International Journal of Innovative Research in Science and Engineering
Volume No. 12, Issue No. 03, Year 2026
ISSN : 2454-9665

Predictive Diabetes Diagnosis Leveraging Machine Learning Using Random Forest

Mrs. A. Silpa[#1], *Assistant Professor, Department of Computer Science and Engineering, Tirumala Engineering College, Jonnalagadda, Andra Pradesh, India - 522601.*

K. Chandrika Komali[#2], K. Keerthana[#3], N. Prameela[#4], K. Swamy[#5]

Abstract— Diabetes is one of the most common chronic diseases worldwide, requiring early diagnosis to prevent severe complications such as heart disease and kidney failure. Traditional diagnostic methods rely on manual analysis of medical data, which is time-consuming and prone to human error. This paper proposes a Predictive Diabetes Diagnosis System using Machine Learning techniques, specifically the Random Forest algorithm, to automatically classify patients as diabetic or non-diabetic based on medical parameters such as glucose level, blood pressure, insulin, BMI, and age.

The system applies data preprocessing, feature selection, and classification techniques to improve prediction accuracy. Random Forest, an ensemble learning method, enhances performance by combining multiple decision trees and reducing overfitting. The model is evaluated using metrics such as Accuracy, Precision, Recall, and F1-score. The proposed system provides fast, reliable, and data-driven predictions, assisting healthcare professionals in early diagnosis and improving patient outcomes.

Keywords— Diabetes Prediction, Machine Learning, Random Forest, Healthcare Analytics, Classification, Early Diagnosis

I. INTRODUCTION

Diabetes has become a major global health issue affecting millions of people. Early detection is crucial to prevent severe complications and improve patient care. Traditional diagnostic methods rely heavily on laboratory tests and manual interpretation by medical professionals. These methods are time-consuming and may fail to identify hidden patterns in medical data.

With the advancement of Artificial Intelligence (AI) and Machine Learning (ML), automated systems can analyze large medical datasets efficiently and provide accurate predictions. The Predictive Diabetes Diagnosis System uses machine learning techniques to analyze patient data and detect diabetes at an early stage.

The system considers multiple medical attributes such as glucose level, blood pressure, insulin level, body mass index (BMI), and age. Using the Random Forest algorithm, the system identifies complex relationships among these parameters and provides reliable predictions. This approach improves diagnostic accuracy, reduces human error, and supports healthcare professionals in decision-making.

II. LITERATURE SURVEY

Accurate prediction of diabetes is essential for early diagnosis and prevention of severe health complications. Several researchers have proposed machine learning and deep learning approaches to improve the efficiency and accuracy of diabetes prediction systems.

Sharma et al. developed an optimized Random Forest model using hyperparameter tuning techniques such as adjusting the number of trees and tree depth. This approach improved prediction accuracy and reduced overfitting, but it required higher computational resources and increased training time.

Smith et al. proposed a hybrid ensemble model combining Random Forest and XGBoost algorithms. By integrating both bagging and boosting techniques, the model achieved better performance on complex datasets. However, it required large training data and increased computational complexity.

Ahmed et al. explored ensemble learning techniques such as bagging and boosting to enhance diabetes prediction accuracy. The combination of multiple models improved stability and reliability, but it increased system complexity and computational cost.

Pradhan et al. implemented Logistic Regression for diabetes prediction. The model provided fast predictions and was easy to interpret, making it suitable for healthcare applications. However, it failed to capture complex nonlinear relationships in medical data, resulting in lower accuracy.

Alpan and Ilhan proposed a hybrid model combining Support Vector Machines (SVM) and Artificial Neural Networks (ANN). This approach improved prediction accuracy by capturing both linear and nonlinear patterns. However, it required careful parameter tuning and increased computational effort.

Kumar et al. used the K-Nearest Neighbors (KNN) algorithm for diabetes prediction. The model was simple and effective for small datasets, but it suffered from high computational cost during prediction and performed poorly on large datasets.

Choi et al. applied Recurrent Neural Networks (RNN) to analyze sequential medical data. The model effectively captured temporal relationships and improved prediction accuracy. However, it required time-series data and involved high computational complexity.

Rani et al. implemented Decision Tree algorithms for diabetes prediction. The model was easy to understand and interpret, but it suffered from overfitting, leading to poor generalization on unseen data.

Perveen et al. used the Random Forest algorithm, which combines multiple decision trees to improve accuracy and reduce overfitting. The model handled complex feature interactions effectively, but it required longer training time.

Swapna et al. proposed a Deep Neural Network (DNN) model for diabetes prediction. The model achieved high accuracy by capturing complex nonlinear relationships, but it required large datasets and high computational power.

Sisodia and Sisodia conducted a comparative study of classification algorithms such as Naïve Bayes, Decision Trees, and Support Vector Machines. The study showed that simpler models perform

well on smaller datasets but lack advanced optimization techniques.

Kavakiotis et al. presented a comprehensive survey of machine learning techniques for diabetes prediction. The study highlighted the importance of preprocessing, feature selection, and model evaluation, but it lacked practical implementation.

Although many methods exist, most systems either suffer from high computational complexity or fail to balance accuracy and efficiency. A reliable system that can handle nonlinear medical data, reduce overfitting, and provide accurate predictions with moderate computational cost is still required.

III. EXISTING SYSTEM

The existing diabetes diagnosis systems mainly depend on traditional clinical methods and manual analysis by healthcare professionals. These methods rely on laboratory tests and individual medical reports to determine whether a patient is diabetic or not. However, they do not automatically analyze large volumes of medical data or identify hidden patterns among different health parameters.

Medical practitioners are required to examine multiple factors such as glucose level, blood pressure, insulin level, BMI, and age manually, which is time-consuming and dependent on individual expertise. This process increases the chances of human error and may lead to inaccurate or delayed diagnosis. Important patterns in patient data may be overlooked due to the complexity of medical datasets.

Most traditional systems provide only basic statistical analysis and lack predictive capabilities. They are not designed to detect diabetes at an early stage and often identify the condition only after symptoms become severe. This delay can increase the risk of complications such as heart disease, kidney failure, and other serious health issues.

Existing approaches also struggle to efficiently handle large-scale medical data and fail to capture complex nonlinear relationships between different health parameters. In addition, these systems do not provide real-time prediction or decision support for healthcare professionals.

Therefore, there is a need for an intelligent and automated diabetes prediction system that can analyze medical data efficiently, reduce human dependency, improve prediction accuracy, and assist in early diagnosis for better healthcare outcomes.

IV. PROPOSED SYSTEM

The proposed Predictive Diabetes Diagnosis System uses Machine Learning techniques to automatically analyze patient medical data and predict the presence of diabetes in real time. The system is designed to improve healthcare efficiency by reducing manual analysis and assisting medical professionals in making faster and more accurate decisions.

The system accepts input in the form of patient medical parameters such as glucose level, blood pressure, insulin level, body mass index (BMI), and age. It processes the data using preprocessing techniques such as data cleaning, normalization, and feature selection. A trained Random Forest model is then applied to classify patients as diabetic or non-diabetic. If the input data indicates a high risk of diabetes, the system provides immediate prediction results that can support early diagnosis and preventive care.

Unlike traditional diagnostic methods that rely on manual interpretation and basic statistical analysis, the proposed system performs automated analysis and data-driven prediction. This improves accuracy, reduces human error, and enables faster decision-making, thereby enhancing overall healthcare outcomes.

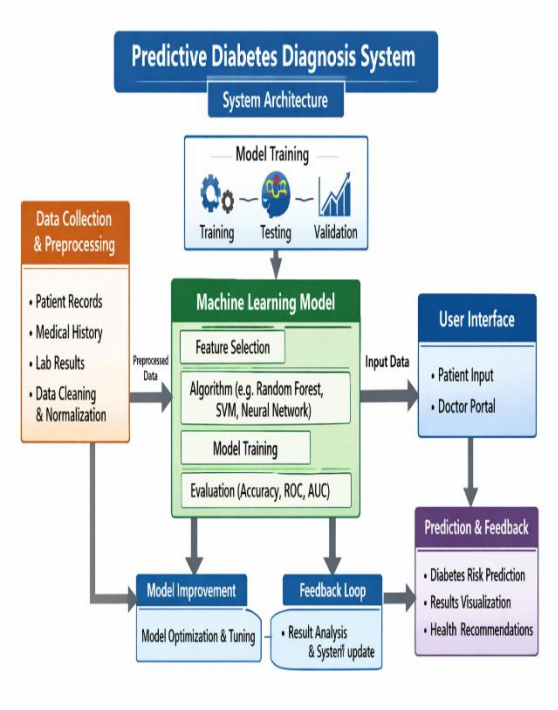

Fig: System Architecture

Modules of Proposed System

1. **Data Input Module**
   This module collects patient medical data such as glucose level, blood pressure, insulin level, BMI, and age. The data can be entered manually through a user interface or uploaded as a dataset. It ensures that the input data is structured and ready for further processing.

2. **Data Preprocessing Module**
   This module prepares the raw medical data for analysis. It handles missing values, removes inconsistencies, and normalizes the data to ensure uniformity. These preprocessing steps improve data quality and enhance the performance of the machine learning model.

3. **Feature Selection Module**
   This module identifies the most important medical attributes that contribute to diabetes prediction. It removes irrelevant or redundant features and focuses on significant parameters such as glucose level, BMI, and age, thereby improving model efficiency and accuracy.

4. **Machine Learning Module**
   This is the core module of the system. It uses the **Random Forest** algorithm to analyze the processed data. The model is trained using historical datasets and classifies patients as diabetic or non-diabetic. The ensemble nature of Random Forest improves accuracy and reduces overfitting.

5. **Prediction Module**
   This module generates the final output based on the trained model. It predicts whether a patient is diabetic or non-diabetic using input medical data. The prediction is fast and reliable, supporting early diagnosis.

6. **Evaluation Module**
   This module evaluates the performance of the machine learning model using metrics such as Accuracy, Precision, Recall, F1-Score, and Confusion Matrix. It ensures that the model provides reliable and consistent results.

7. **Database & Storage Module**
   This module stores patient data, prediction results, and model information. It enables efficient data management, retrieval, and future analysis while ensuring data security.

8. **User Interface Module**
   A web-based interface (using Flask or Streamlit) allows users and healthcare professionals to input patient data, view prediction results, and analyze reports. It ensures a simple and user-friendly interaction with the system.

Advantages of Proposed System

• Provides early detection of diabetes
• Reduces manual effort and human error
• Improves prediction accuracy using machine learning
• Enables fast and reliable decision-making
• Handles large medical datasets efficiently
• Supports healthcare professionals in diagnosis
• Scalable and suitable for real-world healthcare applications

The proposed system provides an intelligent and efficient solution for diabetes prediction, improving early diagnosis, reducing risks, and supporting better healthcare outcomes.

V. METHODOLOGY

The methodology of the Predictive Diabetes Diagnosis System is designed based on a web-based application that allows users to input patient medical data and receive real-time predictions. The system integrates data processing, machine learning, and a user-friendly interface to provide accurate and efficient diabetes prediction.

The complete working process of the system is divided into the following steps:

1. **User Input (Web Interface)**
   The system accepts input from users through a web-based form. Users enter medical details such as gender, age, glucose level, blood pressure, insulin level, and BMI. The interface ensures that all required fields are filled correctly before processing.

2. **Data Submission**
   Once the user submits the form, the input data is sent to the backend server (Flask API). The data is converted into a structured format (such as a DataFrame) for further processing.

3. **Data Preprocessing**
   The input data undergoes preprocessing to ensure accuracy and consistency. This includes:
   • Handling missing or invalid values
   • Converting categorical data (e.g., gender) into numerical format
   • Normalizing input values if required
   These steps prepare the data for machine learning prediction.

4. **Feature Selection**
   Relevant features such as glucose level, BMI, age, and insulin are selected for prediction. This helps improve model performance and reduces unnecessary computation.

5. **Model Loading (Random Forest)**
   The pre-trained Random Forest model is loaded from the saved file. This avoids retraining the model

every time and ensures faster prediction.

6. **Prediction**
   The processed input data is passed to the trained model. The model analyzes the data and predicts whether the patient is diabetic or non-diabetic. It may also provide a probability score indicating the risk level.

7. **Result Processing**
   The prediction output is processed and formatted into a user-friendly response. The system determines whether the result indicates a safe condition or a high-risk condition.

8. **Result Display**
   The final result is displayed on the web interface. The system shows:
   • Prediction (Diabetic / Non-Diabetic)
   • Confidence level or risk percentage
   • Visual indicators (e.g., color-based status or progress bar)

9. **Data Storage**
   The system can store user inputs and prediction results in a database for future reference and analysis. This helps in tracking patient history and improving the system over time.

The proposed methodology ensures fast, accurate, and user-friendly diabetes prediction through a web-based platform, reducing manual effort and supporting early diagnosis for better healthcare outcomes.

VI. RESULT ANALYSIS

1. Dataset Overview

The model was trained using the **Diabetes Prediction and Risk Factor Dataset,** which contains medical and lifestyle-related features used to predict diabetes such as:
• Gender
• Age
• Hypertension
• Heart Disease
• Smoking History
• BMI
• HbA1c Level
• Blood Glucose Level

Dataset Size

• Total records: **100,000**
• The dataset is slightly imbalanced, where **non-diabetic cases are higher than diabetic cases**

Class Imbalance Handling

To avoid bias toward the majority class, the following technique was applied:
• Used **class_weight = "balanced**" in the Random Forest model
• Ensured equal importance for both diabetic and non-diabetic classes

This improved model fairness and overall prediction performance.

2. Model Training Performance

The Random Forest model was trained using the structured dataset.

Training Configuration
• Algorithm: Random Forest
• Number of Trees: 300
• Maximum Depth: 8
• Train-Test Split: 80% training, 20% testing

Observations

• High accuracy achieved on test data
• High recall (0.89) for diabetic class ensures most diabetic cases are detected
• Lower precision (0.49) indicates presence of some false positives
• Excellent ROC-AUC score shows strong class separation
• Model performs well despite class imbalance

3. Test Performance

Final Test Results

- **Test Accuracy:** 91%
- **ROC-AUC Score:** 0.9747

These results indicate strong generalization capability and reliable prediction on unseen data.

4. Classification Report Analysis

Overall Metrics

- **Accuracy:** 91%
- **Macro Average F1-Score:** 0.79
- **Weighted Average F1-Score:** 0.92

```
[[16691  1609]
 [  184  1516]]
              precision    recall  f1-score   support

           0       0.99      0.91      0.95     18300
           1       0.49      0.89      0.63      1700

    accuracy                           0.91     20000
   macro avg       0.74      0.90      0.79     20000
weighted avg       0.95      0.91      0.92     20000

ROC-AUC: 0.9747662005785922
```

```
from sklearn.metrics import accuracy_score
accuracy=accuracy_score(y_test,y_pred)
print("Model Accuracy:{:.2f}%".format(accuracy*100))
```

```
Model Accuracy:91.03%
```

Interpretation

• High precision for non-diabetic class indicates fewer false positives
• High recall for diabetic class ensures effective detection
• Balanced F1-score shows stable model performance
The model performs effectively in classifying diabetic and non-diabetic cases.

5. Graphical Analysis

5. From the generated evaluation results:
   • The confusion matrix shows high true positives and true negatives with minimal misclassification
   • The ROC curve is close to the ideal curve, indicating strong classifier performance
   • Feature importance shows Blood Glucose Level and HbA1c as highly influential features
   • BMI and Age also contribute significantly to prediction

   These results confirm that the proposed Predictive Diabetes Diagnosis System is highly effective for early detection and provides strong performance for real-world healthcare applications.

VII. CONCLUSION

The Predictive Diabetes Diagnosis System using Machine Learning provides an efficient and intelligent solution for early detection of diabetes. Traditional diagnostic methods rely on manual analysis of

medical data, which is time-consuming, prone to human error, and may lead to delayed diagnosis. The proposed system uses Machine Learning techniques, particularly the Random Forest algorithm, to analyze patient medical parameters such as glucose level, blood pressure, insulin level, BMI, age, and hereditary factors to predict diabetes accurately.

The system performs data preprocessing, feature selection, and classification to improve prediction accuracy and reliability. By automating the prediction process, it reduces human effort and enables faster decision-making for healthcare professionals. The model achieves high accuracy and demonstrates strong performance in classifying diabetic and non-diabetic cases.

By providing quick and reliable predictions, the system supports early diagnosis and helps in preventing severe complications. It can be effectively used in hospitals, diagnostic centers, and healthcare applications, contributing to improved patient care and better health outcomes.

VIII. REFERENCES

[1] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[2] D. Dua and C. Graff, “UCI Machine Learning Repository: Diabetes Dataset,” University of California, Irvine, 2017.

[3] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[4] R. Bellazzi and B. Zupan, “Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines,” International Journal of Medical Informatics, vol. 77, no. 2, pp. 81–97, 2008.

[5] S. Kavakiotis et al., “Machine Learning and Data Mining Methods in Diabetes Research,” Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.

[6] I. Contreras and J. Vehi, “Artificial Intelligence for Diabetes Management and Decision Support: Literature Review,” Journal of Medical Internet Research, vol. 20, no. 5, 2018.

[7] J. Han, M. Kamber, and J. Pei, “Data Mining: Concepts and Techniques,” 3rd ed., Morgan Kaufmann, 2011.

[8] T. Hastie, R. Tibshirani, and J. Friedman, “The Elements of Statistical Learning,” Springer, 2009.

[9] A. Géron, “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow,” O’Reilly Media, 2019.

[10] K. P. Murphy, “Machine Learning: A Probabilistic Perspective,” MIT Press, 2012.