A PREDICTIVE MODEL FOR DIABETES DETECTION USING RANDOM FOREST AND LOGISTIC REGRESSION

[1]Sachin Chawhan, [2] Kamal Suthar, [3] Sowmya, [4] Keerthan
[1]AssistantProfessor, [234]Students
Department of Computer Science & Engineering
Siddhartha Institute of Technology & Sciences, Narapally
sachinchawhan_cse@siddhartha.co.in, 23TQ1A0522@siddhartha.co.in,
23TQ1A0511@siddhartha.co.in, 23TQ1A0510@siddhartha.co.in,

Abstract

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from inadequate insulin production or ineffective utilization of insulin by the body. It is one of the most prevalent and rapidly increasing health issues worldwide. Early detection and proper management of diabetes are crucial to prevent serious complications such as cardiovascular diseases, kidney failure, nerve damage, and vision impairment. Traditional diagnostic approaches primarily rely on laboratory tests and clinical evaluations, which may delay early identification of the disease.

With the rapid growth of data analytics and artificial intelligence, machine learning techniques have become powerful tools for disease prediction using historical medical data. This project focuses on developing an efficient predictive model for diabetes detection using Random Forest and Logistic Regression algorithms. The system aims to analyze various patient health parameters and accurately classify individuals as diabetic or non-diabetic.

The model is trained using a healthcare dataset that includes key medical attributes such as number of pregnancies, glucose level, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. These features play a significant role in assessing the risk of diabetes. The proposed system enhances early diagnosis, supports medical decision-making, and contributes to improved healthcare outcomes.

I. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels due to insufficient insulin production or the body's inability to effectively utilize insulin. It is one of the most widespread health problems worldwide and poses a significant threat to public health. If not diagnosed and managed at an early stage, diabetes can lead to severe complications such as cardiovascular diseases, kidney failure, blindness, and nerve damage. According to global health reports, the number of people affected by diabetes continues to rise rapidly, making early detection and prevention extremely important.

Traditional methods of diagnosing diabetes mainly depend on laboratory tests and clinical expertise. While these methods are reliable, they can be time-consuming, costly, and may not always provide early identification of individuals at high risk. As a result, there is a growing need for more efficient and intelligent systems that can assist in early prediction and diagnosis.

With the advancement of Artificial Intelligence (AI) and Machine Learning (ML), predictive analytics has emerged as a powerful approach in healthcare. Machine learning models can process large volumes of medical data and identify hidden patterns that may not be easily detected through conventional methods. This enables early detection and improves decision-making in medical practice.

Among various machine learning techniques, Random Forest and Logistic Regression are widely used for classification tasks due to their accuracy and efficiency. These algorithms analyze important medical attributes such as glucose level, body mass index (BMI), age, blood pressure, insulin level, and skin thickness to predict whether an individual is diabetic or non-diabetic.

II. Literature Survey

Several research studies have been conducted in the field of diabetes prediction using machine learning techniques. These studies highlight the effectiveness of different algorithms and approaches in improving prediction accuracy and early diagnosis.

Smith et al. (2023) analyzed the Pima Indians Diabetes Dataset using machine learning algorithms such as Logistic Regression, Random Forest, and Support Vector Machines. Their study emphasized the importance of data preprocessing techniques like normalization and feature selection, with results showing that Random Forest achieved higher classification accuracy compared to other models .

Ahmed et al. (2022) developed a predictive model using Random Forest classifiers on patient datasets containing attributes such as glucose level, BMI, and insulin. The use of cross-validation improved the reliability of the model, and the results demonstrated that Random Forest outperformed traditional statistical approaches .

Patel (2022) focused on the use of Logistic Regression for diabetes prediction. By applying feature selection techniques, the study reduced data dimensionality and achieved good classification accuracy, proving that Logistic Regression is effective for medical diagnosis problems .

Gupta et al. (2023) conducted a comparative analysis of multiple machine learning algorithms including Random Forest, Decision Tree, and Logistic Regression. Their findings indicated that Random Forest provided the highest accuracy due to its ensemble learning capability

III. System Analysis

System analysis focuses on understanding the need for an efficient and accurate diabetes prediction system. Diabetes is a rapidly increasing health issue worldwide, requiring early detection to avoid severe complications. Traditional diagnostic methods depend on laboratory tests and clinical expertise, which may delay timely diagnosis. There is a growing demand for automated systems that can analyze patient data quickly and accurately. Machine learning provides an effective solution by identifying patterns in medical datasets. The system must handle features such as glucose level, BMI, age, blood pressure, and insulin levels. Data preprocessing techniques like normalization and feature selection are essential for improving model

performance. The system should ensure high accuracy, reliability, and scalability. It should also support healthcare professionals in decision-making. Overall, the system aims to provide a fast, cost-effective, and intelligent solution for diabetes prediction.

Existing System

The existing system for diabetes detection mainly relies on traditional medical practices such as laboratory testing and physician diagnosis. Doctors analyze patient symptoms and medical reports to determine whether a person has diabetes. These methods include blood glucose tests, fasting tests, and oral glucose tolerance tests. Although accurate, these approaches require time, specialized equipment, and expert interpretation. In many cases, diagnosis occurs only after symptoms become severe. Existing systems do not effectively utilize historical healthcare data for predictive analysis. They lack automation and real-time prediction capabilities. Manual analysis can also lead to human errors and inconsistencies. Additionally, these systems are not easily accessible in remote or underdeveloped areas. Therefore, traditional systems are limited in providing early detection and preventive care.

Disadvantages of Existing System

- Time-consuming diagnostic process
- Requires laboratory tests and medical experts
- High cost of medical testing
- No early prediction or risk assessment
- Limited use of historical data
- Prone to human errors
- Not easily accessible in rural areas

Proposed System

The proposed system is a machine learning-based diabetes prediction model using Random Forest and Logistic Regression algorithms. It aims to provide early detection by analyzing patient medical data. The system uses a dataset containing important features such as glucose level, BMI, age, blood pressure, insulin level, and skin thickness. Data preprocessing techniques such as cleaning, normalization, and feature selection are applied to improve accuracy. Logistic Regression is used for simple and interpretable classification, while Random Forest enhances prediction accuracy through ensemble learning. The system automatically classifies patients as diabetic or non-diabetic. It reduces the need for manual analysis and speeds up the diagnosis process. The model is trained and tested to ensure reliable performance. It can be integrated into healthcare systems for real-time prediction.

Advantages of Proposed System

- Early detection of diabetes
- High prediction accuracy
- Reduces manual effort
- Fast and efficient diagnosis
- Cost-effective solution
- Utilizes historical healthcare data

IV. Methodology

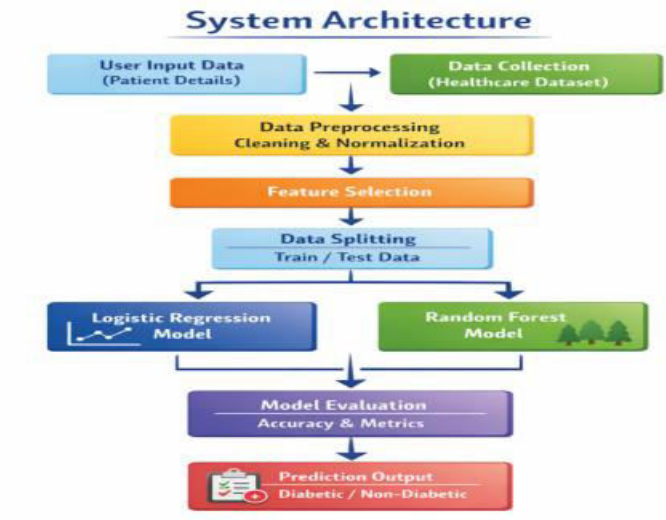
The proposed diabetes prediction system follows a structured methodology to ensure accurate and efficient results. Initially, a healthcare dataset containing important medical attributes such as glucose level, blood pressure, insulin level, body mass index (BMI), age, and pregnancies is collected. The data then undergoes preprocessing, which includes handling missing values, removing noise, and normalizing the data to improve model performance. After preprocessing, feature selection techniques are applied to identify the most relevant attributes that significantly contribute to diabetes prediction.

The dataset is then divided into training and testing sets to evaluate the performance of the models. Two machine learning algorithms, Logistic Regression and Random Forest, are used to build the predictive model. Logistic Regression is applied for its simplicity and interpretability, while Random Forest is used for its ability to improve accuracy through ensemble learning. Both models are trained using the training data and evaluated using performance metrics such as accuracy, precision, and recall.

System Architecture

The system architecture for the diabetes prediction model is designed in a step-by-step process to ensure efficient data handling and accurate prediction. Initially, patient data is collected either from a healthcare dataset or user input, which includes attributes such as glucose level, BMI, blood pressure, age, and insulin level. This data is then passed to the preprocessing stage, where missing values are handled, noise is removed, and normalization is performed to improve data quality.

After preprocessing, feature selection is applied to identify the most relevant attributes that significantly influence diabetes prediction. The processed data is then divided into training and testing datasets to evaluate model performance effectively. In the next stage, two machine learning models—Logistic Regression and Random Forest—are applied. Logistic Regression provides a simple and interpretable model, while Random Forest improves prediction accuracy through ensemble learning.



V. Results and Output

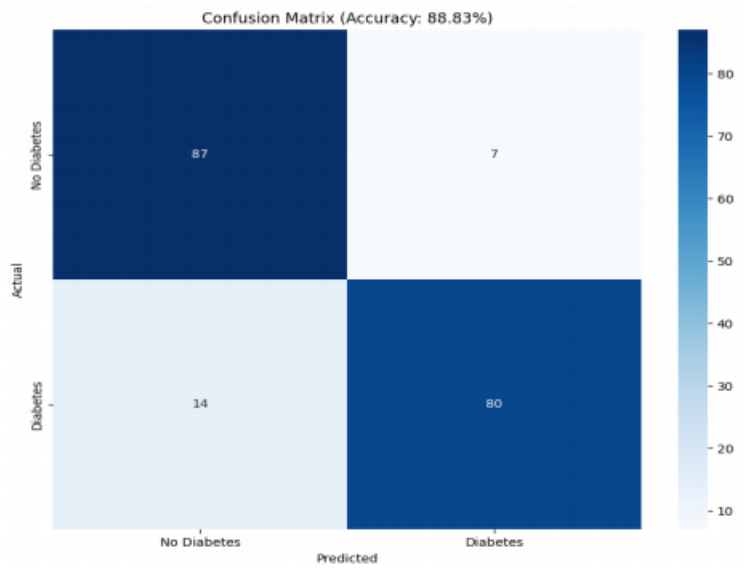
```
DIABETES PREDICTION SYSTEM
Enter Gender (M/F): M
Note: Pregnancies set to 0 for male patient
Glucose Level (mg/dL): 76
Blood Pressure (mm Hg): 120
Skin Thickness (mm): 56
Insulin Level (mu U/ml): 76
BMI Value: 76
Diabetes Pedigree Function: 98
Age (years): 34

RESULT: 🟢 NEGATIVE
Confidence: 52.60%
```

```
Fine-tuning RandomForest...
Fitting 5 folds for each of 81 candidates, totalling 405 fits

Best parameters: {'max_depth': 12, 'min_samples_leaf': 1, 'min_samples_split': 4, 'n_estimators': 500}
Best cross-validation score: 92.91%
Tuned model test accuracy: 88.30%
```

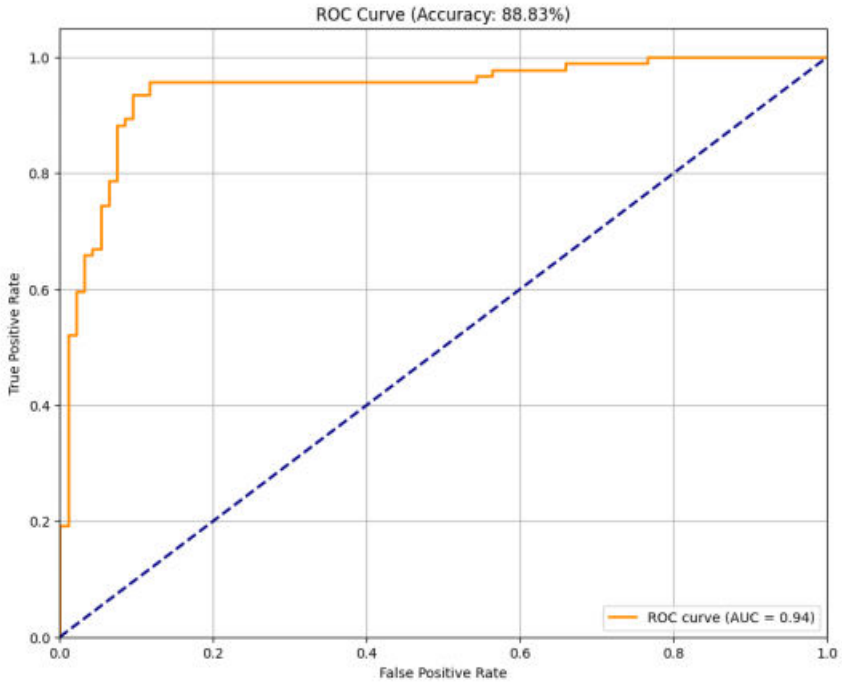
```
Note: No gender column found. Assuming females have pregnancies > 0
Male patients: 111, Female patients: 657

✅ Updated dataset saved with predictions!

Summary:
Has_Diabetes
No     632
Yes    136
Name: count, dtype: int64

Sample output:
   Pregnancies  Glucose   BMI Gender Has_Diabetes  Confidence_%
0            6      148  33.6      F           No         22.61
1            1       85  26.6      F           No          2.14
2            8      183  23.3      F           No         27.66
3            1       89  28.1      F           No          0.37
4            0      137  43.1      M          Yes         95.80
```

```
DIABETES PREDICTION SYSTEM
Enter Gender (M/F): F
Number of Pregnancies: 0
Glucose Level (mg/dL): 87
Blood Pressure (mm Hg): 111
Skin Thickness (mm): 68
Insulin Level (mu U/ml): 50
BMI Value: 86
Diabetes Pedigree Function: 90
Age (years): 20

RESULT: 🟢 NEGATIVE
Confidence: 69.76%
```

```
Training RandomForest...
Accuracy: 88.30%
ROC-AUC: 0.9478
CV Score: 92.37% (+/- 3.50%)

Training GradientBoosting...
Accuracy: 87.77%
ROC-AUC: 0.9520
CV Score: 93.58% (+/- 4.15%)

Training LogisticRegression...
Accuracy: 84.04%
ROC-AUC: 0.9061
CV Score: 83.82% (+/- 4.39%)

Training XGBoost...
Accuracy: 87.77%
ROC-AUC: 0.9477
CV Score: 92.91% (+/- 2.93%)

==================================================
MODEL PERFORMANCE SUMMARY
==================================================
                    accuracy   roc_auc   cv_mean    cv_std
RandomForest        0.882979  0.947770  0.923740  0.017493
GradientBoosting    0.877660  0.952014  0.935767  0.020741
XGBoost             0.877660  0.947714  0.929101  0.014628
LogisticRegression  0.840426  0.906066  0.838201  0.021928
```




```
DIABETES PREDICTION SYSTEM
Enter Gender (M/F): F
Number of Pregnancies: 3
Glucose Level (mg/dL): 87
Blood Pressure (mm Hg): 100
Skin Thickness (mm): 87
Insulin Level (mu U/ml): 54
BMI Value: 87
Diabetes Pedigree Function: 40
Age (years): 35

RESULT:  NEGATIVE
Confidence: 63.44%
```



VI. Conclusion

This project highlights the importance of integrating medical science with data engineering to improve healthcare systems. By utilizing Logistic Regression for its simplicity and interpretability, along with Random Forest for its high predictive accuracy, the developed model provides a balanced and reliable approach for diabetes prediction. The system effectively analyzes patient data and identifies individuals at risk of diabetes, enabling early diagnosis and better preventive care.

The results demonstrate that machine learning techniques can significantly enhance the efficiency and accuracy of disease prediction. Although such systems cannot replace the expertise of medical professionals, they serve as valuable tools to support clinical decision-making and reduce manual effort.

In the future, the integration of real-time data through IoT devices and advanced analytics can further improve prediction capabilities. This advancement can help in continuous monitoring, early intervention, and ultimately contribute to making diabetes more manageable and potentially preventable on a global scale.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, “An advanced movie recommender using collaborative filtering and sentiment analysis,” International

Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.