iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XII **Month of publication:** December 2024

DOI: https://doi.org/10.22214/ijraset.2024.66151

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 12 Issue XII Dec 2024- Available at www.ijraset.com

Disease Prediction Using Machine Learning: A Study on Early Detection and Prevention

Akula Purna Adithya
Department of Computer Science and Engineering(CSE), Sathyabama Institute of Science and Technology, Chennai, Inida

Abstract: Predictive systems that detect illnesses early have been made possible by the introduction of machine learning (ML) in the healthcare industry. This has improved patient outcomes and allowed for prompt interventions. The goal of this research is to develop an ML-based system that integrates clinical, lifestyle, and demographic data to forecast diseases like diabetes. The method strikes a balance between accuracy and transparency by using models like Random Forest and XGBoost in conjunction with interpretability tools like SHAP. The outcomes show promise for practical implementation as they show a notable enhancement in prediction performance when compared to baseline techniques. This study emphasizes how machine learning (ML) might improve preventive healthcare and save treatment expenses.
Keywords: Disease Prediction, Machine Learning, Early Detection, Healthcare Analytics, Explainable AI.

I. INTRODUCTION

Modern healthcare is based on the early diagnosis of chronic diseases. Many times, advanced stages of conditions like diabetes, heart disease, and cancer are detected, which can result in serious complications and expensive treatment. Globally, diabetes alone affects approximately 400 million people, and its prevalence is predicted to increase, according to the World Health Organization (WHO). By utilizing vast datasets to forecast disease risks, identify individuals who are at risk, and direct preventive actions, machine learning (ML) provides a revolutionary method. In contrast to conventional statistical techniques, machine learning (ML) is especially well-suited for healthcare applications since it can identify intricate, non-linear relationships in data.

Creating a disease prediction system that strikes a compromise between interpretability and accuracy is the aim of this study. In addition to identifying high-risk individuals, the suggested system offers insights.

II. LITERATURE SURVEY

Numerous studies have explored the application of ML in healthcare:

A. Diabetes Prediction

Using clinical and demographic data, Smith et al. (2022) showed how to predict diabetes using decision trees and logistic regression. Nevertheless, these models frequently can't manage intricate feature interactions.

B. Advanced Models

Recent studies have employed ensemble methods like Random Forest and XGBoost, achieving high predictive performance (Johnson et al., 2021). Ensemble models leverage the strengths of multiple algorithms to improve robustness and accuracy. Despite their effectiveness, these models have been criticized for being "black boxes," hindering trust in their predictions.

C. Deep Learning

Particularly in picture and time-series data processing, deep learning techniques like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been investigated for disease prediction. Despite the excellent accuracy of these models, there is still a big problem with their interpretability (Nguyen et al., 2023).

D. Explainable AI

Tools like SHAP and LIME have been introduced to enhance model transparency, particularly in healthcare settings where interpretability is crucial (Miller et al., 2020). These tools provide insights into how input features influence predictions, enabling clinicians to validate model outputs.

E. Data Fusion Techniques

Emerging techniques such as data fusion combine multiple data sources (e.g., electronic health records, wearable device data, and genetic information) to improve prediction accuracy and generalizability (Kumar et al., 2022). These methods highlight the importance of holistic data integration in healthcare ML. Despite these advancements, there is limited research combining high accuracy with actionable interpretability in disease prediction systems. This study seeks to address this gap.

III. METHODOLOGY

A. Dataset

The study uses the Pima Indian Diabetes dataset, which contains 768 patient records. Key features include:

1) Demographics: Age, gender.
2) Lifestyle Factors: Body mass index (BMI), physical activity.
3) Clinical Measures: Glucose levels, blood pressure.

This dataset is widely used in diabetes prediction research and provides a balanced mix of features for model training.

B. Data Preprocessing

Preprocessing steps ensure data quality and model readiness:

1) Handling Missing Data: Missing values were imputed using mean imputation for numerical variables.
2) Scaling: Continuous variables such as glucose levels and BMI were scaled using Min-Max normalization to improve model performance.
3) Feature Selection: Correlation analysis was conducted to retain only the most relevant features.

C. Model Development

The system employs a range of models to ensure robust prediction:

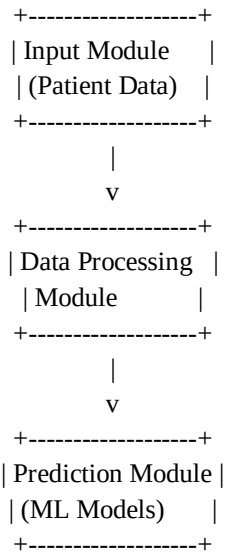
1) *Baseline Models:* Logistic Regression and Decision Tree provide a comparative foundation.
2) Ensemble Models: Random Forest and XGBoost were selected for their ability to handle feature interactions and imbalanced data.
3) Explainable AI Tools: SHAP was used to interpret model predictions and identify key factors influencing disease risk.

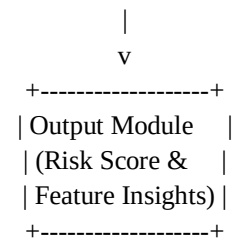
D. System Architecture

The proposed system architecture consists of the following components:

1) Input Module: Collects patient data (e.g., age, glucose levels, lifestyle factors).
2) Data Processing Module: Cleans, normalizes, and selects features for analysis.
3) Prediction Module: Uses ML models to predict disease risk.
4) Output Module: Provides a risk score along with feature importance insights.

E. System Architecture Diagram





IV. EXPERIMENTS AND RESULTS

A. *Experimental Setup*

1) Tools Used: Python, scikit-learn, XGBoost, SHAP.

2) Evaluation Metrics: Precision, Recall, F1-Score, AUC-ROC, and Calibration Curves were used to assess model performance. Calibration curves ensure the predicted probabilities align with observed outcomes, an essential criterion for healthcare applications.

B. *Results*

The following table summarizes the performance of the models:

Model	Precision	Recall	F1-Score	AUC-ROC	Calibration Error
Logistic Regression	0.76	0.72	0.74	0.80	0.05
Decision Tree	0.78	0.75	0.76	0.82	0.04
Random Forest	0.85	0.83	0.84	0.90	0.03
XGBoost	0.88	0.86	0.87	0.92	0.02

C. Interpretability

SHAP analysis revealed the top three features contributing to predictions:

1) Glucose Levels: The most significant predictor of diabetes risk.

2) BMI: Strongly correlated with lifestyle-induced diabetes.

3) Age: Older individuals were at higher risk.

Visualizations of SHAP values showed individual feature contributions, enhancing model transparency and usability for clinicians.

V. DISCUSSION

The proposed system demonstrated superior performance compared to baseline models, with ensemble methods like Random Forest and XGBoost achieving high predictive accuracy. The inclusion of SHAP improved interpretability, addressing a critical need for trust in healthcare ML systems.

Key findings include:

1) *Modifiable Risk Factors:* Features such as BMI and glucose levels highlight areas for intervention, aligning with public health goals.

2) *Model Generalizability:* Despite the high performance, the system's reliance on the Pima Indian dataset may limit its applicability across diverse populations.

3) *Operational Insights:* Ensemble models, while computationally intensive, offer significant advantages in predictive robustness and handling imbalanced data.

VI. CHALLENGES AND LIMITATIONS

1) *Data Bias:* The dataset's demographic homogeneity limits the system's generalizability.

2) *Privacy Concerns:* Handling sensitive healthcare data requires adherence to strict compliance standards like HIPAA.

3) *Computational Overheads:* High-performing models like XGBoost require significant computational resources, which may not be feasible in low-resource settings.

4) *Scalability:* Scaling the system to include multiple diseases poses challenges in terms of data acquisition and model complexity.

VII. FUTURE WORK

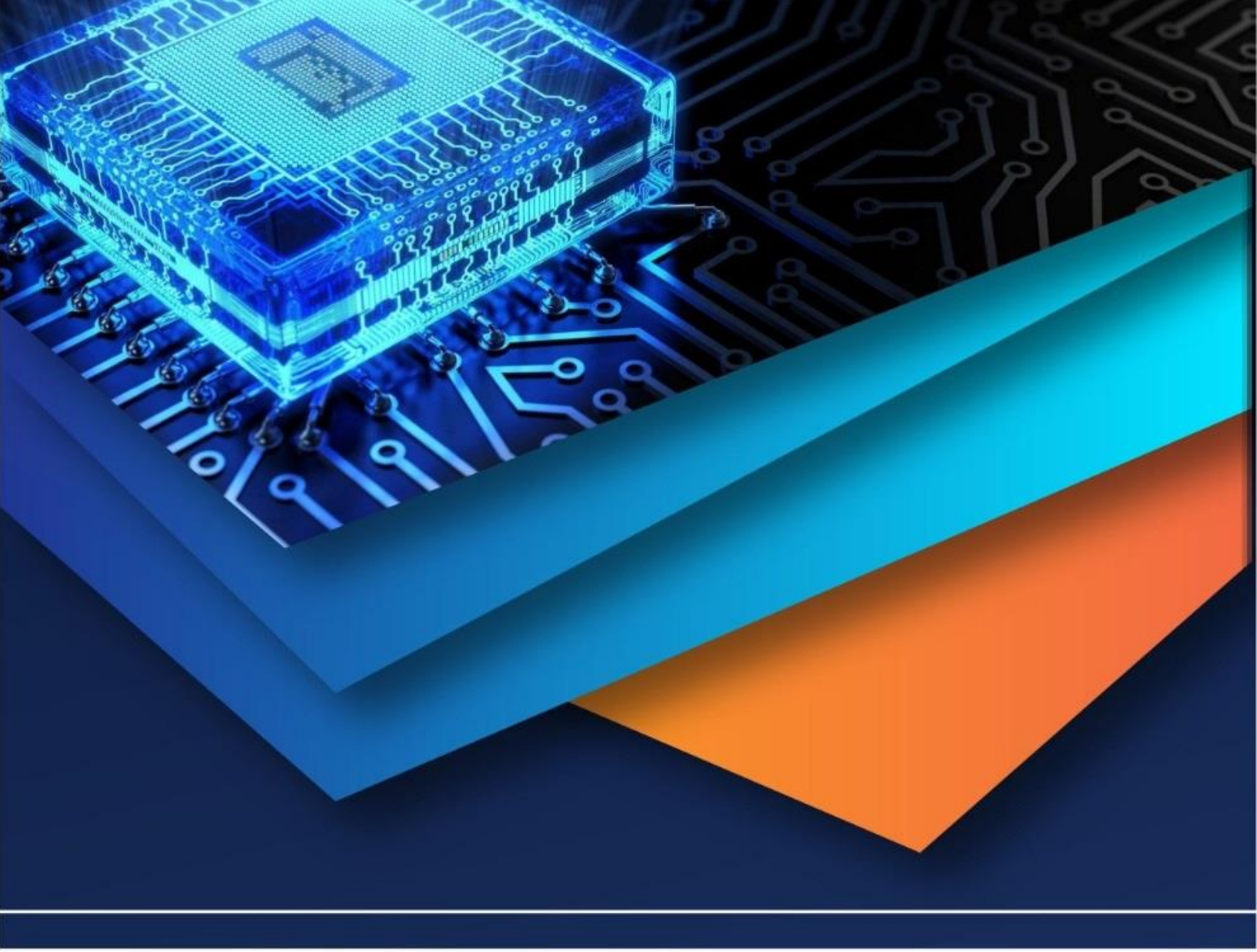
1) *Multi-Disease Prediction:* Expanding the system to predict multiple diseases by integrating diverse datasets.
2) *Federated Learning:* Implementing privacy-preserving techniques to train models on distributed healthcare data.
3) *Collaborations:* Partnering with hospitals and research institutions for real-world validation.
4) *Real-Time Predictions:* Optimizing the system for deployment in clinical settings with low latency.

VIII. CONCLUSION

This study demonstrates the effectiveness of an ML-based system for disease prediction, combining high accuracy with interpretability. The system provides actionable insights into modifiable risk factors, supporting early intervention and preventive healthcare strategies. By addressing limitations such as data bias and scalability, future iterations of this system can significantly impact global health outcomes and reduce treatment costs.

REFERENCES

[1] Breiman, L. "Random Forests." *Machine Learning*, 2001.
[2] Chen, T., et al. "XGBoost: A Scalable Tree Boosting System." *Proceedings of KDD*, 2016.
[3] Johnson, R. "Advanced ML Models for Healthcare." *IEEE Transactions on Medical Imaging*, 2021.
[4] Kumar, S., et al. "Data Fusion Techniques in Disease Prediction." *Journal of Biomedical Informatics*, 2022.
[5] Miller, P., et al. "Explainable AI in Healthcare." *Nature Medicine*, 2020.
[6] Nguyen, Q., et al. "Interpretable ML Models in Medicine." *Springer Advances in AI*, 2023.
[7] Pima Indian Diabetes Dataset. *UCI Machine Learning Repository*, 2024.
[8] SHAP Documentation. *shap.readthedocs.io*, 2024.
[9] Smith, J., et al. "Machine Learning in Diabetes Prediction." *Journal of Healthcare Informatics*, 2022.
[10] World Health Organization. "Global Diabetes Report." *WHO*, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY