ISSN 2277-2685
IJESR/July-September. 2026/ Vol-16/Issue-3/137-143
B Harshitha Reddy *et. al.*, /International Journal of Engineering & Science Research
Full Length Article

Diabetes Prediction Using Machine Learning

B Jyothsna[1], B Harshitha Reddy[2], M Keerthana[3], G Keerthi Sree[4]
[1]Associate Professor; Department of Electronics and Communication Engineering Bhoj Reddy Engineering College for Women Hyderabad, India.
[2,3,4]B. Tech Students; Department of Electronics and Communication Engineering Bhoj Reddy Engineering College for Women Hyderabad, India.
Mail Id; harshithar895@gmail.com[2], mallekedikeerthana0417@gmail.com[3], keerthisree2005@gmail.com[4]
Accepted 27-06-2026

Abstract
Diabetes is one of the most common chronic diseases affecting people globally. It occurs when the body is unable to properly regulate blood sugar levels, leading to severe health complications such as heart disease, kidney failure, and nerve damage. Early detection of diabetes plays a crucial role in controlling the disease and preventing long-term complications.
In this project, a Machine Learning-based approach is used to predict whether a person is diabetic or not. The model is built using the Support Vector Machine (SVM) algorithm with a linear kernel, which is highly effective for classification problems. The dataset used consists of various medical parameters such as glucose level, blood pressure, insulin level, Body Mass Index (BMI), age, and other relevant health attributes.
The system includes data preprocessing, feature scaling using Standard Scaler, and model training and testing using split datasets. The trained model is deployed through a Django web application, allowing users to enter their health details and instantly receive diabetes prediction results. The system also stores patient records for future reference and monitoring.
This project demonstrates how Machine Learning can be effectively applied in healthcare to support early diagnosis, improve clinical decision-making, and enhance patient care through accurate and efficient diabetes prediction.

Keywords
Diabetes Prediction, Machine Learning, Support Vector Machine (SVM), Standard Scaler, Classification, Healthcare Analytics, Django, Medical Diagnosis, Blood Glucose Analysis, Early Disease Detection.

Introduction

Diabetes mellitus is one of the fastest-growing chronic metabolic disorders worldwide and represents a major public health challenge. The disease occurs when the body is unable to produce sufficient insulin or effectively utilize the insulin it produces, resulting in persistently elevated blood glucose levels. The increasing prevalence of diabetes is closely associated with urbanization, sedentary lifestyles, unhealthy dietary patterns, obesity, aging populations, and hereditary factors. As the number of affected individuals continues to rise, diabetes has become a significant cause of morbidity, mortality, and healthcare expenditure across both developed and developing nations.

Persistent hyperglycemia can gradually damage multiple organs and body systems if the condition is not identified and managed at an early stage. Diabetes is strongly associated with severe complications including cardiovascular disorders, chronic kidney disease, peripheral neuropathy, diabetic retinopathy, and impaired wound healing. These complications substantially reduce patients' quality of life while increasing the economic burden on healthcare systems. Consequently, identifying individuals who are at risk before the onset of advanced symptoms has become an important objective in modern healthcare.

Conventional diabetes diagnosis primarily depends on laboratory investigations such as fasting plasma glucose tests, oral glucose tolerance tests, and glycated hemoglobin (HbA1c) measurements, followed by clinical assessment. Although these diagnostic procedures provide reliable results, they require medical infrastructure, trained professionals, and laboratory facilities that may not always be readily available, particularly in resource-limited regions. Furthermore, diagnosis is often performed after noticeable clinical manifestations appear, limiting opportunities for early preventive intervention.

The rapid growth of artificial intelligence and data-driven healthcare has created new opportunities for improving disease prediction and clinical decision

support. Machine Learning (ML) techniques are capable of extracting meaningful relationships from large healthcare datasets, enabling accurate prediction of disease risk based on multiple clinical attributes. These intelligent approaches assist healthcare professionals by supporting faster diagnosis, improving decision-making, and facilitating personalized preventive care.

Among the various supervised machine learning algorithms, the Support Vector Machine (SVM) has gained considerable attention for binary classification tasks because of its strong generalization capability, robustness, and effectiveness when working with multidimensional medical data. The algorithm determines an optimal separating hyperplane that maximizes the margin between different classes, thereby improving classification performance while minimizing prediction errors. Owing to these characteristics, SVM has been successfully applied in numerous healthcare applications, including disease diagnosis and risk assessment.

This research presents a diabetes prediction system based on a Support Vector Machine employing a linear kernel. The proposed model is developed using a dataset containing important clinical attributes, including glucose concentration, blood pressure, insulin level, Body Mass Index (BMI), age, skin thickness, pregnancies, and diabetes pedigree function. Prior to model development, the dataset undergoes preprocessing and feature normalization using the Standard Scaler technique to improve model performance and ensure consistent feature representation. The processed data are subsequently divided into training and testing subsets for model development and evaluation.

To enhance practical usability, the trained prediction model is integrated into a Django-based web application that enables users to enter their medical information and obtain immediate prediction results. The application also maintains patient records for future reference and monitoring. By combining machine learning with a user-friendly web interface, the proposed system provides an accessible tool for supporting early diabetes screening and assisting healthcare professionals in clinical decision-making. The study demonstrates the potential of machine learning techniques to improve diagnostic efficiency, encourage timely medical intervention, and contribute to preventive healthcare practices.

Literature Survey

Diabetes prediction has become an active area of research due to the rapid growth of machine learning and artificial intelligence in healthcare. Numerous studies have focused on developing intelligent prediction systems capable of identifying individuals at risk of diabetes using clinical and demographic information. Various supervised learning algorithms, data mining techniques, and hybrid models have been investigated to improve prediction accuracy, reduce diagnosis time, and support clinical decision-making. The following literature summarizes the major contributions in this field.

1. Clinical Prediction Models for Diabetes

Several researchers have investigated clinical prediction models that combine patient history, laboratory reports, and physiological measurements to estimate diabetes risk. These studies demonstrate that predictive analytics can assist healthcare professionals in disease diagnosis, risk assessment, self-management, and preventive healthcare by utilizing patient-specific clinical information.

2. Machine Learning-Based Diabetes Prediction

A considerable number of studies have explored the application of machine learning algorithms for diabetes prediction. Commonly used techniques include Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB), Artificial Neural Networks (ANN), and K-Nearest Neighbors (KNN). Comparative analyses indicate that machine learning models are capable of identifying complex relationships among clinical variables, thereby improving prediction performance over conventional statistical approaches.

3. Early Prediction of Type-2 Diabetes

Recent research has focused on the early prediction of Type-2 Diabetes Mellitus (T2DM) using historical medical records and patient health parameters. These studies emphasize that early identification of high-risk individuals enables timely medical intervention, lifestyle modification, and continuous monitoring, thereby reducing the likelihood of severe diabetic complications.

4. Data Mining Approaches for Diabetes Diagnosis

Data mining techniques have been extensively employed to discover hidden patterns and meaningful relationships within healthcare datasets. Researchers have utilized classification, clustering, association rule mining, and feature selection methods to improve the accuracy and efficiency of diabetes prediction models. These approaches have demonstrated significant potential for supporting intelligent clinical decision systems.

5. Risk Prediction Frameworks

Several frameworks have been proposed for early-stage diabetes risk assessment by integrating data preprocessing, feature engineering, classification algorithms, and performance evaluation. These frameworks aim to provide reliable prediction systems

that assist physicians in identifying individuals who require further clinical examination and preventive treatment.

Working Methodology

Existing System

The conventional diagnosis of diabetes primarily depends on laboratory investigations and clinical evaluation performed by healthcare professionals. Common diagnostic procedures include fasting blood glucose tests, postprandial blood glucose tests, oral glucose tolerance tests, and glycated hemoglobin (HbA1c) analysis. Physicians evaluate these test results together with patient characteristics such as age, body weight, family medical history, and lifestyle before confirming the presence of diabetes.

Although these diagnostic procedures provide reliable clinical evidence, they present several practical limitations. The diagnosis process generally requires laboratory facilities, trained medical personnel, and multiple patient visits, increasing both the time and cost of healthcare. Such requirements may limit access to timely diagnosis, particularly in rural and resource-constrained regions where healthcare infrastructure is inadequate.

Another limitation of the conventional approach is its dependence on manual clinical interpretation, which may introduce variability among healthcare providers. Furthermore, traditional diagnostic systems typically offer limited predictive capability and are primarily intended to confirm existing disease rather than identify individuals who are at high risk before clinical symptoms become apparent. These limitations emphasize the need for an intelligent prediction system capable of providing rapid, automated, and accessible diabetes risk assessment.

2.3 Proposed System

The proposed system employs Machine Learning techniques to predict diabetes using patient health information. A Support Vector Machine (SVM) with a linear kernel is selected as the classification algorithm because of its high classification accuracy, robustness, and suitability for binary classification problems involving medical datasets. The complete prediction model is deployed through a Django web application, enabling users to obtain prediction results through a simple web interface.

System Overview

The proposed framework integrates data preprocessing, feature scaling, machine learning classification, and web-based deployment into a unified prediction platform. Users provide their clinical information through the application interface, after which the trained model analyzes the data and predicts whether the individual is diabetic or non-diabetic. The system is designed to deliver accurate predictions within a short response time while maintaining ease of use.

Flow Chart

The operational workflow of the proposed diabetes prediction system is illustrated in Figure 2.4.1. The process begins with user registration and data entry through the web interface. The entered clinical parameters undergo preprocessing and feature scaling before being supplied to the trained Support Vector Machine classifier. Based on the classification result, the system predicts whether the individual is diabetic or non-diabetic, displays the prediction, and stores the patient information for future reference.

Figure 1: Flow Chart of Diabetes Prediction Using Machine Learning

Methodology

Data Collection

The prediction model is trained using a publicly available medical dataset containing records of diabetic and non-diabetic individuals. The dataset consists of clinically significant attributes including glucose concentration, blood pressure, insulin level, skin thickness, Body Mass Index (BMI), diabetes pedigree function, age, and pregnancy count. Collecting a reliable and representative dataset is essential because the quality of training data directly influences the predictive capability and generalization performance of the machine learning model.

Data Preprocessing

Raw medical datasets frequently contain incomplete records and inconsistent values that can negatively affect prediction accuracy. Therefore, preprocessing is performed before model training. Invalid zero values present in attributes such as glucose level, blood pressure, BMI, insulin, and skin thickness are treated as missing observations and replaced using appropriate statistical imputation methods such as mean or median substitution. This process enhances data consistency and improves the reliability of the classification model.

Feature Scaling

Since the clinical attributes possess different numerical ranges, feature scaling is performed prior to model training. Standardization using the StandardScaler technique transforms each feature to have zero mean and unit variance. This normalization process prevents attributes with larger magnitudes from dominating the learning process and significantly improves the classification performance of the Support Vector Machine. Proper feature scaling also contributes to faster convergence and enhanced model stability during training.

Software Implementation

This chapter describes the software technologies, development environment, programming tools, and implementation components used to develop the proposed diabetes prediction system. The application combines machine learning techniques with a web-based platform to provide accurate and efficient diabetes prediction. The implementation consists of data preprocessing, model training, deployment, and user interaction modules, all integrated into a unified system. Appropriate software tools and libraries were selected to ensure reliability, scalability, and ease of maintenance.

Software Specifications

Operating System: Windows 11

The application is developed and executed on the Windows 11 operating system, which provides a stable and compatible environment for software development. It supports the installation and execution of Python, Django, Visual Studio Code, and other required development tools. The operating system also ensures smooth integration between the machine learning model and the web application.

Programming Language: Python 3.12

Python 3.12 serves as the primary programming language for implementing the proposed system. Its simple syntax, extensive standard libraries, and strong support for scientific computing make it an ideal choice for machine learning applications. Python facilitates efficient data processing, model development, and integration with web technologies.

Framework: Django

The web application is developed using the Django framework, a high-level Python framework designed for rapid and secure web application development. Django manages user requests, URL routing, backend processing, and template rendering while providing built-in security mechanisms for authentication, session management, and database interaction. Its modular architecture simplifies application development and maintenance.

Machine Learning Library: Scikit-learn

Scikit-learn is employed for implementing the Support Vector Machine (SVM) classifier used in diabetes prediction. The library provides optimized algorithms for model training, prediction, evaluation, and performance analysis. It also includes utilities for preprocessing, feature scaling, and dataset partitioning, enabling efficient development of the classification model.

Data Processing Library: Pandas

Pandas is utilized for loading, organizing, and preprocessing the medical dataset. The library offers powerful data structures that simplify handling missing values, performing statistical operations, filtering records, and preparing structured datasets suitable for machine learning. Efficient preprocessing improves the quality of input data and contributes to enhanced prediction accuracy.

Model Persistence Library: Joblib

The trained Support Vector Machine model and preprocessing objects are stored using the Joblib library. Saving the trained model eliminates the need for repeated training whenever the application is executed, thereby reducing computational overhead and improving system response time. Joblib provides efficient serialization and rapid loading of large machine learning models.

Front-End Technologies: HTML and CSS

The graphical user interface is developed using HTML and CSS. HTML defines the structure and content of the web pages, while CSS enhances their appearance through styling, layout management, and responsive design. The combination of these technologies provides a clean, intuitive, and user-friendly interface for entering medical information and viewing prediction results.

Development Environment: Visual Studio Code

Visual Studio Code (VS Code) is used as the primary integrated development environment for software implementation. The editor offers syntax highlighting, intelligent code completion, debugging facilities, integrated terminal support, extension management, and version control features that improve programming efficiency and simplify project maintenance.

Results and Discussion

The proposed diabetes prediction system was successfully implemented and tested using clinical information entered through the web application. Users provide essential medical attributes, including the number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, Body Mass Index (BMI), diabetes pedigree function, and age. These clinical parameters serve as the input variables for the prediction model.

Before classification, the submitted data undergoes preprocessing to ensure compatibility with the trained machine learning model. The preprocessing stage includes validation of user inputs and feature normalization using the StandardScaler technique. Standardization transforms the input variables into a common numerical scale, thereby improving the stability and classification performance of the Support Vector Machine.

The standardized feature vector is subsequently processed by the trained Support Vector Machine employing a linear kernel. Based on the relationships learned from historical medical data, the classifier predicts whether the individual belongs to the diabetic or non-diabetic category. The prediction is generated within a short response time, enabling efficient real-time decision support.

The prediction result is immediately displayed through the web interface in an easy-to-understand format, allowing users to interpret the outcome without requiring technical expertise. The rapid prediction capability demonstrates the suitability of the proposed system for preliminary diabetes screening and healthcare decision support.

In addition to prediction, the application automatically records each patient's clinical information together with the prediction result, date, and time in an Excel file. Maintaining these records facilitates future review, monitoring, statistical analysis, and follow-up by healthcare professionals or system administrators. The lightweight storage mechanism also eliminates the need for a dedicated database while providing an organized repository of prediction history.

Overall, the experimental implementation demonstrates that the proposed machine learning framework successfully integrates data preprocessing, intelligent classification, and web-based deployment into a single healthcare application. The system provides accurate prediction, efficient processing, simplified record management, and an intuitive user interface, making it suitable for practical diabetes risk assessment.

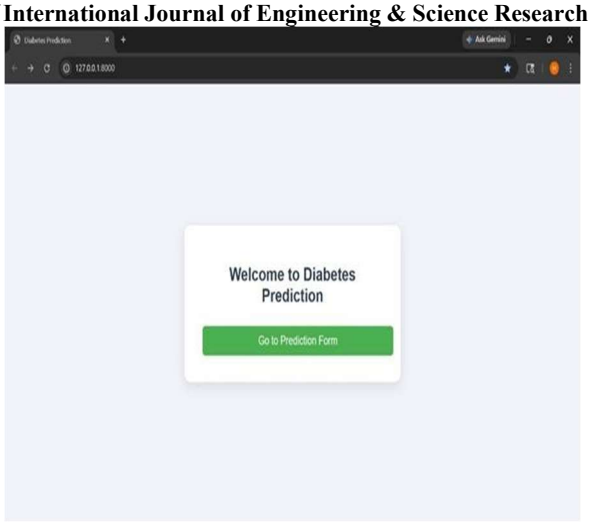


Figure 2: Home Page of the Diabetes Prediction System

Figure 2 illustrates the home page of the developed web application. It functions as the primary interface through which users access the diabetes prediction system. The homepage presents a clean and intuitive design with navigation options that guide users to the prediction module. The simple interface enhances usability and enables users with limited technical knowledge to operate the application efficiently.

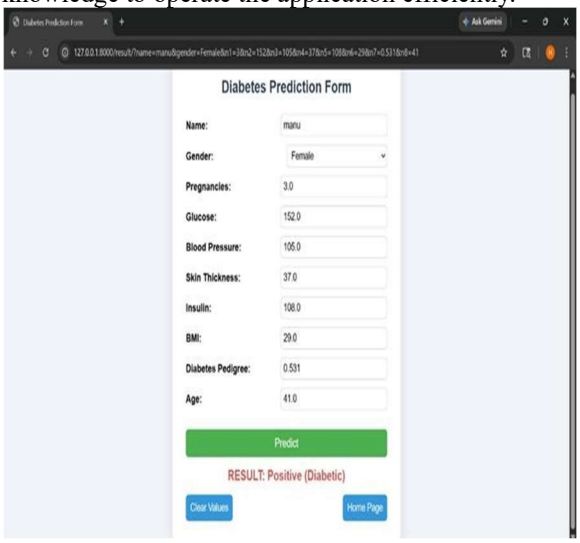
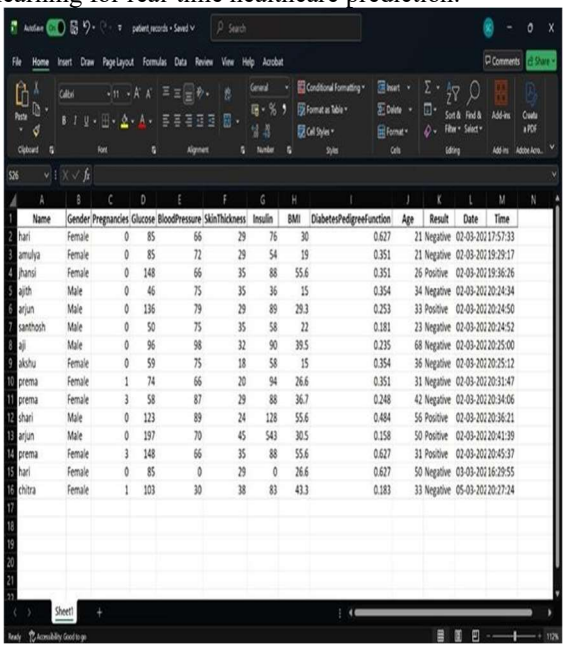


Figure 3: Diabetes Prediction Interface

Figure 3 presents the prediction interface of the proposed system. Users enter their clinical information, including pregnancies, glucose level, blood pressure, skin thickness, insulin concentration, Body Mass Index (BMI), diabetes pedigree function, and age. After submission, the entered values are preprocessed and analyzed by the trained Support Vector Machine model. The application then displays the prediction result, indicating whether the individual is likely to be diabetic or non-diabetic. This interface

demonstrates the practical implementation of machine learning for real-time healthcare prediction.



Name	Gender	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Result	Date	Time
hari	Female	0	85	66	29	76	30	0.627	21	Negative	02-03-2021	7:57:33
amulya	Female	0	85	72	29	54	19	0.351	21	Negative	02-03-2021	9:29:17
jhansi	Female	0	148	66	35	88	55.6	0.351	26	Positive	02-03-2021	9:36:26
ajith	Male	0	46	75	35	36	15	0.354	34	Negative	02-03-2021	0:24:34
arjun	Male	0	136	79	29	89	29.3	0.253	33	Positive	02-03-2021	0:24:50
santhosh	Male	0	50	75	35	58	22	0.181	23	Negative	02-03-2021	0:24:52
aji	Male	0	96	98	32	90	39.5	0.235	68	Negative	02-03-2021	0:25:00
akshu	Female	0	59	75	18	58	15	0.354	36	Negative	02-03-2021	0:25:12
prema	Female	1	74	66	20	94	26.6	0.351	31	Negative	02-03-2021	0:31:47
prema	Female	3	58	87	29	88	36.7	0.248	42	Negative	02-03-2021	0:34:06
shari	Male	0	123	89	24	128	55.6	0.484	56	Positive	02-03-2021	0:36:21
arjun	Male	0	197	70	45	543	30.5	0.158	50	Positive	02-03-2021	0:41:39
prema	Female	3	148	66	35	88	55.6	0.627	31	Positive	02-03-2021	0:45:37
hari	Female	0	85	0	29	0	26.6	0.627	50	Negative	03-03-2021	6:29:55
chitra	Female	1	103	30	38	83	43.3	0.183	33	Negative	05-03-2021	0:27:24

Figure 4: Patient Records Stored in Excel

Figure 4 illustrates the Excel-based repository used for storing patient information and prediction outcomes. Each record includes the patient's input parameters, predicted diagnosis, date, and time of prediction. The organized tabular format enables efficient data management, facilitates future analysis, and supports clinical follow-up. The storage mechanism also provides a simple and cost-effective alternative to database management for maintaining prediction records during system operation.

Conclusion

This research presented a machine learning-based diabetes prediction system using the Support Vector Machine (SVM) algorithm with a linear kernel. The developed model utilizes important clinical parameters, including glucose level, blood pressure, insulin level, Body Mass Index (BMI), diabetes pedigree function, age, skin thickness, and the number of pregnancies, to classify individuals as diabetic or non-diabetic. Before classification, the dataset undergoes preprocessing and feature standardization to improve the robustness, consistency, and predictive performance of the model. The trained prediction model was successfully integrated into a Django-based web application, enabling users to enter their medical information and receive instant prediction results. In addition, the application stores patient records along with the prediction outcome and timestamp, facilitating future analysis and efficient record management. The experimental implementation demonstrates that the proposed system provides a reliable, efficient, and user-friendly solution for preliminary diabetes screening. By integrating machine learning with web technologies, the application supports early disease detection, reduces the time required for initial assessment, and assists healthcare professionals in clinical decision-making. The proposed framework highlights the potential of intelligent healthcare systems to improve accessibility, encourage preventive healthcare practices, and support timely medical intervention. Although the developed system demonstrates satisfactory prediction performance, further improvements can be achieved through the use of larger and more diverse datasets, advanced feature engineering methods, and sophisticated machine learning techniques, thereby enhancing prediction accuracy and the generalization capability of the model for real-world healthcare applications.

Future Scope

The proposed diabetes prediction system offers considerable opportunities for future enhancement to improve its predictive capability, scalability, and practical applicability. Future research may employ larger and more diverse datasets collected from multiple healthcare institutions, enabling the model to learn from a broader range of patient characteristics and improving its generalization performance across different populations. In addition, advanced machine learning and deep learning algorithms such as Random Forest, XGBoost, Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks can be explored and compared with the existing Support Vector Machine classifier to further improve prediction accuracy and overall system performance. The current Excel-based storage mechanism can also be replaced with secure database management systems such as MySQL, PostgreSQL, or MongoDB to provide better scalability, improved data integrity, efficient retrieval, and effective management of patient records. Furthermore, the application can be deployed on cloud platforms to enable secure online access for hospitals, clinics, and individual users, thereby improving system availability, centralized management, and remote accessibility. Future versions may also incorporate an enhanced user interface with responsive web design, interactive dashboards, and improved navigation features to provide a better user experience for both healthcare professionals and patients. Additional functionalities, including graphical data visualization, statistical reports, and trend analysis, can support physicians in monitoring disease progression and making informed clinical decisions. Considering the sensitive nature of

healthcare information, future implementations should integrate advanced security mechanisms such as data encryption, secure authentication, role-based access control, and compliance with healthcare data protection standards to ensure confidentiality and privacy. The prediction system may also be integrated with Internet of Things (IoT) devices and wearable healthcare sensors to enable automatic collection of physiological data and support continuous real-time diabetes monitoring. Finally, the proposed framework can be extended beyond diabetes prediction to include the diagnosis of other chronic diseases, such as cardiovascular disease, hypertension, kidney disease, and liver disorders, thereby evolving into a comprehensive intelligent healthcare platform capable of supporting multi-disease prediction and clinical decision-making.

References

[1] A. Sheik Abdullah, V. N. P. Shashank, and D. A. L. Hudson, "Disseminating the Risk Factors With Enhancement in Precision Medicine Using Comparative Machine Learning Models for Healthcare Data," *IEEE Access*, vol. 12, 2024.

[2] P. G. Jacobs *et al*., "Artificial Intelligence and Machine Learning for Improving Glycemic Control in Diabetes: Best Practices, Pitfalls, and Opportunities," *IEEE Reviews in Biomedical Engineering*, vol. 17, 2024.

[3] A. A. Linkon *et al*., "Evaluation of Feature Transformation and Machine Learning Models on Early Detection of Diabetes Mellitus," *IEEE Access*, vol. 12, 2024.

[4] N. N. N. Nazirun *et al*., "Prediction Models for Type 2 Diabetes Progression: A Systematic Review," *IEEE Access*, vol. 12, 2024.

[5] A. Ahmed, M. A. Hossain, and M. A. Rahman, "A Machine Learning Approach for Diabetes Prediction and Diagnosis," *IEEE Access*, vol. 8, 2022.

[6] P. P. Priya, S. V. Reddy, and K. S. Reddy, "Diabetes Prediction using Support Vector Machine and Machine Learning Techniques," *IEEE International Conference on Intelligent Systems and Control*, 2023.