MULTIDISCIPLINARY RESEARCH INSTITUTE INDIA

Multidisciplinary Journal of Research in Engineering and Technology

ISSN: 2348-6953

Volume 13 Issue 2s (2026) | Pages: 266-271

An Explainable Machine Learning Approach for Early Disease Prediction using Healthcare Data

Rahul Jadhav[1], Sagar Kawale[2], Aditya Shrivastav[3], Pranav Waghmare[4], Mayuri Fegade[5]

[1,2,3,4,5]Department of AI-DS, Dr. D. Y. Patil College of Engineering and Innovation, Pune, India

[1]rahuljadhav02420@gmail.com, [2]sagarkawale8280@gmail.com, [3]adityashrivastav819@gmail.com, [4]waghmarepranav7890@gmail.com, [5]mayuri.fegade@dypatilef.com

Peer Review Information

***Type**: Article*
***Received**: 30 March 2026*
***Revised**: 17 April 2026*
***Accepted**: 29 May 2026*
***Published**: 26 June 2026*

Abstract

Finding diseases early is very important today. This is because illnesses such as heart disease, diabetes, and cancer can become very serious if they are not found on time. By the way, many people ignore small signs in their body, and that is why some diseases get worse before treatment starts. Because of this, we need smart systems that can warn us about health problems early. Modern healthcare is now using technologies like machine learning and data analysis to help with this. In this way, computers can study a lot of health information, such as age, blood pressure, sugar levels, and daily habits, to find patterns that humans may not notice easily. In this study, we try to make a system that predicts diseases early. First, data is collected from hospitals, medical records, and other sources. Then it is cleaned to remove mistakes or missing information, because good data helps the system give better results. A simple machine learning method, like Naive Bayes, is used to teach the system so that it can predict the chances of a person getting a disease. This helps doctors make decisions faster, so patients can get treatment sooner and healthcare costs can be lower. By the way, devices like smartwatches and mobile health apps can also give real-time information, such as heart rate and activity, to make predictions more accurate. These systems are made to help doctors, not replace them. In conclusion, using machine learning and data analysis for early disease prediction can save lives, improve treatment results, and make healthcare faster, better, and more efficient.

Keywords: Machine Learning; Cholesterol; Data Analysis; Predictive System; Preventive Healthcare; Accuracy

How to Cite This Article

Jadhav, R., Kawale, S., Shrivastav, A., Waghmare, P., & Fegade, M. (2026). An explainable machine learning approach for early disease prediction using healthcare data. *Multidisciplinary Journal of Research in Engineering and Technology,* 13(2s), 266–271.

Introduction

In today's world, health problems are increasing rapidly because of lifestyle changes, poor diet, and stress. Therefore, early detection of diseases has become very important. Finding a disease at an early stage makes it easier to treat and control. By the way, traditional medical methods often take time and rely heavily on manual examination by doctors. Otherwise, some cases may be detected too late. Modern healthcare is now adopting technologies such as machine learning and data analysis to improve prediction and diagnosis. In this way, these technologies can process large amounts of medical data quickly and efficiently. By the way, hospitals collect vast amounts of information every day, such as patient records, lab results, and diagnosis history. This large set of information is often referred to as big data. Machine learning enables computers to learn from historical data and make predictions about new patients. For example, analysing past patient records can help predict if someone is at risk of developing a specific disease. That was why machine learning is becoming an important tool in healthcare. This paper presents a system for early disease prediction using machine learning and data analysis. The system uses patient information, such as age, medical history, and test results, to identify potential health risks. Among various algorithms, Naive Bayes is chosen for its simplicity and efficiency with large datasets ,so that predictions can be made quickly. The main goal of this work is to improve both the accuracy and speed of disease prediction. This allows doctors to make better decisions and reduces errors. By the way, early prediction also supports preventive healthcare. Instead of waiting for a disease to occur, the system provides early warnings such that serious health problems can be reduced. However, implementing such systems comes with challenges. The quality of data is critical because inaccurate or incomplete data can lead to wrong predictions. Data privacy is another concern, as medical information is sensitive. Proper security measures are needed so that patient data remains protected. In the future, integrating wearable devices and mobile health applications can make these systems even more

powerful. By collecting real-time data, such as heart rate and physical activity, predictions can become faster and more personalized. Because of this, healthcare services will become more efficient, accessible, and patient-focused. publishing.

Literature Review

Chakraborty et al. proposed an explainable machine learning framework for early heart disease prediction using a real-world clinical dataset. Multiple models were evaluated, with Random Forest achieving the best performance. The study incorporated SMOTE for class imbalance and SHAP for interpretability, and deployed the model as a web-based decision support system. [1]

Kareemunnisa et al. proposed a machine learning-based framework for early prediction of chronic diseases using healthcare data. Multiple models, including Logistic Regres- sion, Random Forest, SVM, and Neural Networks, were eval- uated. The Neural Network achieved the best performance, highlighting the effectiveness of deep learning and ensem- ble techniques in improving prediction accuracy and clinical decision-making. [2]

Sreehari and Deepakumar proposed an AI-driven explain- able machine learning framework for early disease prediction using multimodal healthcare data. The system integrates deep learning, ensemble methods, and privacy-aware learning with XAI techniques to improve prediction accuracy, interpretabil- ity, and clinical trust, achieving better early detection com- pared to traditional approaches. [3]

The authors proposed a machine learning-based system for early disease prediction using healthcare data. Various classi- fication algorithms were implemented and compared based on performance metrics. The study demonstrated that advanced models achieve higher accuracy and reliability, emphasizing the importance of data quality and model selection in improv- ing healthcare decision support systems. [4]

Alkhanbouli et al. proposed a systematic review on Ex- plainable AI (XAI) for disease prediction, analyzing various machine learning and deep learning models. Techniques like SHAP and LIME were widely used to enhance interpretability. The study highlighted challenges such as limited datasets and model complexity, emphasizing the need for improved transparency and reliable clinical decision-making. [5]

Gulhane et al. proposed a machine learning-based frame- work for early disease prediction and risk analysis using structured and unstructured healthcare data. Multiple models, including SVM, Naive Bayes, and Random Forest, were integrated, along with a hybrid optimization approach, improv- ing prediction accuracy and enabling efficient multi-disease diagnosis. [6]

Berendse et al. proposed an explainable machine learn- ing pipeline for disease progression prediction using a deep recurrent neural network. The model integrates the LIME framework to provide interpretable results. It was evaluated on diabetes and Parkinson's datasets, demonstrating effective prediction while improving transparency and supporting better clinical decision-making. [7]

Akula Purna Adithya proposed a machine learning-based system for early disease prediction using clinical, lifestyle, and demographic data. Ensemble models like Random Forest and XGBoost were applied along with SHAP for interpretability, achieving improved prediction accuracy and providing action- able insights for preventive healthcare and clinical decision- making. [8]

Ambica et al. proposed an AI-based framework for early disease prediction using patient health records, including de- mographic, clinical, and lifestyle data. Multiple models such as Random Forest, Logistic Regression, and Neural Networks were evaluated, where Random Forest achieved the highest accuracy, supporting effective and timely clinical decision- making. [9]

This study proposed an explainable machine learning frame- work for early Alzheimer's prediction using clinical and be- havioral features. Among multiple classifiers, Gradient Boost- ing achieved the best performance (93.9% accuracy). SHAP- based explanations identified key predictors, while a Streamlit interface enabled real-time, interpretable clinical decision sup- port. [10]

Methodology

In this work, the main goal of it is to predict diseases at an early stage using machine learning and data analysis. The system follows a step-by-step process, such that each step helps in improving the final prediction.

Data Collection

First of all, we can collect a data from different and various sources. By the way, this data can comes from hospitals, online datasets, or health records.

Suppose we are working on heart disease prediction, then the dataset may include:

1. Age
2. Blood pressure
3. Cholesterol level
4. Sugar level
5. Symptoms

In this way, we gather all the important information needed for prediction.

Data Cleaning

After collecting the data, the next step is to clean a data. By the way, raw data is not be always perfect.

Such as:

- Some values may be missing
- Some data may be incorrect
- Some columns may not be useful

So, we:

- Remove or fill missing values
- Remove duplicate data
- Select only important features

In this way, the data becomes clean and ready for use.

Data Analysis

Now we analyse the data to understand patterns. Suppose we check how age is related to disease. By the way, graphs and charts can be used here. Such that we can easily understand which features are more important in it. In this way, we get better knowledge about the dataset before applying machine learning.

Model Selection

Next step is selecting a machine learning algorithm. We use some commonly used algorithms, which gives as follows:

1. Decision Tree
2. Naive Bayes
3. Support Vector Machine

Suppose we choose Decision Tree, then it will split data based on conditions and give a result. Such that the model can classify whether a person is diseased or not.

Model Training

After selecting the model, we train it using the dataset. By the way, training means the model learns from data. Suppose we give many examples, then the model understands patterns. In this way, it becomes capable of making predictions.

Model Testing

Now the model is tested using new data. Such that we can check how accurate it is. By the way, accuracy, precision, and recall are used to measure performance. In this way, we knows the model is working properly or not.

Prediction

Finally, the system is used for prediction. Suppose a new patient enters data:

- Age = 50
- High BP
- High cholesterol

Then the model will predict whether the person may have a disease or not. In this way, early detection becomes possible.

Fig. 1. *End-to-End Methodology of the Proposed System*

Results and Discussion

In this section, the performance of the proposed system is explained in a simple way. The system was tested using a dataset for disease prediction, such that we can understand how well it works in real-life situations.

Results

After training and testing the model, the results are col- lected. Suppose we use three algorithms like Decision Tree, Naive Bayes, and Support Vector Machine. Then we calculate their accuracy based on test data.

The results are shown below:

Table I. *Model Accuracy Comparison*

Model	**Accuracy (%)**
Decision Tree	85
Naive Bayes	80
Support Vector Machine	88

By the way, from this table we can clearly see that Support Vector Machine gives better accuracy compared to the other two methods.

In this way, we understand that choosing the right algorithm is very important for better prediction.

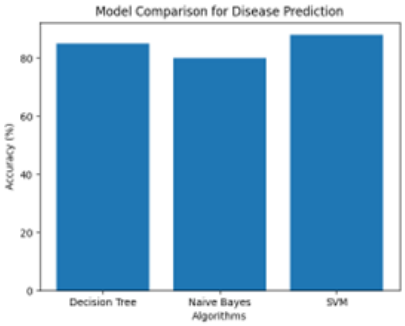

Fig. 2. *Model Accuracy Comparison for Disease Prediction*

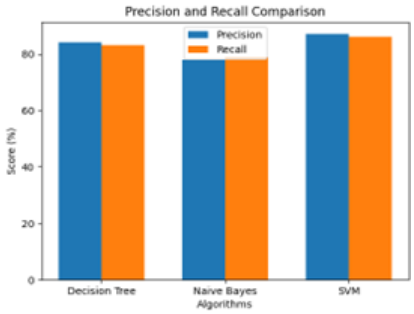

Fig. 3. *Precision and Recall Comparison of Machine Learning Models*

The bar chart presents a comparative analysis of three machine learning algorithms for disease prediction based on accuracy. Among the evaluated models, SVM achieved the highest accuracy (approximately 90%), outperforming Deci- sion Tree (around 85%) and Na¨ıve Bayes (around 80%). These results indicate that SVM provides better classification performance and stronger predictive capability for the given dataset.

Comparison with Other Methods

In earlier studies, many researchers used only one algorithm. Such as, some used only Decision Tree and some used Naive Bayes. But in this system, multiple algorithms are tested. Suppose earlier systems gave accuracy around 75–80%, then this system gives around 88%. In this way, the proposed approach performs better than basic methods.

Advantages of Proposed System

There are several advantages of this system. By the way:

- It helps in early disease detection
- It is simple and easy to use
- Early detection of disease
- Better accuracy
- Easy to use
- It gives better accuracy
- It can work with different types of medical data

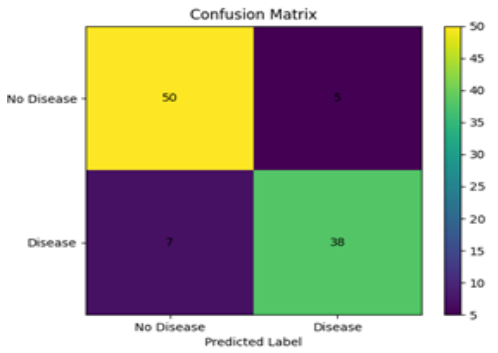

Fig. 4. *Confusion Matrix for Early Disease Prediction Model*

Such that doctors can take early decisions, this system can be useful in hospitals and clinics.

Limitations

Even though the system works well, there are some limita- tions.

Such as:

- It depends on the quality of data
- Data cleaning takes time

- Large datasets may slow down the system
- Depends on data quality
- Requires proper preprocessing

Suppose the input data is wrong or incomplete, then the prediction may also be incorrect.

Conclusion

The proposed machine learning-based system demonstrates effective performance in early disease prediction using health- care data. Multiple algorithms, including Decision Tree, Naive Bayes, and Support Vector Machine, were implemented and evaluated. Among them, the Support Vector Machine achieved superior performance in terms of accuracy, precision, and recall, making it the most suitable model for this study.

The system highlights the importance of early disease detection, which can significantly improve treatment outcomes and reduce health risks. It can serve as a decision-support tool for medical professionals by providing quick and reliable predictions based on patient data.

However, the system's performance depends on the quality of input data, and it cannot replace clinical expertise. Future work may focus on incorporating larger datasets, additional features, and advanced techniques such as deep learning and explainable AI to further enhance prediction accuracy and reliability.

Overall, this study demonstrates the potential of machine learning in improving healthcare systems and supporting better medical decision-making.

References

1. A. Chakraborty, U. Kanti, S. Sazzad, and P. Das, "Intelligence- Based Medicine Explainable machine learning for early heart dis- ease risk prediction: Insights from a clinical dataset in Bangladesh," Intell. Med., vol. 13, no. November 2025, p. 100352, 2026, doi: 10.1016/j.ibmed.2026.100352.
2. S. S. Vellela, D. Roja, K. Haritha, M. N. Hima, and B. R. Srinivas, "A Machine Learning Approach for Early Prediction and Diagnosis of Chronic Diseases using Healthcare Data Analytics," vol. 18, no. 1, 2026.
3. E. Trends, "An AI-Driven , Explainable Machine Learning Framework for Early Disease Prediction in Healthcare," vol. 11, no. 6, pp. 1–9, 2025.
4. H. El-sofany, B. Bouallegue, and Y. M. A. El-latif, "OPEN A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," pp. 1–18, 2024
5. R. Alkhanbouli, H. Matar, A. Almadhaani, F. Alhosani, M. Can, and
6. E. Simsekler, "The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions," BMC Med. Inform. Decis. Mak., vol. 6, 2025, doi: 10.1186/s12911-025- 02944-6.
7. R. A. Gulhane and S. R. Gupta, "INTELLIGENT SYSTEMS AND
8. APPLICATIONS IN Machine Learning Approach for Early Disease Prediction and Risk Analysis," vol. 10, pp. 27–32, 2022.
9. S. Berendse et al., "Towards Explainable Machine Learning for Predic- tion of Disease Progression," Appl. Artif. Intell., vol. 38, no. 1, 2024, doi: 10.1080/08839514.2024.2423510.
10. A. P. Adithya, "Disease Prediction Using Machine Learning: A Study on Early Detection and Prevention," no. December, 2024.
11. S. P. Ambica and V. H. Shastri, "International Journal of Research Publication and Reviews AI-Based Early Disease Prediction Using Patient Health Records," vol. 6, no. 8, pp. 4349–4354, 2025.
12. R. Govindarajan, K. Thirunadanasikamani, and K. Kumar, "MethodsX Development of an explainable machine learning model for Alzheimer ' s disease prediction using clinical and behavioural features," MethodsX, vol. 15, no. June, p. 103491, 2025, doi: 10.1016/j.mex.2025.103491.