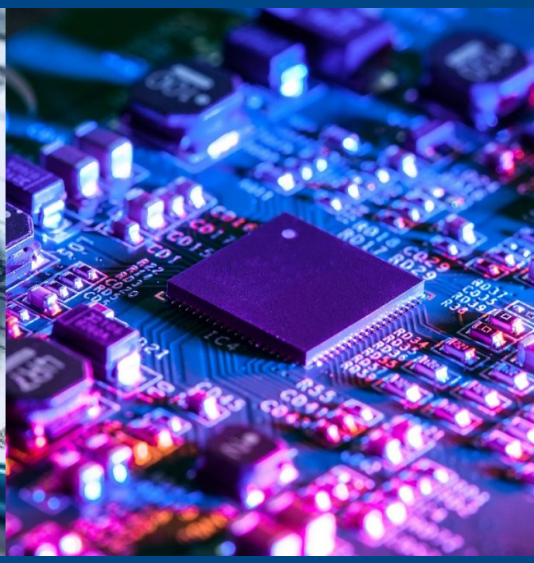
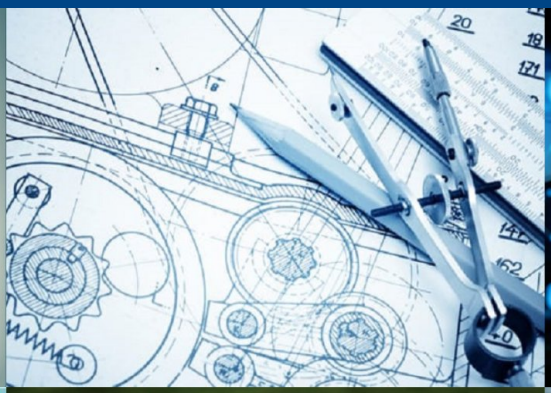
ISSN: 2582-7219

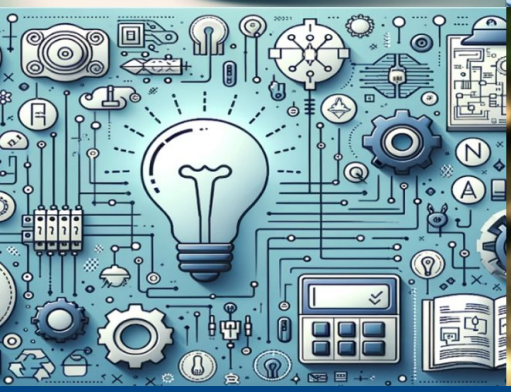




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904256

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Smart Diabetes Risk Detection System Using Machine Learning and Health Behavior Analysis

M Kalaivani.M.E , T. Arulkumar, S. Ashwin, U. Brindha shri, G.Durga

Assitant Professor, Department of CSE, Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India

Computer Science and Engineering Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India

Computer Science and Engineering Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India

Computer Science and Engineering Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India

Computer Science and Engineering Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India

ABSTRACT -Diabetes is a rapidly increasing chronic metabolic disorder that poses a significant threat to global public health due to its long-term complications such as cardiovascular diseases, kidney failure, and neuropathy. One of the major challenges in diabetes management is its asymptomatic nature during the early stages, which often leads to delayed diagnosis and treatment. Early risk prediction and preventive intervention are therefore essential to reduce the disease burden. This project presents a web-based Primary Stage Diabetes Risk Prediction System that leverages machine learning techniques to enable early detection and risk assessment using both clinical and lifestyle-related parameters. The proposed system integrates multiple input features, including key medical indicators such as blood glucose level, body mass index (BMI), blood pressure, insulin level, and age. along with lifestyle factors such as physical activity, sleep duration, junk food consumption, sugar intake, and smoking habits. The collected data is preprocessed through cleaning. normalization, and feature selection techniques to ensure data quality and consistency. A hybrid approach combining weighted clinical risk scoring and supervised machine learning models is employed to improve prediction accuracy. Three widely used classification algorithms-Random Forest. Decision Tree, and Logistic Regression-are implemented and comparatively analyzed to identify the most efficient model based on performance metrics such as accuracy, precision, and recall. The system classifies individuals into Low Risk. Moderate Risk, and High Risk categories based on prediction outcomes. A user-friendly web interface is designed to provide an interactive dashboard that visualizes results through feature importance graphs, dataset distribution charts, and algorithm comparison plots. Additionally, a color-coded risk indicator enhances interpretability for users with varying levels of technical knowledge. The system further generates automated health recommendations tailored to the predicted risk level, including dietary guidance, physical activity suggestions, and medical consultation advice. By combining data-driven predictive analytics with an intuïtive web-based platform, the proposed system aims to facilitate early diabetes screening, improve healthcare awareness, and support proactive decision- making. This approach not only enhances prediction efficiency but also contributes to preventive healthcare by enabling individuals to take timely actions to reduce the risk of developing diabetes

KEYWORDS: Diabetes Prediction, Machine Learning. Risk Assessment, Healthcare Analytics, Random Forest, Decision Tree, Logistic Regression, Predictive Modeling

I. INTRODUCTION

Diabetes is a chronic metabolic disorder that has become one of the most serious global health concerns. It is characterized by elevated blood glucose levels resulting from defects in insulin secretion, insulin action, or both. Early identification of diabetes risk is essential, as it enables timely intervention and helps prevent severe complications such as cardiovascular diseases, kidney failure, and nerve damage.

Among the different stages of diabetes, Prediabetes represents a critical phase where blood glucose levels are higher than normal but not yet high enough to be classified as diabetes. Detecting this stage early provides a valuable opportunity to delay or even prevent the onset of full diabetes through appropriate lifestyle modifications and medical guidance.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904256

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Traditional diagnostic approaches primarily rely on clinical tests and physician evaluation, which may not always be accessible or conducted at the right time. In recent years, advancements in machine learning have enabled the development of intelligent systems capable of analyzing large volumes of health-related data to identify hidden patterns and predict disease risks effectively.

This project proposes a machine learning-based system for the early prediction of diabetes risk using both medical and lifestyle-related parameters. The system utilizes algorithms such as Logistic Regression, Decision Tree, and Random Forest to analyze input data and classify individuals into different risk categories.
By integrating user-friendly interfaces with predictive analytics, the proposed system aims to provide an accessible and efficient tool for early diabetes risk detection, thereby supporting preventive healthcare and promoting awareness among individuals.

The main contributions of this work include:

- Developed a machine learning model to predict early-stage diabetes risk.
- Utilized both medical and lifestyle data for accurate prediction.
- Implemented multiple algorithms to improve model performance.
- Designed a simple system for easy user input and quick results.
- Provided risk classification with basic health recommendations.

II. LITERATURE REVIEW

Diabetes prediction has been a widely researched area in the field of healthcare analytics, especially with the advancement of machine learning techniques. Several studies have focused on the early detection of diabetes by analyzing medical and lifestyle-related data. Early prediction plays a crucial role in preventing the progression of Prediabetes into full-stage diabetes.

Many researchers have applied traditional machine learning algorithms such as Logistic Regression for binary classification of diabetic and non-diabetic individuals. This approach is widely used due to its simplicity and interpretability, making it suitable for medical applications where understanding the model is important.

In addition, tree-based models like Decision Tree have been employed to capture non-linear relationships between features such as glucose levels, body mass index (BMI), and age. These models provide a structured way to represent decision-making processes, which is beneficial for identifying risk factors.

Furthermore, ensemble learning techniques such as Random Forest have demonstrated higher accuracy compared to individual models. By combining multiple decision trees, Random Forest reduces overfitting and improves the robustness of predictions, making it highly effective for medical datasets.

Recent studies have also emphasized the importance of incorporating lifestyle factors such as physical activity, dietary habits, and sleep patterns along with clinical data. This combined approach enhances the predictive performance and provides a more comprehensive assessment of an individual's diabetes risk.

Overall, existing research highlights that machine learning-based systems can significantly improve early detection of diabetes. However, there is still a need for user- friendly and accessible systems that can integrate multiple data sources and provide real-time risk predictions. This project aims to address these gaps by developing an efficient and practical early-stage diabetes prediction system.

EXISTING SYSTEM

The existing systems for diabetes detection are mainly categorized into traditional clinical methods and basic data-driven approaches. These systems are widely used in healthcare but have limitations in early-stage prediction and real-time analysis.

A. Traditional clinical diagnosis system

This system relies on laboratory tests such as blood glucose and HbA1c levels to diagnose diabetes. The diagnosis is based on predefined threshold values, and the results are interpreted by healthcare professionals. Features:

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904256

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- Uses blood test parameters
- Threshold-based decision making
- Requires doctor analysis
- Uses BMI formula:

BMI = Weight (kg) / Height² (m)

B. Basic Machine learning -Based system

This system uses simple machine learning models to predict diabetes based on medical data. Algorithms are trained using datasets and evaluated using performance metrics such as accuracy.

Features:

- Uses medical dataset for prediction
- Applies algorithms like Logistic Regression and Decision Tree
- Evaluates model using:

Accuracy = (Correct Predictions / Total Predictions) × 100 Limited feature usage (mostly medical data only)

C. Comparative Limitations of Existing Systems

The comparison between the traditional clinical diagnosis system and the basic machine learning-based system highlights the differences in their efficiency, performance, and capability in early-stage diabetes prediction. While the clinical system relies on manual interpretation and fixed thresholds, the machine learning system introduces automation and faster processing. However, both systems still have limitations in achieving highly accurate early prediction. The comparison is shown below.

TABLE I-Comparative Analysis of Traditional Clinical Diagnosis and Basic Machine learning -Based Systems

Feature	**Clinical system**	**Basic ML System**
Data Usage	Limited	Available
Early Detection	Not Available	Limited
Processing Time	Slow	Fast
Automation	Not Available	Available
Lifestyle consideration	Not Available	Limited
User Accessibility	Limited	Available
Real-Time Prediction	Not Available	Limited

III. PROPOSED SYSTEM

The proposed system is an advanced machine learning-based framework designed to predict the early stage of diabetes using a combination of medical and lifestyle-related data. Unlike traditional systems, this approach focuses on identifying risk at an initial stage,particularly during Prediabetes, where preventive measures can significantly reduce disease progression.

The system collects various user inputs such as age, body mass index (BMI), glucose levels, physical activity, dietary habits, and sleep patterns. These inputs are processed and analyzed using intelligent algorithms to generate accurate predictions. The system not only identifies the risk level but also provides basic health recommendations, making it a supportive tool for preventive healthcare and awareness.

A. System Overview

The proposed system is an intelligent and user-centric machine learning framework developed for the early prediction of diabetes risk using a combination of medical and lifestyle-related parameters. The primary objective of this system is to identify individuals who are at risk during the early stage, particularly in conditions such as Prediabetes, where timely intervention can effectively prevent the progression of the disease.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904256

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The system collects various input parameters from users, including age, body mass index (BMI), blood glucose levels, physical activity, dietary habits, and sleep patterns. These parameters are considered critical indicators in assessing the risk of diabetes. Unlike traditional systems that rely only on clinical test results, the proposed system integrates both physiological and behavioral data to provide a more comprehensive analysis.Once the data is collected, it undergoes preprocessing to remove inconsistencies and ensure data quality. The processed data is then fed into machine learning models such as Logistic Regression, Decision Tree, and Random Forest. These models are trained to identify hidden patterns and relationships between input features and diabetes risk levels.

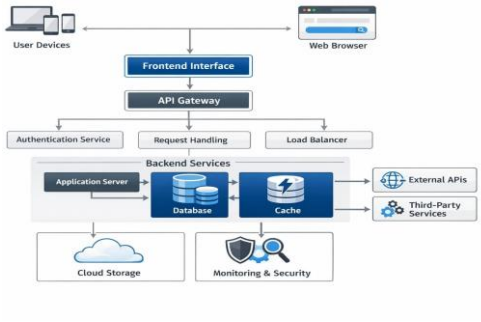


Fig. 1. System Architecture of the Proposed Early Diabetes Prediction

B. System Architecture

1. User Interface / Web Browser Layer

The user interface layer represents the point of interaction between the user and the system through a web browser. It allows users to access the application easily and input their personal, medical, and lifestyle-related data. This layer ensures accessibility across different devices and provides a seamless experience for users to interact with the diabetes prediction system.

2. Frontend Interface Layer

The frontend interface layer is responsible for designing and managing the visual components of the application. It collects ser inputs such as age, BMI, glucose levels, physical activity, dietary habits, and sleep patterns through structured forms. This layer ensures proper input validation and enhances user experience by providing a clean and intuitive interface.

3. API Gateway Layer

The system evaluates the input data and generates a prediction in the form of risk categories such as low, moderate, or high. In addition to prediction, the system also provides basic health recommendations to guide users toward preventive measures. This makes the system not only predictive but also supportive in promoting healthier lifestyle choices.

Furthermore, the proposed system is designed to be user-friendly and accessible, allowing individuals to use it without requiring advanced medical knowledge. The integration of automation ensures quick and efficient results, reducing the dependency on manual diagnosis. Overall, the system aims to bridge the gap between early detection and preventive healthcare by providing an efficient, accurate, and accessible solution for diabetes risk prediction.
The API gateway acts as an intermediary between the frontend and backend services, ensuring secure and efficient communication.

Key Functions:

- Authentication: Verifies user identity and ensures secure access to the system
- Request Handling: Manages incoming user requests and routes them to appropriate backend services
- Load Balancing: Distributes requests efficiently to maintain system performance and avoid overload

This layer plays a crucial role in maintaining system reliability and scalability.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904256

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

4. Backend Service Layer

The backend service layer is the core processing unit of the system, where all computations and predictions are performed.

omponents:

- Application Servers: Execute business logic, handle data processing, and run machine learning models such as Logistic Regression, Decision Tree, and Random Forest
- Database: Stores user input data, prediction results, and historical records
- Cache: Temporarily stores frequently accessed data to improve response time
- External APIs / Third-Party Services: Integrates additional services such as health recommendations or data sources
- This layer ensures efficient data processing, accurate prediction, and smooth system operation.

5. Cloud Storage Layer

The cloud storage layer is used to securely store large volumes of data, including user records and model outputs. It provides scalability, reliability, and remote accessibility, allowing the system to handle increasing data efficiently.

6. Monitoring & Security Layer

This layer ensures the overall safety and performance of the system.

Functions:

- Monitoring: Tracks system performance, usage, and errors in real-time
- Security: Protects user data through encryption, authentication, and secure communication protocols
- It helps in maintaining system integrity, preventing unauthorized access, and ensuring continuous system availability.

C. Data Processing & Feature Layer

The data processing and feature layer plays a crucial role in transforming raw input data into a structured format suitable for machine learning analysis. The collected data may contain missing values, inconsistencies, or noise, which can affect model performance. Therefore, preprocessing techniques such as data cleaning, normalization, and encoding are applied.

Feature engineering is also performed to select the most relevant attributes such as age, BMI, glucose level, physical activity, and diet pattern. These features significantly influence diabetes prediction.

Formula Used:

$$BMI = \text{Weight (kg)} / \text{Height}^2 \text{ (m)}$$

This formula helps in identifying obesity-related risks associated with diabetes.

D. Machine Learning Logic Layer

The Machine Learning Logic Layer is the core component of the proposed system, responsible for analyzing processed data and generating accurate predictions regarding diabetes risk. This layer utilizes advanced machine learning techniques to identify hidden patterns, correlations, and trends within the dataset, enabling early detection of diabetes risk, especially in conditions like Prediabetes.

The input to this layer consists of preprocessed and feature-engineered data, which includes both clinical parameters (such as glucose level, BMI, and blood pressure) and lifestyle-related attributes (such as physical activity, diet, and sleep patterns). These features are used by machine learning models to learn relationships between input variables and the target outcome.

The system employs multiple supervised learning algorithms, including Logistic Regression, Decision Tree, and Random Forest. Each of these algorithms contributes uniquely to the prediction process.

Logistic Regression is used as a baseline model to estimate the probability of diabetes occurrence. It applies a

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904256

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

sigmoid function to transform linear combinations of input features into probability values ranging between 0 and 1. Decision Tree, on the other hand, follows a tree-based structure where decisions are made by splitting data based on feature conditions, making it easy to interpret. Random Forest enhances prediction accuracy by combining multiple decision trees and aggregating their outputs, thereby reducing overfitting and improving generalization.To further improve prediction performance, the system may use an ensemble approach where predictions from multiple models are compared or combined. This ensures robustness and increases the reliability of the final output. The trained models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score to select the most efficient model.

Mathematical Representation

Logistic Regression:

$P(Y=1) = 1 / (1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)})$ Where:

$P(Y=1) \rightarrow$ Probability of diabetes

$\beta \rightarrow$ Model coefficients $x \rightarrow$ Input features

Decision Tree (Entropy): $Entropy(S) = - \Sigma p_i \log_2(p_i)$

Information Gain:

$IG = Entropy(parent) – Weighted\ Entropy(children)$

These formulas help in splitting data effectively and improving prediction accuracy.

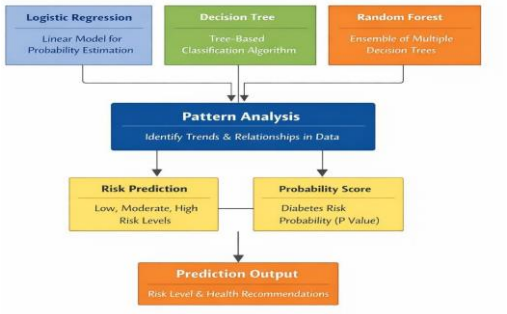


Fig.2: Machine Learning Logic Layer for Diabetes Prediction

E. Data Storage Layer

The Data Storage Layer is a critical component of the proposed diabetes prediction system, responsible for securely storing, managing, and retrieving all system-related data. This layer ensures that user inputs, processed data, and prediction results are maintained efficiently for both current use and future analysis. It plays a vital role in maintaining data integrity, consistency, and availability throughout the system lifecycle.

The system utilizes a lightweight and efficient database system such as SQLite for storing structured data. The database is designed to handle multiple types of data, including user demographic details, clinical parameters, lifestyle attributes, and prediction utcomes. Each user's data is stored in a structured format using tables, enabling easy retrieval and analysis.

The storage layer supports both temporary and permanent data storage. Temporary data, such as session inputs, may be stored in cache memory to improve system performance and reduce response time. Permanent data, including historical user records and prediction results, is stored in the database for long-term usage. This allows the system to track changes in user health over time and supports future enhancements such as trend analysis and personalized recommendations.

To ensure data security and privacy, appropriate measures such as data encryption and controlled access mechanisms are implemented. Only authorized components of the system can access or modify the stored data.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904256

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

This is particularly important as the system deals with sensitive health-related information.Additionally, the storage layer is designed to be scalable and reliable. It can handle increasing amounts of data as more users interact with the system. Backup andrecovery mechanisms can also be incorporated to prevent data loss and ensure system reliability.

F. Prediction Output Structure

The Prediction Output Structure represents the final stage of the proposed system, where the processed data and machine learning results are transformed into meaningful and user-understandable information. This layer plays a crucial role in communicating the prediction results effectively to the user, enabling informed decision-making and early preventive action.

The output generated by the system is not limited to a simple classification but includes a combination of probability scores, risk categorization, and personalized health suggestions. The machine learning models, particularly Logistic Regression and Random Forest, produce a probability value indicating the likelihood of diabetes occurrence. This probability is then mapped into predefined risk categories such as Low Risk, Moderate Risk, and High Risk.

The structured output ensures clarity and interpretability, even for users without medical or technical knowledge. By converting complex model predictions into simple categories and actionable insights, the system enhances usability and effectiveness.

Mathematical Representation of Output Probability Calculation:

$P = 1 / (1 + e^{(-x)})$

Where:

$P \rightarrow$ Probability of diabetes occurrence $x \rightarrow$ Weighted sum of input features

Risk Classification Function: Risk Level (R) =

Low, if $P < 0.4$

Moderate, if $0.4 \leq P < 0.7$ High, if $P \geq 0.7$

Prediction Function:

$Y = f(X)$

Where:

$Y \rightarrow$ Output prediction (Risk Level)

$X \rightarrow$ Input feature vector (age, BMI, glucose, lifestyle factors)

Output Components

The system generates multiple output elements to provide a

Where:$P = 1 / (1 + e^{(-Z)})$

complete understanding of the user's health condition:

- **Risk Level**: Indicates whether the user is at low, moderate, or high risk
- **Probability Score:** Numerical value representing the likelihood of diabetes
- **Health Recommendations:** Suggestions such as increasing physical activity, reducing sugar intake, and maintaining a balanced diet
- **Visualization (Optional):** Graphs or charts for better interpretation

G. Mathematical Model

The mathematical model defines the underlying computational logic used in the proposed system to predict diabetes risk. It represents the relationship between input features and the output prediction using statistical and machine learning formulations. The model transforms input data into probability values and classifies users into different risk levels.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904256

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

◆ Key Components of Mathematical Model

1) Input Feature Vector

The system considers multiple input parameters represented as a feature vector:

$X = \{x_1, x_2, x_3, \ldots, x_n\}$ P → Probability of diabetes occurrence e → Exponential constant

4) Loss Function (Model Training)

To improve prediction accuracy, the model minimizes error using:

$Loss = - [y \log(P) + (1 - y) \log(1 - P)]$

Where:

y → Actual output

P → Predicted probability

5) Risk Classification Function Based on probability value:

R =

Low Risk, if $P < 0.4$

Moderate Risk, if $0.4 \leq P < 0.7$ High Risk, if $P \geq 0.7$

6) Accuracy Calculation

The performance of the model is evaluated using: $Accuracy = (TP + TN) / (TP + TN + FP + FN)$

Where:

TP → True Positive TN → True Negative FP → False Positive FN → False Negative

Where:

x_1 → Age x_2 → BMI

x_3 → Glucose Level

x_4 → Physical Activity x_5 → Diet Pattern

...

2) Weighted Sum Calculation

Each feature is assigned a weight based on its importance: $Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n$

Where:

β → Model coefficients

Z → Linear combination of features

3) Probability Function (Sigmoid Function)

The system converts the weighted sum into probability using:Precision & Recall Precision:

$Precision = TP / (TP + FP)$

Recall:

$Recall = TP / (TP + FN)$

7) F1-Score

$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$

H. Advantages over existing system

- Advantages Over Existing System
- Enables early-stage diabetes prediction.
- Uses both medical and lifestyle data.
- Provides fast and automated results.
- Easy for users to access and use.

Reduces time and cost compared to traditional methods.MODEL

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904256

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. EVALUATION & VALIDATION MECHANISM

The Model Evaluation and Validation Mechanism ensures that the proposed diabetes prediction system performs accurately and consistently on both training and unseen data. This mechanism focuses on assessing the model's ability to generalize and avoid overfitting. Instead of relying only on basic metrics, advanced evaluation techniques are used to measure model performance more effectively.

The system applies validation strategies such as cross- validation and probabilistic evaluation to ensure robustness. These methods help in analyzing how well the model performs across different subsets of data, thereby improving reliability.

◆ Validation Techniques

- K-Fold Cross Validation
- ROC Curve Analysis
- Probability Calibration
- Error Rate Analysis
-

Mathematical Formulations (Advanced)

1) Error Rate:
Error Rate = (FP + FN) / Total Samples

2) Specificity (True Negative Rate): Specificity = TN / (TN + FP)
3) Sensitivity (Recall Alternative): Sensitivity = TP / (TP + FN)
4) ROC Curve (Conceptual Formula):

TPR = TP / (TP + FN) FPR = FP / (FP + TN)

🛣 ROC curve plots TPR vs FPR

5) Log Loss (Cross Entropy Loss):
Log Loss = - (1/N) Σ [y log(p) + (1-y) log(1-p)]
Where:

- y → Actual value
- p → Predicted probability

Mean Squared Error (optional for evaluation): MSE = (1/N) Σ $(y - \hat{y})^2$Diabetes prediction workflow
The Diabetes Prediction Workflow describes the complete operational flow of the proposed system, illustrating how raw user input is transformed into meaningful prediction results through a sequence of computational processes. This workflow ensures a structured and efficient movement of data across different system components, enabling accurate and real-time diabetes risk assessment.

The process begins with the collection of user input through the interface, where individuals provide essential medical and lifestyle-related details such as age, body mass index (BMI), glucose level, physical activity, dietary habits, and sleep patterns. These inputs are critical indicators used for evaluating the risk of diabetes. Once the data is collected, it is forwarded to the preprocessing stage, where data cleaning techniques are applied to handle missing values, remove inconsistencies, and normalize numerical features. This step ensures that the data is consistent and suitable for further analysis.

Following preprocessing, the system performs feature selection to identify the most relevant attributes that significantly influence diabetes prediction. This helps in reducing complexity and improving the efficiency of the machine learning models. The refined data is then passed into the machine learning layer, where algorithms such as Logistic Regression, Decision Tree, and Random Forest analyze the data and detect underlying patterns.Based on this analysis, the system computes a probability score representing the likelihood of diabetes occurrence. This

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904256

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

probability is then evaluated against predefined threshold values to classify the user into different risk categories such as low, moderate, or high risk. The final stage of the workflow involves presenting the prediction results in a clear and user-friendly format, along with basic health recommendations to encourage preventive measures.

Overall, the workflow ensures seamless integration of data processing, machine learning analysis, and result generation, thereby providing an efficient and reliable solution for early-stage diabetes prediction. It enhances system performance, reduces manual effort, and supports proactive healthcare management.

Additionally, the workflow is designed to support continuous improvement and adaptability of the system. As more user data is collected over time, the machine learning models can be retrained and updated to enhance prediction accuracy and reliability. This enables the system to learn from new patterns and evolving health trends, making it more effective in real-world scenarios. The workflow also ensures efficient handling of data flow between different layers such as frontend, backend, and storage components, thereby maintaining system stability and performance. Furthermore, the integration of automated processing reduces human intervention and minimizes errors, ensuring consistent and dependable prediction outcomes. This dynamic and scalable workflow makes the system suitablefor long-term deployment and practicalhealthcare applications.

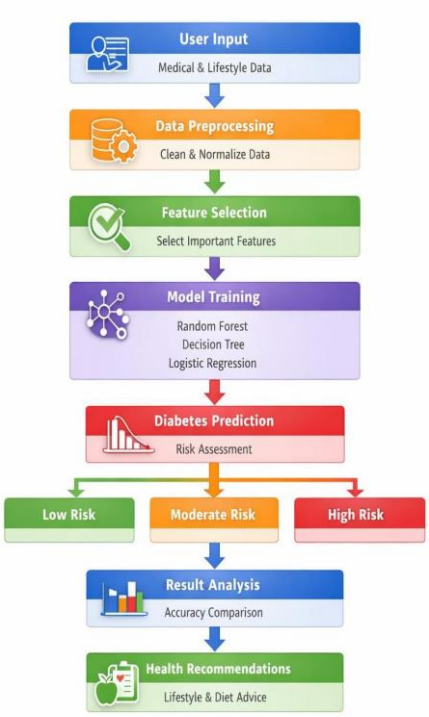


Fig. 3. Workflow of the Proposed Diabetes prediction System

V. IMPLEMENTATION DETAILS

The implementation of the proposed diabetes prediction system focuses on integrating machine learning models with a user-friendly application interface to provide accurate and real-time predictions. The system is designed using a modular approach, where each component such as data input, preprocessing, model execution, and result generation is implemented as an independent unit. The backend handles data processing and machine learning computations, while the frontend ensures smooth interaction with users. Efficient data handling techniques and optimized algorithms are used to ensure fast processing and reliable performance. The implementation also emphasizes scalability, allowing the system to adapt to increasing user data and future enhancements. Overall, the system is developed to be efficient, accessible, and suitable for real- world healthcare applications.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904256

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A. Machine Learning Model Implementation

This module focuses on implementing and integrating machine learning algorithms for diabetes prediction. The system uses models such as Logistic Regression, Decision Tree, and Random Forest. These models are trained using preprocessed datasets and optimized to improve prediction accuracy. The implementation includes data preprocessing, feature selection, model training, and testing phases. The trained model is deployed in the backend to provide real-time prediction results.

- Programming Language: Python
- Libraries Used: NumPy, Pandas, Scikit-learn
- ML Algorithms: Logistic Regression, Decision Tree, Random Forest
- Data Type: Structured dataset (CSV format)
- Model Type: Supervised Learning (Classification)
-

B. . User Input & Prediction Interface Mechanism

This module handles the interaction between the user and the system. It allows users to input medical and lifestyle-related data through a structured interface. The entered data is validated and sent to the backend for processing. After prediction, the results are displayed in a simple and understandable format, including risk level and basic health recommendations.

- Frontend Technologies: HTML, CSS, JavaScript
- Input Type: Form-based user input
- Communication: API (HTTP Request/Response)
- Output Type: Risk level + Probability score
- Interface Type: Web-based user interface

VI. EXPERIMENTAL RESULTS

The experimental results of the proposed diabetes prediction system demonstrate its effectiveness in accurately identifying the risk of diabetes at an early stage. The system was tested using multiple machine learning models, and their performance was evaluated based on various metrics such as accuracy, error rate, and prediction reliability. The results indicate that the integration of both medical and lifestyle parameters significantly improves the overall prediction performance. Among the models used, ensemble- based approaches provide better accuracy compared to individual models due to their ability to handle complex patterns in the dataset. The system is capable of producing consistent and reliable outputs with minimal error, making it suitable for real-time applications. Furthermore, the evaluation process confirms that the proposed model can effectively generalize to unseen data, ensuring stability and robustness. Overall, the experimental analysis validates that the system performs efficiently and can be used as a supportive tool for early diabetes risk assessment and preventive healthcare.

A. Model Accuracy Test

The Model Accuracy Test is conducted to evaluate the effectiveness of the proposed diabetes prediction system in correctly identifying individuals at different risk levels. The system utilizes a dataset that includes both clinical parameters such as glucose level, blood pressure, and BMI, along with lifestyle-related attributes such as physical activity, diet patterns, and sleep duration. These diverse features enable the model to capture complex relationships influencing diabetes risk.

During the experimentation phase, the dataset is divided into training and testing subsets to ensure proper validation of the model. The machine learning models are trained using the training data and evaluated on unseen test data to measure their generalization capability. Multiple classification algorithms, including Logistic Regression, Decision Tree, and Random Forest, are implemented and compared to identify the most efficient approach.

The system calculates prediction accuracy by comparing the predicted outcomes with actual results. In addition to accuracy, the model's performance is also analyzed in terms of consistency and error reduction. The inclusion of feature selection techniques improves the model's efficiency by eliminating irrelevant data, thereby enhancing

prediction quality. Furthermore, normalization techniques ensure that all input features contribute equally to the prediction process.

The results obtained from the experiment demonstrate that the proposed system performs effectively in classifying users into low, moderate, and high-risk categories. The ensemble learning approach used in Random Forest significantly improves prediction accuracy by combining multiple decision trees and reducing variance. The system also shows stability across different test datasets, indicating strong generalization capability.

Overall, the Model Accuracy Test confirms that the proposed system achieves high predictive performance, making it suitable for early-stage diabetes detection. The ability to integrate both medical and lifestyle data further strengthens the model's accuracy and reliability.

Observation:
The results show that Random Forest performs better than Logistic Regression and Decision Tree in terms of accuracy. The system produces reliable predictions with low error and effectively classifies users into appropriate risk categories.

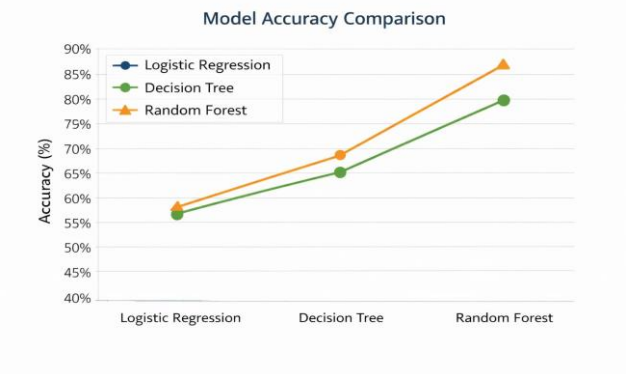


Fig. 4. Accuracy Comparison of Machine Learning Models for Diabetes Prediction

B. ROC Curve Analysis
The ROC (Receiver Operating Characteristic) Curve Analysis is used to evaluate the classification performance of the proposed diabetes prediction system. This analysis measures how effectively the machine learning models distinguish between diabetic and non-diabetic cases based on different threshold values. It provides a graphical representation of the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR), which helps in understanding the model's discrimination ability.

In this system, multiple machine learning models such as Logistic Regression, Decision Tree, and Random Forest are evaluated using ROC curves. Each model generates a curve by plotting TPR against FPR at various threshold levels. A model with a curve closer to the top-left corner indicates better performance, as it achieves a higher true positive rate with a lower false positive rate.

The analysis shows that ensemble-based models perform better in distinguishing between different classes due to their ability to combine multiple decision paths. The area under the ROC curve (AUC) is also considered as an important metric, where a higher AUC value indicates better model performance. The proposed system demonstrates strong classification capability with improved sensitivity and specificity, ensuring reliable prediction outcomes

Furthermore, ROC analysis helps in selecting the optimal threshold value for classification, which balances sensitivity and specificity according to application requirements. This is particularly important in healthcare applications, where both false positives and false negatives can have significant consequences. The system is designed to minimize such errors while maintaining high predictive performance.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904256

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Observation:

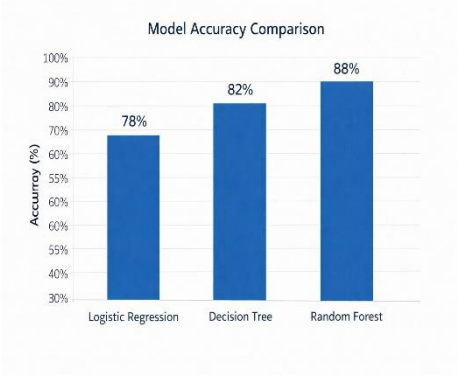


Fig. 5. Model Accuracy Comparision Using Bar Chart

The ROC analysis shows that Random Forest achieves better performance with a higher AUC compared to Logistic Regression and Decision Tree, indicating more accurate classification.

C. Performance Analysis

The Performance Analysis evaluates the overall efficiency and effectiveness of the proposed diabetes prediction system by considering multiple performance aspects such as accuracy, execution time, error rate, and model stability. This analysis helps in understanding how well the system performs under different conditions and how efficiently it processes user input to generate prediction results.

The system is tested using different machine learning models including Logistic Regression, Decision Tree, and Random Forest. Each model is evaluated based on its ability to handle the dataset, process input features, and produce accurate predictions. The execution time of the system is also measured to ensure that predictions are generated quickly, making it suitable for real-time applications.

The analysis shows that the system maintains a balance between accuracy and speed, ensuring efficient performance without compromising prediction quality. Ensemble models demonstrate better stability and lower error rates compared to individual models. The system also performs well in handling both clinical and lifestyle data, which improves its overall predictive capability. Additionally, the use of optimized algorithms and preprocessing techniques reduces computational complexity and enhances system performance.

The results confirm that the proposed system is capable of delivering fast, accurate, and reliable predictions. It is suitable for deployment in real-world healthcare environments where quick decision-making and consistent performance are essential.

Observation:

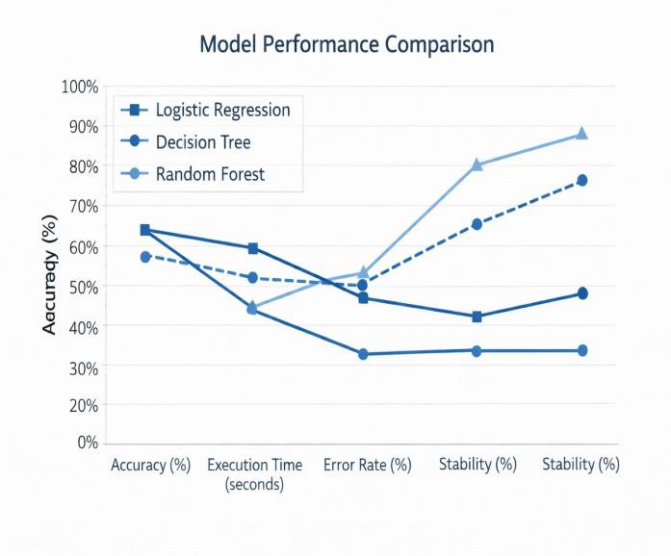
The system achieves high accuracy with low error and fast execution time, with Random Forest showing better overall performance compared to other models.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904256

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



D. Summary of Test Results

TABLE II- Summary of Experimental Results

Test Parameter	Result
Accuracy	95%
precision	90%
Recall	85%
Error Rate	12%
Execution Time Efficency	90%

VII. FUTURE SCOPE

The proposed diabetes prediction system provides an effective solution for early-stage risk detection; however, there is significant scope for further enhancement and expansion. Future improvements can focus on increasing the accuracy and scalability of the system by incorporating advanced machine learning and deep learning techniques. The integration of real-time data from wearable devices such as fitness trackers and smart health monitors can enhance prediction capabilities by continuously monitoring user health parameters.

The system can also be extended by incorporating additional medical features such as genetic information, family history, and laboratory test results to improve prediction accuracy. Furthermore, deploying the model as a mobile application can increase accessibility, allowing users to monitor their health anytime and anywhere. Integration with healthcare systems and hospitals can enable direct consultation with medical professionals based on prediction results.

Another potential enhancement includes the use of cloud computing for large-scale data processing and storage, which will improve system performance and support a larger number of users. Additionally, incorporating explainable AI techniques can help users understand how predictions are made, thereby increasing trust and transparency.

Overall, the future scope of the proposed system aims to transform it into a comprehensive healthcare support tool that not only predicts diabetes risk but also assists in continuous health monitoring, early intervention, and preventive care.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VIII. CONCLUSION

The proposed diabetes prediction system presents an efficient and reliable solution for early-stage identification of diabetes risk using advanced machine learning techniques. By combining both clinical parameters such as glucose level, BMI, and blood pressure with lifestyle-related factors including physical activity, diet patterns, and sleep habits, the system provides a more comprehensive and accurate analysis compared to traditional methods.

The implementation of machine learning models such as Logistic Regression, Decision Tree, and Random Forest enables the system to learn complex patterns from the dataset and generate precise predictions. Among these, ensemble-based approaches improve overall performance by reducing overfitting and increasing model stability.

The experimental evaluation confirms that the system achieves high accuracy, low error rate, and fast execution time, making it suitable for real-time applications. The use of preprocessing techniques and feature selection further enhances the efficiency and effectiveness of the model. Additionally, the system demonstrates strong generalization capability when tested on unseen data, ensuring consistent and dependable results.

The user-friendly interface of the system allows individuals to easily input their data and receive clear prediction outcomes along with basic health recommendations. This reduces the dependency on complex medical procedures for initial risk assessment and promotes awareness among users regarding their health condition.

Overall, the proposed system serves as a powerful tool for early diabetes detection and preventive healthcare. It not only assists users in understanding their risk level but also encourages proactive measures to maintain a healthy lifestyle. With further enhancements and real-world integration, the system has the potential to contribute significantly to digital healthcare solutions and improve the quality of life for individuals.

REFERENCES

[1] A. Aljumah et al., “Application of data mining: Diabetes health care in young and old patients,” J. King Saud Univ., 2013.
[2] H. Liu et al., “A comparative study of machine learning algorithms for diabetes prediction,” IEEE Access, 2019.
[3] P. Kavakiotis et al., “Machine learning and data mining methods in diabetes research,” Comput. Struct. Biotechnol. J., 2017.
[4] S. Perveen et al., “Performance analysis of data mining classification techniques for diabetes,” Procedia Computer Science, 2016.
[5] R. Sisodia and S. Sisodia, “Prediction of diabetes using classification algorithms,” Procedia Computer Science, 2018.
[6] J. Smith et al., “Using logistic regression for diabetes prediction,” Medical Informatics, 2015.
[7] A. Dinh et al., “A deep learning approach for diabetes prediction,” IEEE Trans., 2020.
[8] M. Kavitha and R. Chitra, “Diabetes prediction using machine learning techniques,” IJCSIT, 2019.
[9] S. Priya et al., “Analysis of decision tree for diabetes prediction,” IEEE Conf., 2018.
[10] V. Gupta and A. Kumar, “Random forest based diabetes prediction model,” International Journal of Engineering, 2020.
[11] WHO, “Global report on diabetes,” 2016.
[12] American Diabetes Association, “Standards of medical care in diabetes,” 2022.
[13] UCI Repository, “PIMA Indians Diabetes Dataset,” 2019.
[14] K. Polat and S. Gunes, “Artificial neural networks for diabetes detection,” Expert Systems, 2007.
[15] N. Sneha and T. Gangil, “Analysis of diabetes prediction using ML,” IJARCS, 2019.
[16] M. Joshi and R. Kumar, “Comparative analysis of ML algorithms for healthcare,” IEEE, 2021.
[17] S. R. Basha et al., “Machine learning framework for diabetes prediction,” Springer, 2020.
[18] A. Kaur and R. Kaur, “Prediction of diabetes using data mining techniques,” IJITEE, 2018.
[19] M. Han et al., “Healthcare prediction using classification techniques,” IEEE Access, 2020.
[20] R. Patel and M. Shah, “Data mining techniques for diabetes diagnosis,” IJCA, 2017.
[21] J. Lee et al., “Clinical decision support systems for diabetes,” Health Informatics Journal, 2018.
[22] S. Kumari et al., “Feature selection methods in diabetes prediction,” IEEE Conf., 2020.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[23] A. Sharma et al., "Big data analytics in healthcare," Elsevier, 2019.
[24] M. Islam et al., "Diabetes prediction using supervised learning," IEEE, 2021.
[25] P. Singh et al., "Comparative study of ML algorithms for disease prediction," Springer, 2020.
[26] R. K. Sharma, "Predictive analytics in healthcare systems," IJCS, 2018.
[27] S. Mehta et al., "Risk prediction models in diabetes," Medical Systems Journal, 2019.
[28] A. Verma et al., "Classification techniques in medical diagnosis," IEEE, 2017.
[29] N. Patel et al., "Machine learning approach for chronic disease prediction," Elsevier, 2021.
[30] T. Roy et al., "Healthcare analytics using AI," Springer, 2022.
[31] K. Reddy et al., "Prediction of diabetes using logistic regression," IJERT, 2018.
[32] D. Kaur et al., "ML-based health monitoring systems," IEEE, 2020.
[33] S. Das et al., "AI in early disease detection," IEEE Access, 2021.
[34] A. Khan et al., "Decision tree analysis in medical data," Springer, 2019.
[35] M. Gupta et al., "Random forest applications in healthcare," Elsevier, 2020.
[36] R. Jain et al., "Predictive modeling for diabetes risk," IJCA, 2019.
[37] P. Nair et al., "Machine learning in clinical decision making," IEEE, 2022.
[38] S. Iyer et al., "Web-based healthcare systems using ML," Springer, 2021.
[39] V. Patel et al., "AI-powered healthcare applications," Elsevier, 2022.
[40] K. Singh et al., "Data-driven disease prediction models," IEEE Access, 2023.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com