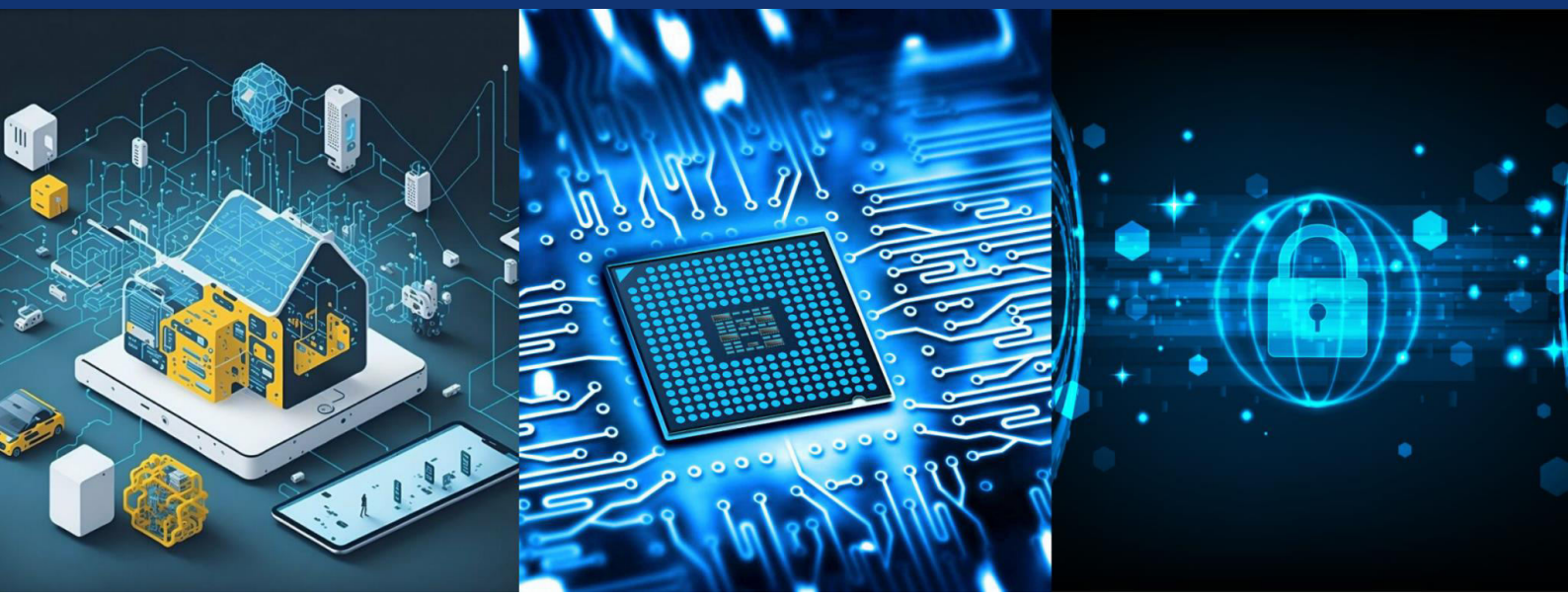
IJIRCCE

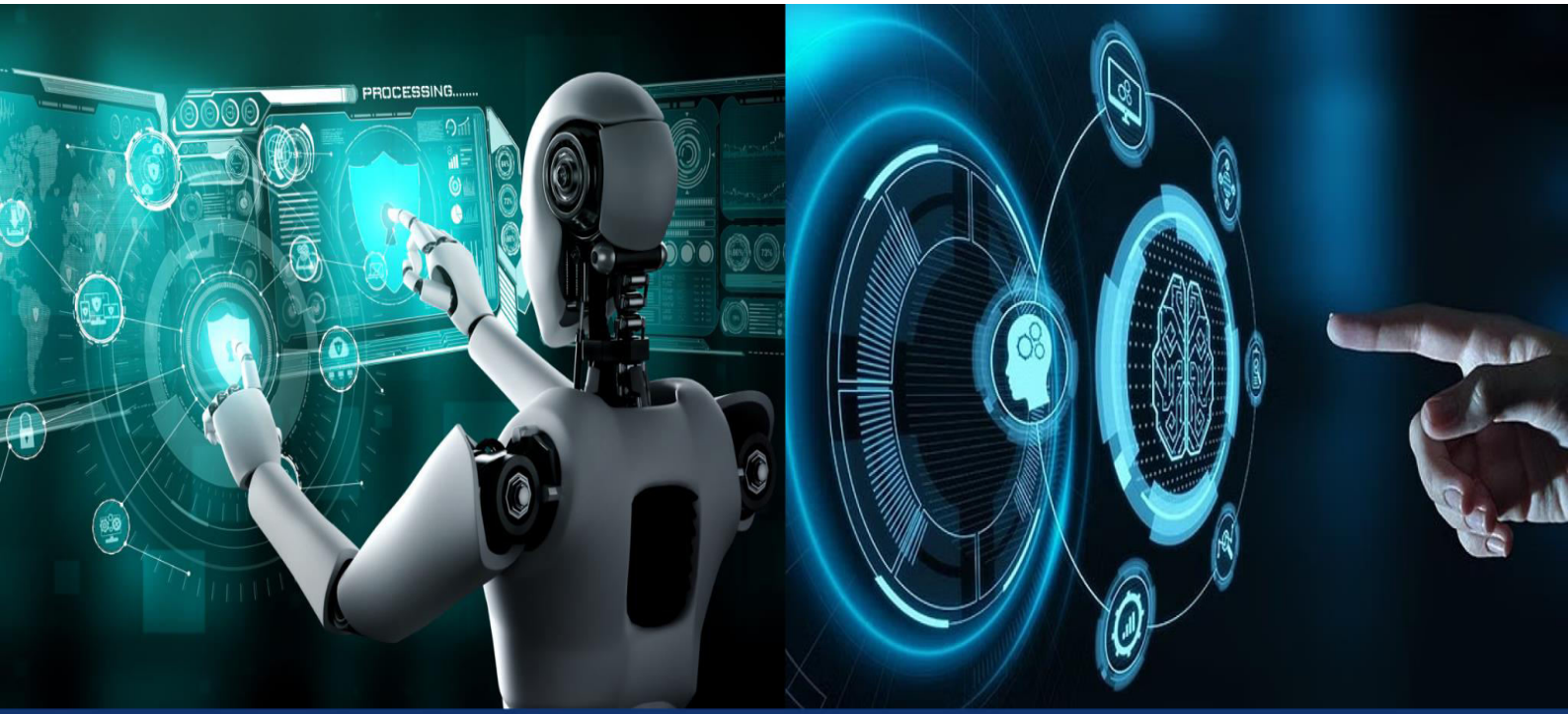
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

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



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304093

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

Optimizing Diabetes Prediction with XGBoost, Random Forest and Clinical Data.

Mrs Gandu Lavanya, D. Sai Pujitha, J. Sai Shiva, K Sai Sriraj, P. Sai Rakesh

Asst. Professor, Department of CSE, School of Engineering, Malla Reddy University, Hyderabad, Telangana, India

Department of CSE, School of Engineering, Malla Reddy University, Hyderabad, Telangana, India

ABSTRACT: Diabetes is a chronic metabolic disorder that affects millions of people worldwide, with significant implications for public health. Early prediction and intervention are crucial to managing the condition effectively and reducing its associated risks. This study explores the use of XGBoost (Extreme Gradient Boosting), a powerful machine learning algorithm, to enhance the prediction accuracy of diabetes in individuals based on clinical data. Various clinical features, including Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index(BMI), Age are analyzed to identify the most influential factors contributing to diabetes onset. The effectiveness of XGBoost is compared with other machine learning models, emphasizing its ability to handle imbalanced data, manage missing values, and provide feature importance insights. Through rigorous experimentation and model tuning, the study demonstrates that XGBoost significantly improves prediction performance, making it a valuable tool for clinical decision support in diabetes diagnosis and prevention. The results suggest that integrating clinical data with advanced machine learning techniques such as XGBoost can lead to more accurate, reliable, and early prediction systems for diabetes management.

KEYWORDS: Diabetes, Machine Learning, XGBoost, Random Forest, Clinical Data, Feature Selection.

I. INTRODUCTION

Diabetes is a chronic condition with significant health risks, making early detection and intervention crucial. Traditional diagnostic methods may lack the precision required for timely management. Machine learning techniques, particularly XGBoost (Extreme Gradient Boosting), offer an advanced solution by improving prediction accuracy using clinical data. XGBoost is known for its ability to handle complex datasets, missing values, and imbalanced data, making it ideal for clinical applications.

This study aims to optimize diabetes prediction by analyzing clinical features such as demographics, lifestyle, and medical history. By leveraging XGBoost, we aim to develop a highly accurate model for early diagnosis. The research also compares XGBoost with other machine learning algorithms, highlighting its effectiveness in improving diabetes prediction and supporting better clinical decision-making.

II. LITERATURE SURVEY

This refers to refining the process of forecasting whether an individual is likely to develop diabetes. The goal is to improve the performance (accuracy, precision, recall, etc.) of the prediction model, making it more reliable and effective in identifying individuals at risk. XGBoost (Extreme Gradient Boosting) is a popular machine learning algorithm known for its high performance in structured data problems. It is a type of gradient boosting technique that builds an ensemble of decision trees to make predictions. XGBoost is particularly known for its speed, scalability, and effectiveness in handling complex datasets. This refers to the health-related data collected from patients or individuals, such as medical history, test results, lifestyle factors, and other clinical parameters (e.g Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index(BMI), Age etc.). This data is typically used to identify risk factors and predict the onset of conditions like diabetes. Diabetes is a chronic metabolic disorder that affects millions of people worldwide, with significant implications for public health. Early prediction and intervention are crucial to managing the condition effectively and reducing its associated risks. This study explores the use of XGBoost (Extreme Gradient Boosting), a powerful machine learning algorithm, to enhance the prediction accuracy of diabetes in individuals based on clinical data. Various clinical features, including Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index(BMI), Age are analyzed to identify the most influential factors contributing to diabetes onset. The effectiveness of XGBoost is compared with other

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

machine learning models, emphasizing its ability to handle imbalanced data, manage missing values, and provide feature importance insights. Through rigorous experimentation and model tuning, the study demonstrates that XGBoost significantly improves prediction performance, making it a valuable tool for clinical decision support in diabetes diagnosis and prevention. The results suggest that integrating clinical data with advanced machine learning techniques such as XGBoost can lead to more accurate, reliable, and early prediction systems for diabetes management. Diabetes is a prevalent metabolic disorder characterized by high blood sugar levels, leading to severe health complications. Early diagnosis is crucial to prevent or delay associated risks. Traditional diagnosis methods rely on clinical tests, which can be time-consuming and resource-intensive. Machine learning techniques offer promising solutions for automated diabetes detection, improving accuracy and reducing diagnostic delays. Among various machine learning models, XGBoost and Random Forest (RF) have demonstrated high predictive performance in diabetes classification.

Several studies have explored different machine learning algorithms to predict diabetes, emphasizing feature selection techniques to enhance classification accuracy. Kandhasamy and Balamurali (2015) compared k-NN, Support Vector Machines (SVM), Random Forest (RF), and J48 decision trees for diabetes prediction. Their study found that k-NN (k=1) and RF achieved 100% accuracy in their dataset. Xu et al. (2017) conducted a comparative study of Naïve Bayes, ID3, RF, and AdaBoost classifiers, reporting that RF outperformed other models with 85% accuracy. Paul and Choubey (2017) developed a hybrid model combining a Radial Basis Function Network (RBFN) with a Genetic Algorithm (GA), which showed superior performance compared to standalone RBFN models. Zou et al. (2018) applied Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR) for feature selection, obtaining 80.84% accuracy using RF. Wu et al. (2018) proposed a data mining-based approach integrating an improved K-Means algorithm and Logistic Regression, achieving 95.42% accuracy. Ayon and Islam (2019) used deep learning on the PIMA Indian Diabetes dataset, achieving 98.35% accuracy. Tigga and Garg (2020) employed a dataset with 18 questionnaire-based features and found that RF provided 94.10% accuracy. Naz and Ahuja (2020) tested multiple classifiers, including Naïve Bayes, Decision Trees, Deep Learning (DL), and Artificial Neural Networks (ANN). Their deep learning model achieved 98.07% accuracy. Islam et al. (2020) analyzed diabetes prediction using Naïve Bayes, Logistic Regression, and RF. The best performance was obtained using RF (99% accuracy) with 16 input features. Gourisaria et al. (2022) tested Random Forest with feature selection techniques, achieving 99.2% accuracy.

III. PROBLEM STATEMENT

Diabetes is a chronic disease that requires accurate and early diagnosis to prevent severe complications. Traditional diagnostic methods rely on clinical assessments, which may be time-consuming and less efficient in handling large datasets. Machine learning techniques like XGBoost and Random Forest can enhance predictive accuracy by analyzing complex patterns in clinical data. However, selecting the best model and optimizing its performance for real-world healthcare applications remains a challenge. This study aims to compare and optimize these machine learning models to improve diabetes prediction and management, leveraging key clinical features for better decision-making.

IV. OBJECTIVES

A. Early Detection

Develop a comprehensive tool capable of detecting of the major disease Pneumonia at an early stage to improve patient outcomes and reduce mortality rates.

B. High Accuracy

Utilize advanced deep learning models, including CNN to achieve high accuracy in disease detection, ensuring reliable diagnostic results for each condition.

C. User – Friendly Interface

Design and implement an intuitive and accessible user interface that allows healthcare professionals and non-experts alike to use the application effectively for disease detection and diagnosis

D. Non -Invasive Diagnostic Tool

Provide a non-invasive diagnostic solution that reduces the need for costly and uncomfortable procedures, making healthcare more accessible and less burdensome for patients.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304093

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

V. PROPOSED SYSTEMS

The current system for diagnosing diabetes relies on clinical diagnostic tests such as plasma glucose measurements, which require blood samples to be taken and analyzed in laboratories. This process can be time-consuming and prone to errors due to specimen collection, evaluation time, and instrumental inaccuracies.

In recent years, artificial intelligence-based systems have been introduced to automate diabetes detection, utilizing machine learning models like k-NN, support vector machines, random forest, Naïve Bayes, logistic regression, deep learning, and ensemble methods. Some studies have focused on hybrid approaches, such as genetic algorithms with radial basis function networks, principal component analysis for dimensionality reduction, and deep learning-based detection models.

- Many approaches require extensive input data, including blood tests, which are not always easily obtainable.
- Some models do not adequately test for statistical significance between independent and dependent variables, raising concerns about feature reliability.
- The computational time and complexity increase when handling large datasets with high-dimensional features.

To address these limitations, the study proposes a new approach combining Multiple Linear Regression (MLR), Random Forest (RF) for feature selection, and XGBoost (XG) for classification. This method aims to enhance prediction accuracy while reducing computation time and reliance on blood test data.

VI. METHODOLOGY

6.1. Data Collection

The dataset used in this study consists of clinical data, including glucose levels, BMI, age, insulin levels, and blood pressure. Data preprocessing includes normalization, outlier detection, and handling missing values to ensure model efficiency. The dataset is sourced from publicly available repositories, ensuring reliability and diverse sample representation.

6.2. Feature Selection

Random Forest is utilized to identify the most significant predictors, reducing the number of input variables from 16 to 9 without compromising accuracy. Feature selection techniques such as Recursive Feature Elimination (RFE) and statistical correlation analysis are employed to refine the dataset further.

6.3. Model Development

XGBoost is implemented as the primary classification algorithm due to its robustness in handling imbalanced datasets. The model
undergoes hyperparameter tuning using techniques such as grid search and Bayesian optimization to enhance performance. A comparative analysis is conducted with other machine learning models, including Random Forest, logistic regression, and support vector machines.

The training phase includes multiple iterations with cross-validation to prevent overfitting. The dataset is split into training (80%) and testing (20%) sets to evaluate generalization capability. Model performance is fine-tuned using learning rate adjustments, tree depth optimization, and feature scaling.

6.4. Web Based Deployment

To provide real-time pneumonia detection, the trained model is integrated into a Flask-based web application. Users can upload their details via the web interface, and the backend processes them using the trained XGBoost and Random Forest model. The results are displayed instantly, indicating whether diabetes is detected. The frontend, developed with HTML, CSS, and JavaScript, provides an intuitive and responsive user experience. Flask handles data uploads, model inference, and result rendering dynamically. This deployment enables remote medical diagnosis, making diabetes detection more accessible to healthcare professionals and individuals without requiring in-person visits.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304093

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

VII. USE CASE DIAGRAM

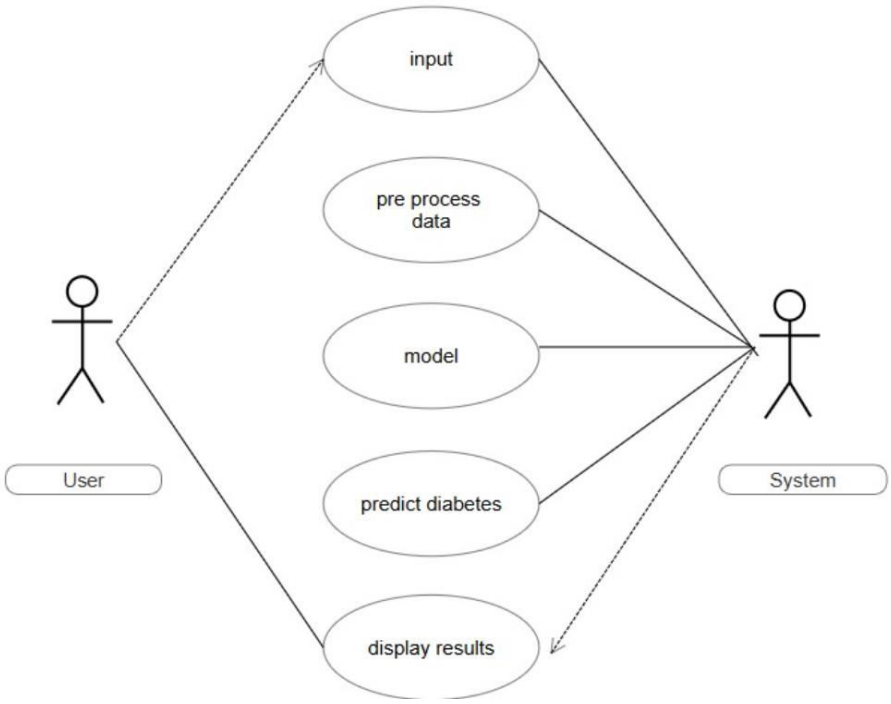


FIGURE 1. Use case diagram

VIII. SYSTEM ARCHITECTURE

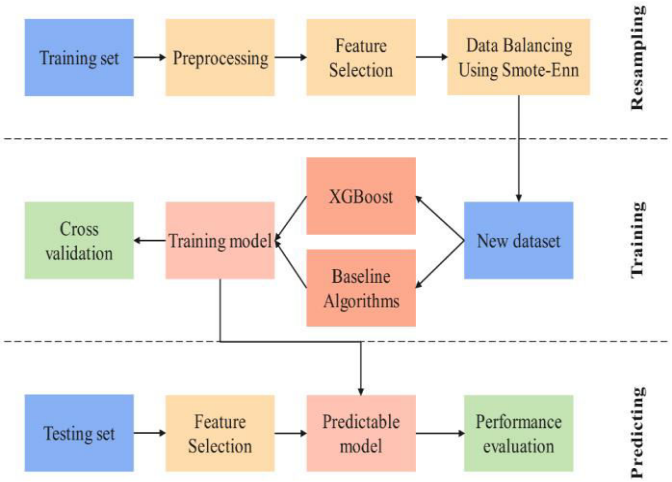


FIGURE 2. System architecture for proposed pneumonia detection model .

IX. RESULTS AND DISCUSSION

The Diabetes detection system achieves accuracy of 80% using the XGBoost and Random Forest approach. This application has the front-end website where the user can enter their details and get the results.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304093

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

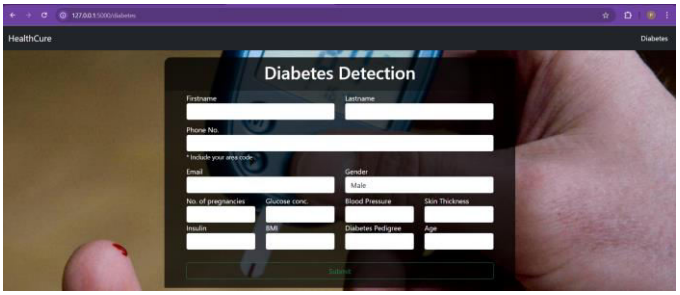


FIGURE 3. Home Page

The user gives the details asked in the form then the model takes those input from the user and predicts whether the person has the diabetes or not. Let's see an example for Negative.

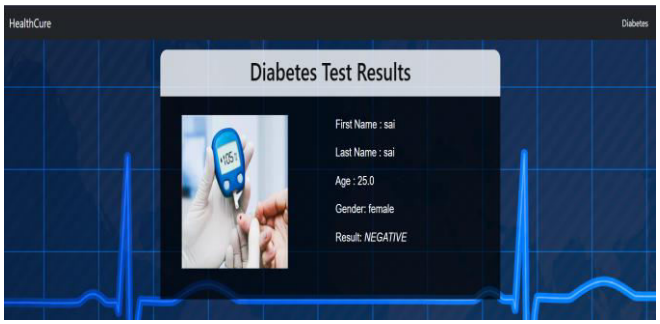


FIGURE 4. Negative Result

Let's see an example for Positive.

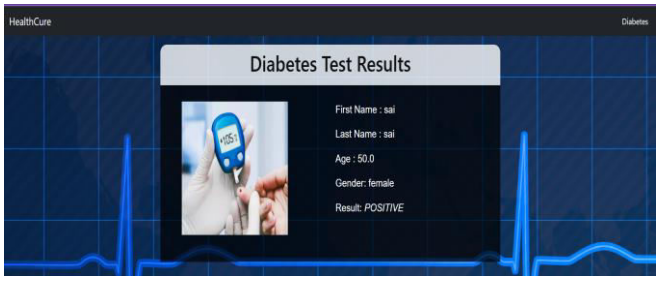


FIGURE 5. Positive Test Results

TABLE 1. RESULTS

Name	Age	Gender	Result
Rajesh	31	Male	Positive
Shiva	47	Male	Positive
Sai	53	Male	Negative
Shivani	35	Female	Positive
Rithu	39	Female	Negative
Geetha	25	Female	Positive

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

X. CONCLUSION

This study demonstrates the effectiveness of XGBoost in optimizing diabetes prediction using clinical data. By employing key preprocessing techniques like data cleaning, normalization, and augmentation, the dataset was prepared for optimal model performance. XGBoost outperformed traditional machine learning algorithms, including Random Forest, providing better handling of imbalanced data, reducing overfitting, and offering valuable feature importance insights. While Random Forest also performed well, XGBoost's superior predictive accuracy and ability to manage complex interactions between features made it a more suitable choice for this task. These capabilities make XGBoost particularly advantageous for clinical applications where accuracy, interpretability, and timely decision-making are crucial. The findings underscore the potential of machine learning, specifically XGBoost, in improving early diabetes detection. By integrating clinical data effectively, this approach can aid healthcare professionals in making more accurate predictions, leading to earlier interventions and better patient outcomes. Future work may involve further model optimization, incorporating additional data sources, and real-world implementation, ultimately contributing to more efficient and effective diabetes management strategies.

REFERENCES

[1] Ayon SI, Islam MM (2019) Diabetes prediction: a deep learning approach. Int J Inform Eng Electro Business(IJIEEB) MECS press 11(2):21–27. https://doi.org/10.5815/ijieeb.2019.02.03
[2] Ayon SI, Islam MM, Hossain MR (2020) Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE J Res 2020. https://doi.org/10.1080/03772063.2020.1713916
[3] Bloomgarden ZT (2020) Diabetes and COVID-19. J Diabetes 12:347–348. https://doi.org/10.1111/1753-0407. 13027
[4] Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
[5] Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, Pratt M, Ekelund U, Yngve A, Sallis JF, Oja P (2003) International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 35(8):1381–1395. https://doi.org/10.1249/01.MSS.0000078924.61453.FB
[6] Damle R, Alavi K (2016) The University Healthsystem consortium clinical database: an emerging resource in colorectal surgery research. Sem Colon Rectal Surg 27(2):92–95. https://doi.org/10.1053/j.scrs.2016.01.006
[7] Dong Y, Ma X, Fu T (2021) Electrical load forecasting: a deep learning approach based on K-nearest neighbors. Appl Soft Comput 99:106900. https://doi.org/10.1016/j.asoc.2020.106900
[8] Frank EA, Shubha MC, D'Souza CJM (2012) Blood glucose determination: plasma or serum? J Clin Lab Anal 26(5):317–320. https://doi.org/10.1002/jcla.21524
[9] Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451
[10] Garcia-Carretero R, Vigil-Medina L, Mora-Jimenez I, Soguero-Ruiz C, Barquero-Perez O, Ramos-Lopez J (2020) Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population. Med Biol Eng Comput 58:991–1002. https://doi.org/10.1007/s11517-020-02132-w
[11] Ghasemi J, Saaidpour S, Brown SD (2007) QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. J Mol Struct THEOCHEM 805(1–3):27–32. https://doi.org/10.1016/j.theochem.2006.09.026
[12] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H (2018) Predicting diabetes mellitus with machine learning techniques. Front Genet 9(515):1–10. https://doi.org/10.3389/fgene.2018.00515
[13] World Health Organization (n.d.) Diabetes. https://www.who.int/health-topics/diabetes#tab=tab_1.
[14] World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia (2006). https://www.who.int/diabetes/publications/diagnosis_diabetes2006/en/
[15] Wu H, Yang S, Huang Z, He J, Wang X (2018) Type 2 diabetes mellitus prediction model based on data mining. Inform Med Unlocked 10:100–107. https://doi.org/10.1016/j.imu.2017.12.006

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details