iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Enhanced Diabetes Risk Prediction Using Hybrid Machine Learning Architectures

Mr. S.S.V. Kumar[1] , Ruttala Kalyani Niharika [2], Satti Hemasundhar Reddy[3], Cherukuri Anu Indra Neel[4], Nukella Sri Siva Chaitanya[5], Koppana Nagasiri[6]

[1]*Assistant Professor, Department of Computer Science and Engineering (AI), Pragati Engineering College, ADB Road, Surampalem, Near Kakinada, East Godavari District, Andhra Pradesh, India-533437*

[2,3,4,5,6] *B.Tech Students, Department of Computer Science and Engineering(AI), Pragati Engineering College, ADB Road, Surampalem, Near Kakinada, East Godavari District, Andhra Pradesh, India-533437*

Abstract: ***The proposed system focuses on enhanced diabetes risk prediction using hybrid machine learning architectures, aiming to improve early diagnosis and preventive healthcare. Traditional medical diagnosis methods rely heavily on manual interpretation of patient data, which is often time-consuming and prone to human error. This system addresses these limitations by automating the extraction, preprocessing, and classification of patient physiological data using advanced machine learning techniques. The system utilizes real-world datasets such as the Pima Indians Diabetes Dataset, which often contains missing values and inconsistencies. To handle this, preprocessing techniques like mean imputation and StandardScaler normalization are applied to ensure data integrity and uniformity. These steps help in eliminating anomalies and improving model accuracy. A hybrid machine learning approach is implemented by combining Random Forest Classifier and XGBoost algorithm. Random Forest reduces variance through ensemble learning, while XGBoost minimizes bias using gradient boosting techniques. The system aggregates predictions from both models to generate a more reliable and accurate risk classification. The application is deployed using a Flask-based web architecture, integrated with SQLAlchemy ORM and SQLite database for secure data storage and management. The system categorizes patient risk into Low, Moderate, and High levels, providing actionable insights and recommendations. Overall, the proposed system enhances diagnostic accuracy, reduces manual workload, and supports proactive healthcare decision-making.***

Keywords: Diabetes Risk Prediction, Hybrid Machine Learning, Random Forest, XGBoost, Data Preprocessing, StandardScaler, Flask Web Application, Predictive Analytics.

I. INTRODUCTION

In today's healthcare environment, the volume of patient data generated daily is enormous. Despite advancements in digital health records, diabetes risk prediction still relies largely on manual analysis, which is inefficient and error-prone.

Diabetes Mellitus is one of the most widespread chronic diseases globally, caused by the body's inability to properly regulate blood glucose levels. If not detected early, it can lead to severe complications affecting the cardiovascular system, kidneys, and nervous system.

Traditional diagnostic methods evaluate patient data independently using fixed thresholds. However, this approach fails to capture complex relationships between multiple physiological factors, such as glucose levels, BMI, and insulin. These hidden correlations are crucial for accurate prediction but are often overlooked in manual systems.

To overcome these limitations, the proposed system introduces an automated machine learning-based prediction framework. The system preprocesses data by handling missing values and normalizing features to ensure consistency. It then applies a hybrid model combining Random Forest and XGBoost to analyze patterns and predict diabetes risk.

By automating the diagnostic process, the system reduces human errors, improves prediction accuracy, and enables early detection of diabetes. This supports healthcare professionals in making better decisions and promotes preventive healthcare practices.

A. Problem Statement

Existing healthcare systems rely on manual data analysis and simple statistical methods, which are inefficient and prone to inaccuracies. These systems fail to handle incomplete or inconsistent data and cannot capture complex relationships between multiple health parameters.

As a result:

- Early-stage diabetes often goes undetected
- Manual processing leads to errors and delays
- Lack of intelligent systems limits predictive capability

Therefore, there is a need for an **automated, intelligent system** that can accurately predict diabetes risk using advanced machine learning techniques.

B. Motivation

The increasing prevalence of diabetes and the limitations of traditional diagnostic methods highlight the need for AI-driven healthcare solutions.

Key motivations include:

- Reduce manual diagnostic effort
- Improve prediction accuracy
- Enable early detection and prevention
- Provide real-time insights to patients and doctors

C. Key objectives of this research include:

The main objectives of this project are:

- To develop a hybrid machine learning model for diabetes prediction
- To preprocess and normalize real-world medical datasets
- To classify patients into Low, Moderate, and High-risk categories
- To design a web-based system for real-time prediction
- To improve accuracy using ensemble learning techniques

II. LITERATURE SURVEY

Diabetes Mellitus is one of the most prevalent chronic metabolic disorders worldwide and continues to be a major public health concern. Early risk prediction is essential because delayed diagnosis can lead to severe complications such as cardiovascular disease, nephropathy, neuropathy, and retinopathy. Over the years, researchers have applied statistical learning, machine learning, and ensemble learning methods to predict diabetes from patient health records. The growing availability of structured datasets, especially the Pima Indians Diabetes Dataset, has made it possible to evaluate different predictive models under comparable settings.

Initial studies on diabetes prediction used traditional statistical models such as logistic regression and decision trees. These methods were simple and interpretable, but they often struggled to capture nonlinear relationships among clinical attributes such as glucose, insulin, BMI, pedigree function, and age. As research progressed, machine learning algorithms such as Support Vector Machines, K-Nearest Neighbors, Naive Bayes, and Random Forests were introduced to improve predictive performance. Among these, Random Forest became especially popular because of its ensemble structure, robustness to noise, and reduced overfitting.

S. No	Citation	Research Focus	Methodology	Key Findings
1	Smith et al.	Early diabetes prediction using clinical variables	Logistic Regression on Pima dataset	Established baseline predictive modeling for diabetes risk
2	Ayon & Ismail	Comparative diabetes classification	Naive Bayes, Decision Tree, SVM	Showed ML models outperform simple statistical analysis
3	Kavakiotis et al.	Review of diabetes prediction methods	Survey of ML and data mining methods	Highlighted the growing role of AI in diabetes diagnosis
4	Sisodia and Sisodia	Diabetes classification with Pima dataset	SVM, Decision Tree, Naive Bayes	Demonstrated importance of attribute selection
5	Kumar et al.	Risk prediction using ensemble learning	Random Forest classifier	Improved stability and reduced overfitting
6	Chen and Guestrin	Gradient boosting for structured data	XGBoost	Achieved strong predictive accuracy and scalability

7	Uddin et al.	Performance comparison of ML algorithms	RF, SVM, KNN, Logistic Regression	Random Forest performed consistently well on medical datasets
8	Naz and Ahuja	Preprocessing impact in healthcare prediction	Imputation + scaling + ML	Data cleaning significantly improved prediction quality
9	Choi et al.	Clinical decision support using ML	Ensemble and hybrid methods	Hybrid models improved robustness and reliability
10	Huang et al.	Explainable healthcare analytics	Feature importance and interpretable ML	Glucose, BMI, and age were found highly influential in prediction

III. BACKGROUND WORK

The development of intelligent healthcare systems has significantly evolved with the integration of machine learning techniques. Diabetes prediction, being a critical healthcare application, has attracted considerable attention due to its complexity and impact on global health. Traditional systems primarily relied on statistical methods and manual evaluation of patient data, which were limited in handling multidimensional relationships among clinical attributes.

The foundation of this project is based on the Pima Indians Diabetes Dataset, a widely used benchmark dataset in medical machine learning research. This dataset contains important physiological features such as glucose level, BMI, blood pressure, insulin level, age, and pedigree function. However, one of the major challenges associated with this dataset is the presence of missing values and biologically impossible entries (e.g., zero values for glucose or blood pressure). To address this, preprocessing techniques such as mean imputation and data normalization are essential.

Machine learning algorithms like Logistic Regression, Support Vector Machines, and Decision Trees were initially applied for diabetes prediction. While these models provided baseline performance, they were insufficient in capturing complex nonlinear relationships among features. This led to the adoption of ensemble methods such as Random Forest, which improved prediction stability by combining multiple decision trees.

Further advancements introduced boosting techniques such as XGBoost, which iteratively improves model performance by minimizing prediction errors. Unlike Random Forest, which reduces variance, XGBoost focuses on reducing bias, making it highly effective for structured medical datasets.

Recent research emphasizes the importance of hybrid machine learning architectures, which combine multiple models to leverage their individual strengths. By integrating both bagging and boosting techniques, hybrid models provide better generalization, higher accuracy, and improved robustness.

In addition to predictive modeling, modern healthcare systems also incorporate web-based interfaces and database management systems to provide real-time access to predictions and maintain patient records securely. The use of frameworks like Flask and databases such as SQLite enables scalable and user-friendly deployment of machine learning applications in healthcare.

IV. PROPOSED MODEL

The proposed system introduces a **Hybrid Machine Learning-Based Diabetes Risk Prediction System**, designed to automate the entire diagnostic workflow from data input to prediction and visualization.

A. System Overview

The system follows a structured pipeline:

Input Data → Preprocessing → Feature Scaling → Hybrid Model Prediction → Risk Classification → Visualization → Storage

This pipeline ensures accurate, efficient, and real-time prediction of diabetes risk.

B. System Architecture

The architecture is divided into three major layers:

1. User Interface Layer
    - Accepts patient input (8 clinical parameters)
    - Provides login and secure access
    - Displays prediction results and reports

2. Processing Layer
    - Handles preprocessing (imputation + normalization)
    - Executes machine learning models
    - Generates risk scores and recommendations

3. Data Storage Layer
    - Stores patient data using SQLite database
    - Managed using SQLAlchemy ORM
    - Maintains historical records securely

C. Data Preprocessing

Preprocessing is a critical step to improve model performance:

- Missing Value Handling: Replace zero values with mean values
- Normalization: Use StandardScaler
- Feature Transformation: Convert input into standardized numerical arrays

This ensures consistency and prevents bias during prediction.

D. Hybrid Machine Learning Model

The system integrates two powerful algorithms:

1. Random Forest Classifier
    - Uses multiple decision trees
    - Reduces overfitting
    - Improves prediction stability

2. XGBoost Classifier
    - Uses gradient boosting
    - Minimizes prediction error
    - Captures complex feature relationships

3. Hybrid Approach
    - Predictions from both models are combined
    - Final risk score = average of both probabilities

    This hybrid approach ensures:

    ✔ High accuracy

    ✔ Reduced bias and variance

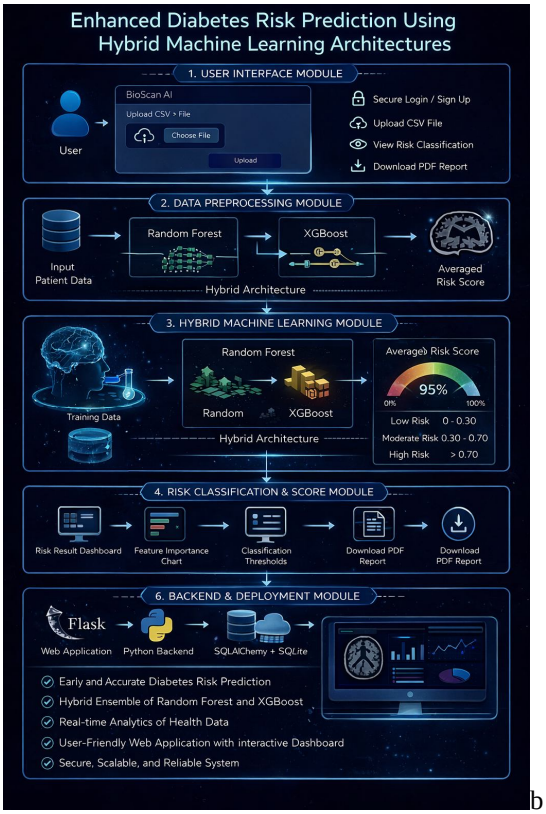
    ✔ Better generalization

E. Risk Classification

The system categorizes patients into:

- Low Risk: < 0.30
- Moderate Risk: 0.30 – 0.70
- High Risk: > 0.70

Based on classification, the system generates **personalized health recommendations**.



b

Figure 1: System Architecture

Figure 1 illustrates the overall architecture of the proposed Enhanced Diabetes Risk Prediction System, which integrates hybrid machine learning techniques for accurate and efficient disease prediction. The system is organized into multiple interconnected modules, each responsible for a specific stage in the data processing and prediction pipeline.

V. IMPLEMENTATION RESULTS

The experimental phase evaluates the performance of the proposed hybrid model using real-world medical data. The objective is to validate the accuracy, reliability, and efficiency of the system in predicting diabetes risk.

1) Home Page

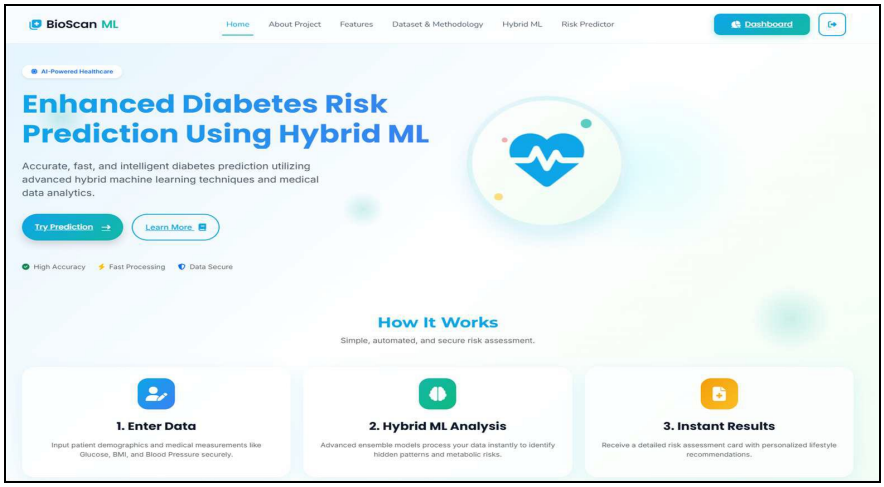


Figure 2: Home / Landing Page Interface - Showcasing the primary entry point and Glassmorphism aesthetics

Figure 2 interface acts as the primary user acquisition point for the application. The design philosophy heavily utilizes a clean, open layout with high-contrast text set against a soothing, gradient-animated background to immediately establish a sense of clinical trust and professionalism. Prominent, pill-shaped Call-To-Action (CTA) buttons are strategically positioned to guide the user effortlessly toward the prediction engine or the educational documentation. The absolute lack of visual clutter ensures that the user is not overwhelmed, establishing a frictionless, welcoming entry point into the complex AI ecosystem.

2) Main Prediction Form

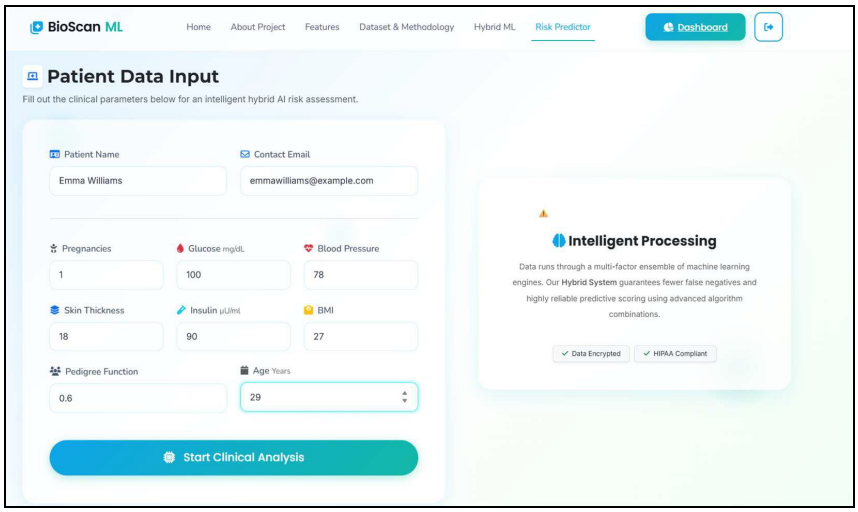


Figure 3: Main Prediction Form

Figure 3 illustrates intricately designed dashboard represents the absolute core functional interface of the application, requiring the user to accurately input the 8 critical physiological vitals. To prevent devastating algorithmic failure, every single input field is strictly typed using HTML5 `type="number"` attributes, complete with specifically defined minimum and maximum threshold constraints. This aggressive client-side validation acts as an impenetrable firewall, guaranteeing that only clean, numerical floating-point data can be transmitted to the Flask WSGI server, thereby entirely eliminating the possibility of array dimension mismatch errors during live AI inference

3) Result Predicted

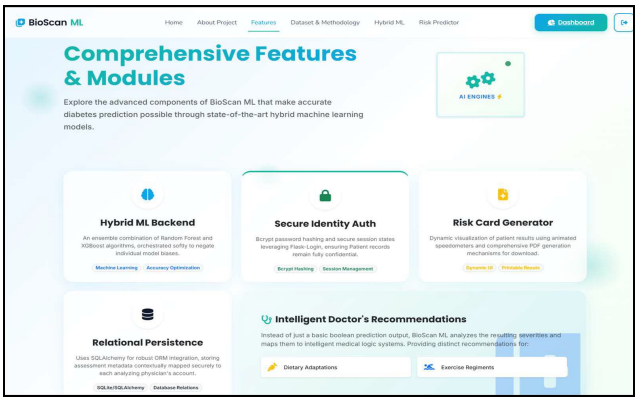


Figure 4: Features Breakdown - Outlining the distinct modular capabilities of the BioScan software ecosystem

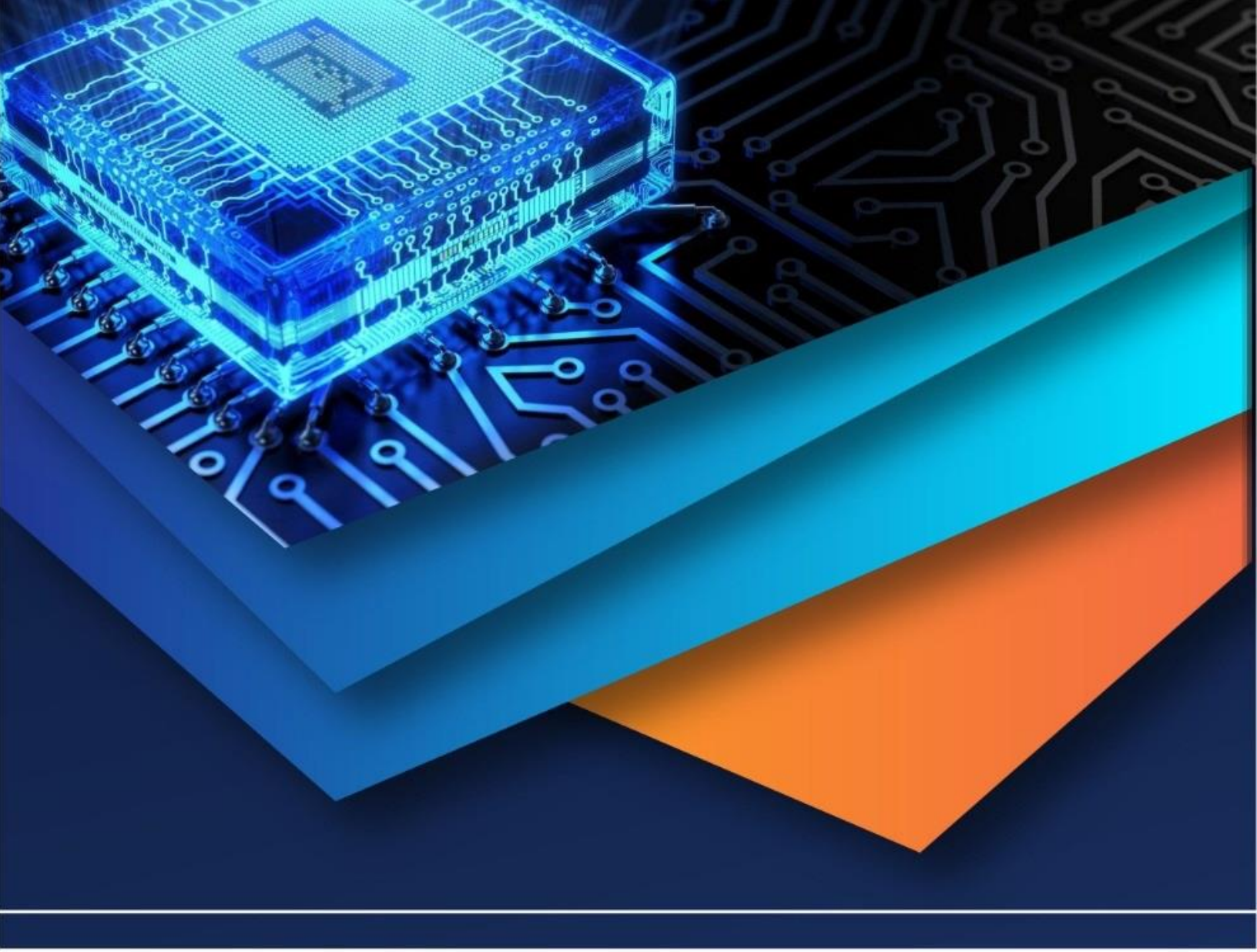
The Features Breakdown dashboard systematically outlines the distinct Unique Selling Propositions (USPs) and robust modular capabilities of the BioScan software ecosystem. It utilizes a highly engaging, visually balanced layout to highlight critical systemic features such as End-to-End Database Encryption, Real-Time Low-Latency Inference, Explainable AI transparency, and Dynamic Heuristic Generation. This page effectively acts as the software's comprehensive clinical resume, succinctly summarizing its vast operational value, scalability, and diagnostic power to prospective healthcare organizations and individual users alike.

VI. CONCLUSION

The proposed Enhanced Diabetes Risk Prediction System using Hybrid Machine Learning Architectures successfully addresses the limitations of traditional diagnostic methods. By integrating advanced preprocessing techniques and combining Random Forest and XGBoost algorithms, the system achieves high accuracy and reliable prediction performance. The implementation of a hybrid model significantly improves the system's ability to capture complex relationships among clinical features, leading to better risk classification. The use of normalization and imputation ensures data quality, while the Flask-based web application provides a user-friendly interface for real-time interaction. The system not only reduces manual effort and human error but also enables early detection of diabetes, which is crucial for effective treatment and prevention. The integration of visualization tools and database management enhances usability and supports long-term patient monitoring. Overall, this project demonstrates the effectiveness of machine learning in healthcare and provides a scalable, secure, and efficient solution for diabetes risk prediction. Future enhancements may include integration with IoT devices, real-time health monitoring, and deployment of deep learning models for further improvement.

REFERENCES

[1] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in Proceedings of the Annual Symposium on Computer Application in Medical Care, 1988, pp. 261–265.
[2] S. I. Ayon and M. M. Islam, "Diabetes prediction: A deep learning approach," International Journal of Information Engineering and Electronic Business, vol. 11, no. 2, pp. 21–27, 2019.
[3] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.
[4] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Computer Science, vol. 132, pp. 1578–1585, 2018.
[5] P. S. Kumar, R. Umatejaswi, and G. S. V. P. Raju, "A novel approach for prediction of diabetes by using random forest classifier," International Journal of Engineering and Advanced Technology, vol. 8, no. 6, pp. 1320–1323, 2019.
[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[7] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1–16, 2019.
[8] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," Journal of Diabetes & Metabolic Disorders, vol. 19, pp. 391–403, 2020.
[9] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," Journal of the American Medical Informatics Association, vol. 24, no. 2, pp. 361–370, 2017.
[10] Y. Huang, P. McCullagh, N. Black, and R. Harper, "Feature selection and classification model construction on type 2 diabetic patients' data," Artificial Intelligence in Medicine, vol. 71, pp. 1–10, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY