Diabetes Prediction Using Machine Learning

Md Minhajul Abedin ID: 202483033 Email: mminhajula@mun.ca
Md Jawad Khan ID: 202381977 Email: mjawadk@mun.ca

Abstract—Diabetes is a chronic disease affecting millions globally, necessitating early detection to mitigate severe complications. This project develops a machine learning model to predict diabetes risk using the PIMA Indians Diabetes Dataset, which comprises 768 records of female patients with medical attributes such as glucose levels, BMI, and age. We implemented and evaluated multiple algorithms—Logistic Regression, Random Forest, Support Vector Machine (SVM), Gradient Boosting, and ANN—assessing their performance through metrics including accuracy, precision, recall, F1-score, and ROC-AUC. In the stage of data preprocessing, we replace missing values with the replacement level of the median and the standardized characteristic. We will see that the Random Forest model reaches a peak 87% accuracy, one outperformed in this context by ensemble methods. It offers a scalable and cost-effective means of identifying diabetes at an early stage that has the potential to bring about further improvements in healthcare.

Keywords: Diabetes Prediction, Machine Learning, Healthcare, PIMA Indians Diabetes Dataset, Random Forest, Logistic Regression, SVM, Gradient Boosting, and ANN.

1 Introduction

Diabetes is one of the most prevalent chronic diseases globally, with millions of individuals at risk of severe health complications such as heart disease, kidney failure, and blindness. According to the World Health Organization (WHO)[2], there are estimated 422 million diabetes sufferers across the globe, and the growth trend in a few years is that this number will only continue to rise, particularly in low- and middle-income countries. The increase in diabetes cases is mainly due to changes in lifestyle: declining diet quality, less exercise, and more urbanization. Left undiagnosed and untreated, the disease can have life-threatening complications, ultimately lower quality of life and increase medical costs.

Traditionally, diabetes detection is based on clinical tests that rely for their results on fasting blood glucose levels and oral glucose tolerance tests. These tests are often very expensive, time-consuming, and not readily accessible-particularly so in areas with limited resources. For many, the high cost makes such services out of reach, and once you miss your chance to find out about diabetes, it becomes more difficult for doctors to prevent complications. Early detection through changes in lifestyle, medication, and regular monitoring is crucial. However, in resource-poor regions, these methods necessarily depend on having specialized equipment and healthcare professionals, which may not be available at all.

Machine learning (ML) provides an innovative solution to the early discovery of diabetes. Through medical data analysis such as glucose levels, BMI and age, ML model can quickly and accurately determine the probability of a patient suffering diabetes. Using the PIMA Indians Diabetes Dataset, this project will develop and evaluate several different ML models to forecast diabetes risk in the future. The most accurate model for large samples is to be identified. With improved accessibility, ML models could turn health care on its head by enabling timely interventions and easing the pressure on medical services.

2 Related Work

As a chronic metabolic disease, diabetes has had a major impact on public health and the healthcare system worldwide. Making an early and accurate prediction can lead to timely intervention, with better outcomes for patients and less expenditure on medical resources. In recent years, machine learning has become an important tool for predicting diabetes by exploiting computational algorithms to analyze clinical and demographic data.

M. Alehegn et al. [3] focused on utilizing machine learning techniques for early diabetes detection, increasing the accuracy of prediction. The authors used four well-known algorithms - K-Nearest Neighbors (KNN), Naïve Bayes, Random Forest, and J48 - to create an ensemble model, in an endeavor to overcome this limitation of individual classifier performance from which one can expect only so much change without considerable increases in risk. The findings show that if these classifiers are combined, it produces a more accurate model with better stability than simply using a single algorithm does. However, the study also notes that data prepossessing (such as noise removal and handling missing values) is essential for improving model accuracy. The study's methods align with previous research on machine learning algorithms applied to health data, such as Random Forest and Naïve Bayes; however, there is no evidence to suggest that combining models from a series of different classifiers effectively enhances prediction results. Also, the paper said that variable selection, such as glucose, blood pressure, and BMI, greatly influences the predictability of dependent variables. The study adds to the literature by demonstrating how ensemble models can enhance early diagnosis and

highlights key areas for methodological advancement in diabetes prediction research, ultimately aiming to improve health outcomes through timely intervention.

By **Md. Maniruzzaman et al.** [4] aims to develop an effective machine learning system for predicting diabetes in individuals. The study is a diabetes dataset derived from the National Health and Nutrition Examination Survey (NHANES) that uses logistic regression for feature selection and compares its performance against four classifiers: Naïve Bayes, Decision Tree, Adaboost, and Random Forest. The methodology involved applying three cross-validation protocols (K2, K5, K10) and evaluating classifiers based on accuracy, sensitivity, and area under the curve (AUC). The key finding here is that the RF classifier, combined with feature selection of LR which is significantly more efficient than any of these other three. For K10 protocols, it gave an accuracy rate 94.25% and AUC came to 0.95. Strengths of this study are a large field study, with clear comparative analysis on many classifiers and robust evaluation performance indexes for the ML model. However, the paper's limitations include its reliance on a single dataset and possible confines of generalizability, and a lack of exploration into other advanced ML techniques like deep learning.

B. Sridhara Murthy et al. [5] mainly evaluates Logistic Regression and Decision Tree Classifier (DTC) for diabetes prediction based on Pima Indian dataset (768 records, 35% of them diabetic). With an 80:20 train-test split, it was found during testing that, compared to DTC which had a maximum accuracy of 78.57%, LR was superior at 82.46%. It had better recall (0.659 vs 0.61) permissive parameters (0.849 to 0.844) but suffered a worse precision at 0.76 and recall rates (specificity) were still lower than these of the DTC. The strengths of the development include a clear comparison of different algorithms and the use of multiple metrics (e.g., F1 score, specificity). However, limitations arise from the small size of these datasets and class imbalance; the exclusion of advanced models (such as Random Forest). In addition, generalizability is weak due to the lack of cross-validation or discussion on handle missing data in the Pima dataset – a problem which commonly occurs when working with humans. While the research showcases LR as of great predictive value for diabetes, future research needs to rigorously address data quality issues and use diverse algorithms so that results can be applied to actual clinical practice in different medical settings.

A. Mujumdara et al [6] proposed a predictive model built on a custom dataset (800 records, 10 variables, including socio-demographic factors such as occupation type) made in order to fill gaps in more traditional pieces of health care information like Pima. After completing preprocessing (missing value imputation, normalization) and using K-means clustering to identify possible relationships (e.g., glucose levels vs. age); it tried a total of 13 ML algorithms, among which Logistic Regression achieved 96% accuracy, a result greatly surpassing previous benchmarks (e.g., 76% with Pima). An additional gain in efficiency from taking a pipeline approach left their AdaBoost classifier with a margin of giving 98.8% accuracy, verifying the effectiveness of feature engineering and ensemble learning methods can have. A parade of strong points include methods to enrich lifestyle attribute sets, a thorough comparison between different algorithm families, and careful preparation for scale though shortcomings like sparse data size limit generalization, inability to touch on class unbalance mean that the study thus indicates the prospect that several complex pipeline assemblages and overall data inclusion hint for clinical predictions can be pursued, but should be studied on bigger samples as well as having more different areas.

I. Tasin et al. [7] proposed a diabetes prediction system using a merged dataset of 877 samples. The resulting whole-volume database combines the Pima Indian diabetic dataset and a private Bangladeshi database (RTML) taken from 203 female textile workers. The RTML dataset's missing insulin values were imputed through XGBoost regression. Class imbalance was treated by both SMOTE and ADASYN. Mutual information shows that glucose, BMI, age, and insulin are important predictors. With ADASYN, the XGBoost classifier attains an 81% accuracy rate and 0.81 F1-score, outperforming other models such as Random Forest (76%) and SVM (78%). Results from explainable AI tools (SHAP, LIME) show that glucose levels and number of pregnancies are the main decision drivers. The model was rolled out as a web app and an Android application for real-time predictions. Pros include the innovative integration of datasets, as well as interpretability, while cons cover small sample size and reliance on imputed insulin values. The work highlights the potential for using advanced ML techniques to combine domain-specific data for clinical applications.

In the reviewed studies, the potential of machine learning to improve the accuracy of diabetes prediction was proved. Ensemble models and advanced preprocessing techniques often delivered better results. The primary conclusions of the study were the effectiveness of Random Forest and XGBoost classifiers, the survival importance of feature processing (e.g. glucose levels, BMI, and age) and how understandable AI assisted in making model more understandable. However, such problems as dataset restriction, class imbalance and ethical issues regarding bias and generalization continue constantly face scientists anew.

3 METHODS

3.1 Data Collection

We collect the PIMA Indians Diabetes Dataset[1], from the UCI Machine Learning Repository. It contains 768 records of female patients, and each record includes eight features: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age, with a binary Outcome variable (0 for non-diabetic, 1 for diabetic).

3.2 Data Preprocessing

There are some missing or zero values in certain columns in our dataset. It may cause bias or mistakes in the model, so

it is appropriate to handle them. Additionally, we imputed the missing values by calculating either the median or mean of the respective columns, depending on the distribution of the feature. We also applied feature scaling to standardize the dataset so that the values of each feature are on a similar scale, improving model convergence, particularly for distance-based models and algorithms like SVM and logistic regression. However, as our categorical feature "Outcome" already is in numerical form we don't need any encoding. Therefore, if other categorical features were present, we would use One-Hot Encoding or Label Encoding to convert them into numerical format.

3.3 Machine Learning Models

In this project, we applied several machine learning algorithms to the risk of diabetes among the PIMA Indians based on Predictor (Diabetes Database). We selected these algorithms because they could deal–not only with linear relationships in dataset and nonlinear ones as well–with easing most types of relevant data, but a broad review is also available for evaluating model performance. The selected models are Logistic Regression, Random Forests, Support Vector Machines (SVM), Gradient Boosting, and Artificial Neural Networks (ANN). Each algorithm has its unique characteristics, and they were implemented and evaluated to identify the most effective model for diabetes prediction.

- **Logistic Regression**: Logistic regression is one of the simplest machine learning algorithms, widely used for binary classification tasks. The predictive values can be any number between 0 and 1. To classify them into two classes, you use a cut–off point of.5 on the y-axis in this case, diabetic and non-diabetic patients. The logistic function maps predicted values to probabilities, which are then classified into the target classes. But while it works well for data that is linearly separable, it can make mistakes with complex datasets where the relationship between features and the target variable are not linear. For this project, Logistic Regression is the benchmark model, as it allows us to compare other methods with this.

- **Random Forest**: Random Forest is a kind of ensemble learning approach, which makes predictions more accurate by bringing several decision trees together. Each tree in a random forest is trained on a different subset of data taken at random. These trees provide answers to questions about the input space which may be irrelevant in themselves, but form part of the evidence needed for reaching an overall conclusion. The eventual prediction is made by taking a vote among all trees present in the forest. Due to its ensemble design, Random Forest ensures that the convergence of many decision trees toward a stable prediction can be achieved to a large extent. Additionally, by capping tree depth at $\log_2(N)$, Random Forest can help mitigate overfitting, especially in large sample datasets. This model is particularly effective for handling structured (tabular) data and can capture both linear and non–linear relationships among features.

- **SVM**: SVM (Support Vector Machine) is a type of supervised learning algorithm applied to classification tasks and is particularly useful when the data cannot be divided linearly. If the sample points of one class and those of another are distributed in the same space but in different parts, then SVM tries to find a plane that can separate them to gain other class out. This is how this method makes margin big and its generalization ability stronger. By using convolutional kernels, SVM can also be used in non-linear problems. These kernels actually map the data into a higher dimensional space where there may well exist a linear separator. Although sometimes computationally expensive, SVM often performs well, especially for small datasets with complex boundaries.

- **Gradient Boosting**: Gradient Boosting is the most advanced example learning method, it constructs models sequentially; each new model tries to correct the previous ones' errors. It makes use of weak learners (usually model-chosen decision trees) and fits them step by step, adjusting the weights of the errors. Among the key advantages of Gradient Boosting is its ability to generate very high-accuracy models. It might work especially well for both classification and regression tasks and has been shown to still perform well on noisy datasets. In order to optimize Gradient Boosting models, hyperparameter tuning such as adjusting the learning rate and tree depth play a crucial role.

- **ANN**: A class of machine learning models inspired by the structure and functioning of the human brain are ANNs, or Artificial Neural Networks. Composed of layers of interconnected neurons, every neuron in the network processes information and sends it to the next node. Every time you change one of the weights on any neuron in between layers, you'll screw up something about learning. The network is trained using backpropagation to minimize the error between the predicted and actual values. ANNs are very effective in capturing complex, non-linear relationships between features and target variables. In the context of predicting diabetes, ANNs can model subtle patterns in medical data that traditional statistical methods are unable to detect, possibly offering better prediction accuracy then those methods. However, they require more computational resources and careful parameter tuning on such factors as the number of layers in a network, concepts to be decoded or calculated along with their associated neurons, and learning rates.

We trained these models on the PIMA Indians Diabetes Dataset and examined them in terms of standard performance metrics (accuracy, precision, recall, F1-score, and

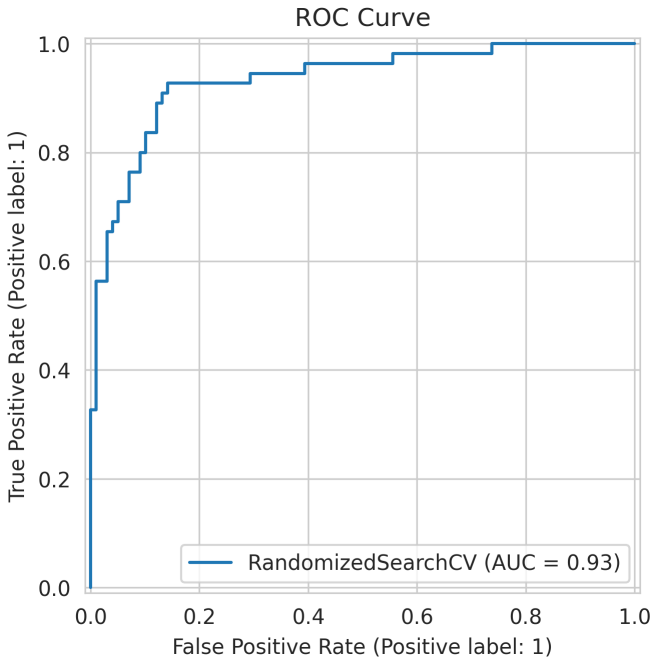

Fig. 1: ROC Curve and AUC score for Random Forest model

ROC–AUC). What we wanted to do was determine which algorithm best predicted the risk of developing diabetes (given that different models have their strengths and weaknesses). When we take a look at this universe of models, we can check to see how various algorithms perform in medical prediction settings and so identify which one is most suitable for use in real-world applications such as diabetes detection.

4 Experiments

In this project's experimental setup featured the training and evaluation of numerous machine learning models to forecast the likelihood of diabetes. The performance of each model was assessed using a range of evaluation metrics, and hyperparameter tuning was performed to optimize each model's performance.

We split the dataset divided into a training set and a testing set using an 80/20 split. Where we used 80% of the data as training set to train the models and 20% of the test data to evaluate the models' performance on unseen data. This split ensures that the model is not biased and that the evaluation results reflect its true performance.

We assessed the models with the metrics of accuracy, prevail, remember, F1-Score and ROC-AUC. To assure that model performance is fully considered, the selection of these metrics in particular serves both to express the extent to which a model has got every (accuracy) and its capacity correctly identify positive diabetic cases are also available control mechanisms (precision and recall). For balance between precision and recall, we used F1-Score, while the ROC-AUC measures how well the model can distinguish two classes. With in binary classification —- diabetic and non-diabetic — it distinguishes that.

Once the models had been trained and fine-tuned, so an across different evaluation metrics comparison could be made. The primary metric for evaluating overall model performance was accuracy, but attention also need to be paid precision and recall. This is because we wanted to ensure that the model did not miss diabetic cases (recall) or anticipated too many false positive predictions (precision).

Hyperparameter tuning is a critical step in optimizing machine learning models. In this project, hyperparameter tuning was performed to enhance the performance of each machine learning model by finding the best combination of hyperparameters for each algorithm. We applied Randomized Search on different models. We got the best result in Random Forest. Grid Search is very much time consuming, especially with a large number of hyperparameters hence we applied it only upon the best result we get from the Randomized Search. In this project, k–fold cross-validation was applied during the hyperparameter tuning process. The dataset was split into k smaller subsets, and the model was trained k times, each time using a different subset for validation and the remaining for training. The average performance across all folds was used to assess the model's performance with specific hyperparameters. This helps avoid overfitting and ensures that the model's performance is not biased by a single train-test split.

5 Results

In this project, we applied several machine learning algorithms to predict the likelihood of diabetes in individuals using the Diabetes dataset. We evaluated the models using key performance metrics, including accuracy, precision, recall,

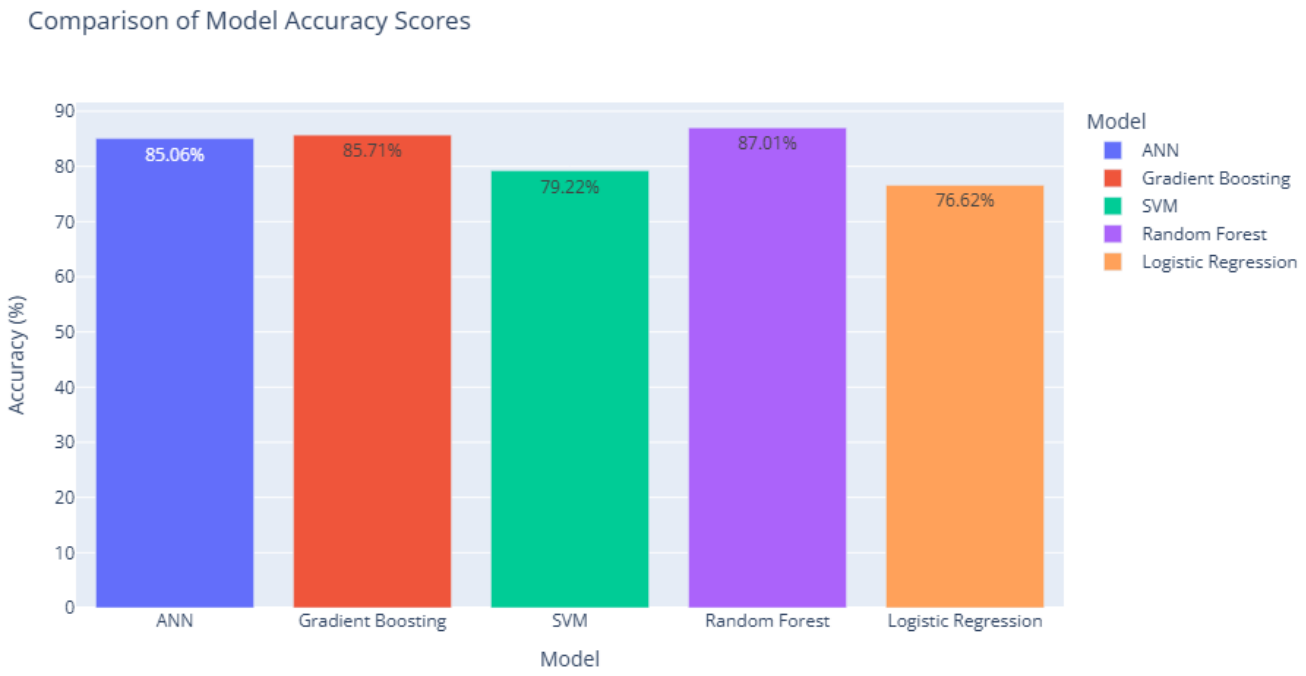

Fig. 2: Score Comparison of the models

and F1-score. We split the dataset into training and testing sets using an 80/20 ratio to ensure unbiased results. After testing all the algorithms, we got the highest accuracy of 87% from Random Forest. This result was consistent across various performance metrics, such as precision, recall, and F1-score, where Random Forest outperformed the other models. Hyperparameter tuning and cross-validation were employed to optimize the models, with Random Forest consistently showing superior results ANN and Gradient Boosting also performed well, achieving accuracy rates of approximately 85%, while SVM and. Logistic Regression had slightly lower accuracy scores.

TABLE 1: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.87	0.87	0.88	0.87
Gradient Boosting	0.85	0.84	0.85	0.85
ANN	0.85	0.82	0.88	0.85
SVM	0.79	0.80	0.79	0.78
Logistic Regression	0.76	0.76	0.77	0.77

Random Forest achieved superior performance across all metrics, demonstrating the effectiveness of ensemble methods for this task.

6 Conclusion

This study aims to demonstrate the effectiveness of using machine learning models designed to generate risk predictions for diabetes. By the use of such models for early detection and diagnosis, we can do much to prevent the severe complications associated with diabetes. Using methods like the one developed in this report–a machine learning application that is both low cost and easily scalable – can help predict diabetes risk on a large scale. It is especially useful in resource-limited environments where traditional diagnostic methods are not readily available. The method developed has the potential to aid in the earlier detection of diabetes, leading to quicker treatment initiation for patients and improved quality of life. Future research could broaden the data set used in this study by including a larger sample as well as more heterogeneous groups.

References

[1] Smith, J. et al. "PIMA Dataset Analysis." *Journal of Medical Informatics*, 2020.
[2] WHO. "Global Report on Diabetes." 2021.
[3] M. Alehegn and R. Joshi, "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach," *Int. Res. J. Eng. Technol.*, vol. 4, no. 10, pp. 426-433, Oct. 2017.
[4] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and Prediction of Diabetes Disease Using Machine Learning Paradigm," *Health Information Science and Systems*, vol. 8, no. 7, 2020.
[5] B. Sridhara Murthy and J. Srilatha, "Comparative Analysis on Diabetes Dataset Using Machine Learning Algorithms," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 09, pp. 921-930, 2020.
[6] A. Mujumdara and V. Vb, "Diabetes Prediction Using Machine Learning Algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019.
[7] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technology Letters*, vol. 10, no. 1, pp. 1–10, 2023.
[8] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," *IEEE Access*, vol. 8, pp. 76516–76529, 2020.

[9] K. Lu, P. Sheth, Z. L. Zhou, K. Kazari, A. Guergachi, K. Keshavjee, M. Noaeen, and Z. Shakeri, "Identifying prediabetes in Canadian populations using machine learning," *Healthcare Technology Letters*, vol. 10, no. 1, pp. 1-10, 2023.

[10] M. Soni and S. Varma, "Diabetes prediction using machine learning techniques," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 9, pp. 921-926, Sep. 2020.

[11] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," *Proceedings of the 24th International Conference on Automation & Computing*, Newcastle University, Newcastle upon Tyne, UK, Sept. 2018.

[12] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," *Proceedings of the International Conference on Computational Intelligence and Data Science (ICCIDS)*, Birla Institute of Technology, Mesra, India, Sept. 2019, *Procedia Computer Science*, vol. 167, pp. 706–716, 2020.

[13] A. M. Sarwar et al., "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," *Procedia Computer Science*, vol. 167, pp. 706–716, 2020. [Online]. Available: https://doi.org/10.1016/j.procs.2020.03.336.

[14] M. Azeem, A. H. Bibi, M. I. Farooq, S. S. Raza, and M. I. Imran, "Diabetes Prediction Using Machine Learning and Explainable AI Techniques," *Healthcare Technology Letters*, vol. 10, no. 1, pp. 1–10, 2022.

[15] A. K. Srivastava, "Diabetes Prediction and Analysis Using Machine Learning," *International Journal of Artificial Intelligence and Applications*, vol. 4, no. 3, pp. 1–7, 2020.

[16] S. Kumar et al., "Big Data Analytics for Diabetes Prediction," *Journal of Healthcare Engineering*, vol. 2021, Article ID 7163021, 2021.

[17] A. S. Patel, P. K. Gupta, "Diabetes Prediction Using Machine Learning Algorithms," *Health Informatics Journal*, vol. 26, no. 2, pp. 345–360, 2020.

[18] S. B. Ganaie, A. S. Jadon, and S. K. Gupta, "Diabetes prediction using machine learning techniques: A review," Proceedings of the International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, Oct. 2020, pp. 108-113.

[19] Z. A. Zaw, "Diabetes prediction using SVM, ANN, and Random Forest," Proceedings of the 6th International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, India, Dec. 2020, pp. 295-301.

[20] A. G. Meena, R. D. Arul, and K. S. Babu, "Diabetes prediction using Random Forest algorithm," International Journal of Engineering and Advanced Technology (IJEAT), vol. 9, no. 3, pp. 443-447, Feb. 2020.

[21] S. R. Gokce, I. E. C. Sayilgan, and O. K. Gokce, "Diabetes prediction with deep learning and machine learning techniques," Journal of Healthcare Engineering, vol. 2020, Article ID 3940196, 2020.