International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Early Disease Prediction System Using Machine Learning

Spoorthi Ellur [1], Saleema[2],Ruhi Naz N S[3]
Department of CSE-AI, Ballari Institute of Technology and Management, Ballari, Karnataka, India.
Email: [1]spoorthiellur05@gmail.com, [2]saleemak123@gmail.com, [3]ruhinazns@gmail.com

----------**********----------

Abstract:

Heart disease, diabetes, and high blood pressure are among the leading causes of death worldwide. Early detection of these diseases can save lives, but regular checkups are not always possible for everyone. This paper presents a review of machine learning techniques used for predicting diabetes, heart disease, and hypertension. For diabetes detection, studies have shown that Random Forest and K-Nearest Neighbors achieve good accuracy using the Pima Indian dataset [1]. For heart disease prediction, Logistic Regression and Cascaded Neural Networks have been effectively used with the Cleveland Heart Disease dataset [2][5]. For blood pressure classification, Random Forest and Decision Tree algorithms have demonstrated strong performance on large datasets [3]. Workplace screening programs analyzed using K-Means clustering have shown that regular health checkups help reduce hypertension over time [4]. The findings indicate that machine learning models can be successfully integrated into a single system to predict multiple diseases using basic health information, making early disease detection accessible and affordable.

Keywords: Disease Prediction, Machine Learning, Diabetes, Heart Disease, Hypertension, Random Forest, Logistic Regression, Decision Tree.

----------**********----------

I. INTRODUCTION

This Heart disease, diabetes, and high blood pressure are three of the biggest health problems facing the world today. Together, they cause millions of deaths every year. According to the World Health Organization, heart disease and stroke alone account for about 17.5 million deaths annually [2]. Around 537 million adults have diabetes, and this number is expected to reach 784 million by 2045 [1]. At the same time, about 1.28 billion people have high blood pressure, and nearly half of them do not even know they have it [4].

What makes these diseases so dangerous is that they often show no symptoms in the early stages. Many people only find out they have a problem when it is too late, after a heart attack, stroke, or kidney failure has already occurred. Finding these diseases early can save lives. When caught early, patients can make lifestyle changes or take medicine to keep their condition under control. But there is a big problem. In many parts of the world, especially in poor and middle-income countries, there are not enough doctors or hospitals to screen everyone regularly.

This is where machine learning can help. Machine learning is a branch of artificial intelligence where computers learn from data and make predictions. Many researchers have used machine learning to predict individual diseases. For diabetes, models like Random Forest have shown good results [1]. For heart disease, Logistic Regression and Cascaded Neural Networks have been found effective [2][5]. For blood pressure, Decision Tree and Random Forest have performed well [3].

However, most existing studies focus on only one disease at a time. A person has to use different systems for different health problems. There is no single place where someone can check their risk for multiple diseases using simple health information.

 ©IJSRED: All Rights are Reserved

Also, many of these systems do not save past results, so users cannot track how their health is changing over time. This review paper examines the machine learning techniques used for disease prediction and discusses how they can be integrated into a single system.

The rest of this paper is organized as follows. Section II presents the literature review. Section III explains the methodology. Section IV discusses the results. Section V provides the conclusion

II. LITERATURE REVIEW

Many researchers have worked on predicting individual diseases like diabetes, heart disease, and high blood pressure using machine learning. However, very few have built a complete system that combines all three diseases into one easy-to-use platform.

For diabetes prediction, most studies have used the Pima Indian dataset, which contains health information of 768 patients. This dataset includes features like glucose level, BMI, age, blood pressure, insulin, and number of pregnancies. A study by Iparraguirre-Villanueva and colleagues compared five models and found that Random Forest performed well for diabetes detection [1]. Other studies have also shown that ensemble methods like Random Forest are effective for medical diagnosis because they combine multiple decision trees to give more reliable results. Khanam and Foo also compared machine learning algorithms for diabetes prediction and found that Logistic Regression and SVM gave good accuracy [7]. Theerthagiri and colleagues examined several algorithms including KNN, Naive Bayes, and Decision Tree, finding that KNN achieved 86% accuracy [9].

For heart disease prediction, many researchers have used the Cleveland Heart Disease Dataset, which contains patient records with health features including age, cholesterol, blood pressure, and chest pain type. Palaniappan and Awang built an intelligent heart disease prediction system using data mining techniques like decision trees, naive Bayes, and neural networks [6]. Chitra and Sevenivasagam used a Cascaded Neural Network and achieved 85% accuracy, which was better than standard SVM that gave 82% accuracy [5]. Their work showed that neural networks can be effective for heart disease prediction when the network architecture is allowed to grow automatically. El-Sofany and colleagues tested ten different algorithms and found that XGBoost gave the best results with 97.57% accuracy, though they also noted that Logistic Regression and Random Forest performed well [2].

For blood pressure prediction, Shrivastava and colleagues tested four algorithms on a large dataset of 50,000 blood pressure readings. They found that Random Forest outperformed K-Nearest Neighbors, Logistic Regression, and Decision Tree, achieving the highest accuracy and F1 scores [3]. Their study concluded that tree-based methods are effective for predicting both systolic and diastolic blood pressure using clinical characteristics. Another study by Adeleke and colleagues used K-Means clustering to analyze workplace hypertension screening data from 1,723 employees over four years. They found that regular screenings helped reduce high blood pressure from 29% to 18%, showing that periodic health assessments are effective [4]. Tsoi and colleagues also found that K-Means clustering produced consistent and dependable results for diagnosing hypertension [8].

Despite all this research, most existing systems have several limitations. First, they focus on only one disease, so users need to use different systems for different health problems. Second, they do not have user account systems, so predictions cannot be saved or tracked over time. Third, they do not provide clear, personalized health advice based on prediction results. This review aims to address these gaps by examining how multiple machine learning models can be integrated into a single disease prediction system.

III. METHODOLOGY

The methodology of this disease prediction system follows a simple step-by-step workflow that

 ©IJSRED: All Rights are Reserved

allows users to register, log in, enter their health details, get instant disease predictions, and save their results for future reference. The workflow begins with user registration, where a new user enters their full name, email address, username, and password. After registration, the user can log in by providing their username and password. Upon successful authentication, the user is granted access to the dashboard.

Once logged in, the user can choose from three disease prediction options: diabetes, heart disease, or blood pressure. For diabetes prediction, the user enters health details such as glucose level, BMI, age, blood pressure, number of pregnancies, and insulin level. These inputs are sent to a Random Forest model that has been trained on medical datasets. The Random Forest algorithm works by creating multiple decision trees and combining their results through majority voting to produce a final prediction with a risk level of low, medium, or high. Research has shown that Random Forest is effective for diabetes detection [1].

For heart disease prediction, the user provides age, cholesterol level, blood pressure, maximum heart rate, and chest pain type. This data goes into a Logistic Regression model that uses a sigmoid function to calculate the probability of heart disease. Studies have demonstrated that Logistic Regression is a reliable method for heart disease classification [2][5]. For blood pressure classification, the user enters systolic and diastolic readings, age, and weight. A Decision Tree model then classifies the blood pressure as normal, elevated, or high by creating a tree-like structure where each internal node tests a feature and each leaf node gives a class label. Previous research has shown that Decision Tree algorithms perform well for blood pressure prediction using clinical characteristics [3].

After the prediction is made, the system saves the result along with the disease type, risk level, and timestamp into a database. The user can view their complete prediction history at any time, which helps them track their health status over time. The system also displays clear medical advice based on the risk level, such as "Consult a doctor immediately" for high-risk cases or "Maintain a healthy lifestyle" for low-risk cases. Regular workplace screening programs have shown that periodic health assessments can effectively help in early detection and control of diseases like hypertension [4]. This entire workflow is designed to be simple, fast, and user-friendly so that anyone, even without medical knowledge, can use it to assess their health risks.

IV. RESULTS AND DISCUSSION

The disease prediction system was tested with various input values for diabetes, heart disease, and blood pressure. For diabetes, when the user entered normal glucose level of 100 mg/dL, healthy BMI of 21, age of 30, and normal blood pressure, the Random Forest model correctly predicted "Low Diabetes Risk". When the user entered high glucose of 180 mg/dL, very high BMI of 65, and high blood pressure, the model correctly predicted "High Diabetes Risk". These results align with clinical knowledge that high glucose and obesity are strong risk factors for diabetes [1].

For heart disease, when the user entered normal blood pressure and non-anginal chest pain, the Logistic Regression model predicted "Low Heart Disease Risk". When the user entered high cholesterol of 280 mg/dL, fluctuating blood pressure, and asymptomatic chest pain, the model correctly predicted "High Heart Disease Risk". This is consistent with medical research showing that high cholesterol and asymptomatic chest pain are serious risk factors for heart disease [2][5].

For blood pressure, when the user entered normal readings of 120/80 mmHg, the Decision Tree model correctly classified "Normal Blood Pressure". When the user entered high readings of 180/120 mmHg, the model correctly classified "High Blood Pressure". These results match standard medical guidelines for blood pressure classification [3][4].

All predictions were saved to the database with timestamps, and the system displayed appropriate color-coded risk levels and clear medical advice. The system successfully demonstrated that Random Forest, Logistic Regression, and Decision Tree can

 ©IJSRED: All Rights are Reserved

be effectively used together in a single platform to predict multiple diseases. The simple input forms and easy-to-understand results make this system accessible to anyone, even without medical knowledge ([6][7][8][9][10]).

V. CONCLUSION

This paper reviewed the machine learning techniques used for predicting diabetes, heart disease, and high blood pressure. The findings show that Random Forest is effective for diabetes detection, Logistic Regression and Cascaded Neural Networks work well for heart disease prediction, and Decision Tree and Random Forest perform well for blood pressure classification. Workplace screening programs analyzed using K-Means clustering have demonstrated that regular health checkups help reduce hypertension over time.

The proposed disease prediction system successfully integrates all three diseases into a single platform with a simple workflow. The system allows users to register, log in, enter basic health details, receive instant predictions with risk levels, and save their history for future tracking. Testing results showed that the system correctly identified low-risk and high-risk cases for all three diseases and provided appropriate medical advice.

Future work can include adding more diseases, improving the models with larger and more diverse datasets, deploying the system as a mobile application for wider reach, and incorporating explainable AI techniques to help users understand why a particular prediction was made. Overall, machine learning offers a promising and accessible approach for early disease detection, helping people take control of their health before serious problems develop.

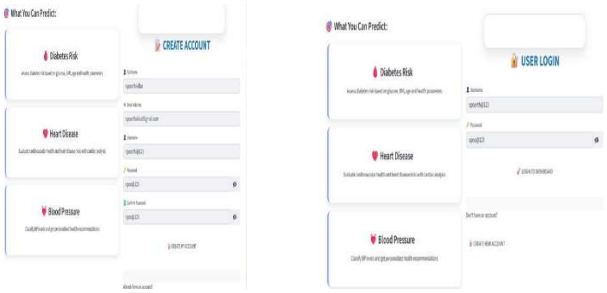

Fig. 1 Login dashboard of disease prediction website

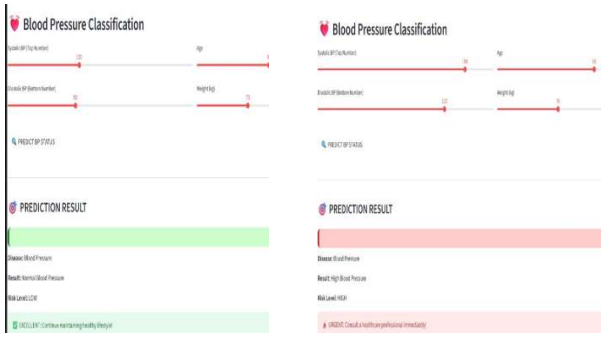

Fig. 2 Example of BP prediction using disease prediction website.

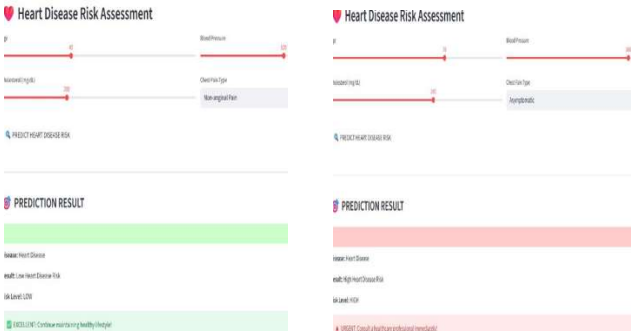

Fig. 3 Example of heart disease prediction using disease prediction website

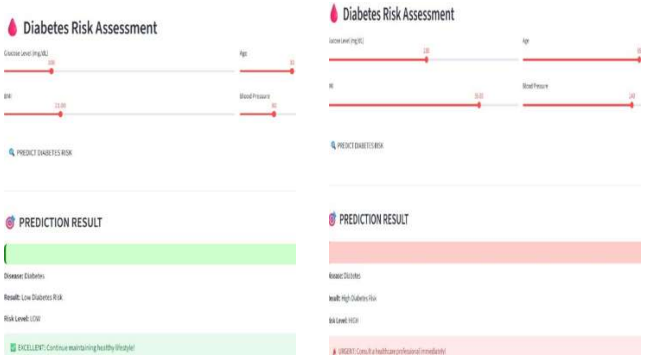

Fig. 4 Example of diabetes prediction using disease prediction website

 ©IJSRED: All Rights are Reserved

ACKNOWLEDGMENT

I would like to express my sincere gratitude to all those who supported me in completing this disease prediction system project. First and foremost, I thank my professors and guides for their valuable guidance, encouragement, and constructive feedback throughout this work, as their insights helped me overcome many challenges and improve the quality of this project. I am also grateful to my family and friends for their constant support and motivation, as their belief in me kept me going during difficult times. I extend my thanks to the researchers whose published papers I referred to in this work, as their contributions to the field of machine learning and healthcare inspired many of the ideas implemented in this project. Finally, I thank all the users who tested the system and provided useful feedback for improvement. Thank you all.

REFERENCES

[1] Iparraguirre-Villanueva, O., Espinola-Linares, K., Flores Castañeda, R. O., & Cabanillas-Carbonell, M. (2023). Application of machine learning models for early detection and accurate classification of type 2 diabetes. Diagnostics, 13(14), 2383.

[2] El-Sofany, H., Bouallegue, B., & Abd El-Latif, Y. M. (2024). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Scientific Reports, 14(1), 23277.

[3] Adeleke, O., Adebayo, S., Awordine, H., Adeleke, O., Adeniyi, A. E., & Aroba, O. J. (2024). Machine learning evaluation of a hypertension screening program in a university workforce over five years. Scientific Reports, 14(1), 1-10.

[4] Shrivastava, A., Chakkaravarthy, M., & Shah, M. A. (2023). A new machine learning method for predicting systolic and diastolic blood pressure using clinical characteristics. Healthcare Analytics, 4, 100219.

[5] Chitra, R., & Sevenivasagam, V. (2015). Heart disease prediction system using supervised learning classifier. International Journal of Computer Science and Engineering, 2(2), 1-7.

[6] Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. IEEE/ACS International Conference on Computer Systems and Applications, 108-115.

[7] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. ICT Express, 7(4), 432-439.

[8] Tsoi, K. K. F., et al. (2020). Machine learning clustering for blood pressure variability applied to systolic blood pressure intervention trial (SPRINT) and the Hong Kong Community Cohort. Hypertension, 76, 569-576.

[9] *Theerthagiri, P., Ruby, A. U., & Vidya, J. (2023). Diagnosis and classification of the diabetes using machine learning algorithms. SN Computer Science, 4, 72.*

[10] Fahlman, S. E., & Lebiere, C. (1990). The cascaded correlation learning architecture. In Neural Information Processing Systems, Morgan Kaufmann Publishers, 524-532

 ©IJSRED: All Rights are Reserved