1–4.

[12] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, "Machine learning tools for long-term type 2 diabetes risk prediction," *IEEE Access*, vol. 9, pp. 103 737–103 757, 2021.

[13] M. A. R. Refat, M. A. 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)*, 2021, pp. 654–659.