IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Research Article)

Check for updates

Symptom-Based Disease Recognition and Test Recommendation Using XGBoost

Prashreet Sharma Majagaiyan *, Kotla Kanaka Mahalakshmi Kavya, S.Y. Surya Venkata Durga Prasad, Komarthi Arun Kumar and U.P Kumar Chaturvedula

Department of Computer Science and Engineering, Aditya collage of Engineering and Technology, Surampalem, India.

International Journal of Science and Research Archive, 2026, 18(03), 398-403

Publication history: Received on 17 January 2026; revised on 02 March 2026; accepted on 04 March 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.18.3.0433

Abstract

Timely diagnosis of a disease is essential to improve the condition of the patient; however, due to the limited access to medical professionals, it often delays the early diagnosis of a health condition. This paper presents a machine learning based Disease Recognition and Test Recommendation System to help the user in predicting any possible disease based on the symptoms reported and perform any appropriate diagnostic test to confirm the diagnosis. The proposed system uses supervised learning algorithms such as Random Forest and XGBoost that are trained using public symptom and disease data. User given symptoms are converted into binary feature vectors and fed to classification model to detect probable diseases with high level of accuracy. A rule-based mapping module is further integrated to recommend the relevant medical tests corresponding to the predicted conditions and fill the gap between the preliminary self-assessment and professional healthcare consultation. The system is tested with standard performance parameters such as accuracy and precision and shows reliable and interpretable results. This work adds a scalable and easy to use decision support system that helps with the early detection of disease and encourages preventive healthcare.

Keywords: Disease Recognition; Machine Learning; Ensemble Learning; Test Recommendation; Xgboost

1. Introduction

In recent times, the field of Machine Learning has received massive boost and applications in various domain specially in the field of healthcare. The ability of the machine learning models to predict the disease early and accurately is essential for the delivery of effective healthcare. However, due to low availability of the medical professionals often leads to delay in consultation and diagnosis. In this respect, the use of machine learning (ML) has been proven to be a promising technique for the analysis of data in the medical field and for the purpose of predictive diagnostics.

Recent development in machine learning has led to the development of models which can predict the disease based on the symptoms reported by the user. But these models use traditional machine learning algorithms like KNN, SVM, Naive Bayes etc having low accuracy. Moreover, these models focus on predicting the disease only, and do not integrate with recommendations on the diagnostic test that are important to confirm medical conditions. To overcome this problem, in this research a Disease Recognition and Test Recommendation system is purposed. The system utilizes algorithms based on ensemble learning such as XGBoost which are trained using public available symptoms-disease dataset.

The system proposed is to offer an accessible, interpretable and scalable solution to make decisions that will aid healthcare actions at early stages. By combining predictive analytics with diagnostic advice, this work is a step towards bettering efficiency, accessibility and preventive care in healthcare systems today.

[*] Corresponding author: Prashreet Sharma Majagaiyan.

2. Related works

The application of machine learning (ML) techniques in healthcare has experienced an incredible growth in the last two decades, which allows better disease prediction, diagnosis, and clinical decision support. Early attempts made by Kononenko [7] discussed the capabilities of ML in medical diagnosis and the history of ML, referring to the problems of incorporating intelligent systems in healthcare. Deo [10] and Wiens and Shenoy [11] stressed even more on the transformative potential of ML in medicine, pointing the ability to process large-scale electronic health records (EHRs) for predictive analytics.

Several studies have been done in the use of ensemble learning techniques and especially decision tree-based models for the prediction of disease. Random Forests has been introduced by Breiman [2] and has high popularity due to its robustness and capability to work in high dimensional data with interpretability and confidence estimates [13]. Gradient boosting methods and XGBoost specifically have proven to be excellent for their scalability and power in many areas and this is discussed by Chen and Guestrin [1]. Friedman [3] formalized the theoretical underpinnings for gradient boosting, providing a framework for iterative function approximation with a minimization of predictive error -- a concept that has been used in modern healthcare applications to a huge extent.

Recent works have demonstrated the usefulness of ML algorithms in diseases recognition. Chen et al. [4] and Rajkomar et al. [5] discussed about the predictive modeling using big data and the usefulness of EHRs when it comes to predicting disease. Beam and Kohane [6] highlighted the importance of utilizing large scale data to gain information from clinical data, and Caruana et al. [8] developed intelligible models to predict risk of hospital readmission and pneumonia, indicating the need for interpretable ML in clinical decision making. Clinical decision support systems (CDSS) reviewed by Wang et al. [9] is one of the critical applications of ML in healthcare intervention guidance.

Several surveys and comparative studies have been done specifically on the prediction of the disease using ML. Agrawal and Choudhary [14], Shanmugam et al. [15] and Pingale et al. [19], reviewed a series of algorithms, focusing in XGBoost, Random Forests, Support Vector Machines (SVMs) and hybrid models for accurate disease detection. Dua and Graff [16] made a widely used datasets from UCI repository available for reproducibility and benchmarking. Similarly, Kourou et al. [12] discussed the applications of ML methods in cancer prognosis and reflects the prediction power of ensemble methods. More recent researches [20]- [24] improved on these findings by using multi-disease prediction along with combination of the symptom-based features and ensemble classifiers for better performance.

XGBoost in particular has become cutting edge algorithm for tasks of recognizing diseases. Its gradient boosting framework is supportive to the use of skewed data and complex non-linear relationships in that it performs better than conventional classifiers according to several studies [1], [21], [22]. In terms of prediction of heart disease, ensemble and hybrid approaches of XGBoost combined with other ML models have shown greater accuracy as well as reduction of false positive rates [25]-[28]. Comparative analyses of ML models in healthcare also confirm the ML superior predicting power of XGBoost and the computational efficiency [29], [30].

In summary, literature has agreed that machine learning and especially XGBoost, is effective to predict disease and provide clinical decision support. The integration of such models with intelligent systems for test recommendations can help improve early diagnosis, resource allocation and ultimately improve patient outcomes. This collection of works serves as a good foundation on the creation of an automated Disease Recognition and Test Recommendation System with the predictive power of XGBoost on patient datasets.

3. Methodology

The proposed system is shown in figure I. In this paper, we propose a framework for classifying human diseases given the symptoms by machine learning and ensemble learning algorithm such as XGBoost. The dataset that contains multiple diseases with their symptoms was collected and processed using techniques such as label encoding to make the model more efficient.

This model can predict more than one disease depending on the symptoms reported by the patient. Further the symptoms are weighted in order to increase the accuracy of the model. In this system the user enters the symptoms and then these symptoms are fed to the machine model to predict the machine learning model about the disease. And the model also suggests the appropriate medical test to be conducted to establish the disease predicted. The algorithm for the prediction of the disease is XGBoost algorithm.

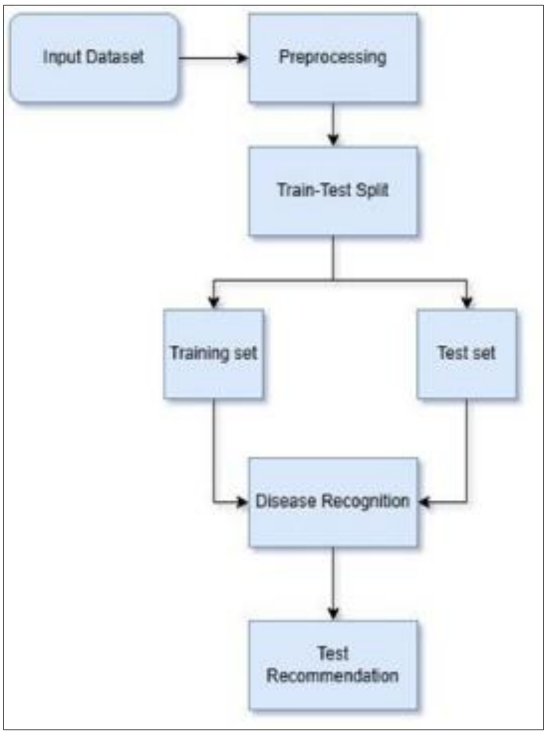

Figure 1 Proposed System

4. Experiment and result

4.1. Data Collection

The datasets that were utilized in this research work are accessible publicly on Kaggle. The dataset consists of mapping between diseases and their symptoms. The dataset consists of more than 40 diseases and their symptoms. There are about 5000 instances of data in the dataset. The dataset contains nominal data having disease name and the symptoms.

4.2. Preprocessing

Preprocessing step is needed in order to increase the model reliability and minimize noise. This step involves data cleaning, feature selection and encoding. As the dataset is the categorical data, they are transformed to the binary feature vector in which each symptom in it is marked by the value of 1 in case of its appearance and 0 in case of the absence. Missing value processing and feature selection in which uncommon and irrelevant features are eliminated are also part of the preprocessing stage.

4.3. Model Selection and Training

To improve the accuracy of the model ensemble learning technique called XGBoost is used.

XGBoost is an acronym of eXtreme Gradient Boosting. It is the ensemble learning method that integrates multiple weak models into a strong model. XGBoost takes decision trees as its base learners and integrates them sequentially to enhance the performance of the model. The trees are trained one after the other to rectify the mistakes of the last tree and this is known as boosting. It constructs additive decision trees sequentially by minimizing a regularized objective function

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, \overline{y_i}) + \sum_{k=1}^{K} \Omega(f_k)$$

Where l is the loss function and $\Omega(fk)$ is the regularization term.

4.4. Disease Prediction

Once the user feeds the system with the symptoms by the interface, the system transforms the input into a binary feature vector and passes it to the trained model. The classifier gives the highest likelihood disease and the probability scores that are associated and this gives the estimation of confidence.

4.5. Test Recommendation Module

When the disease is predicted, the rule-based mapping module suggests applicable diagnostic tests that are related to the disease that has been predicted. This mapping is built based on the conventional medical sources and guarantees meaningful clinically suggestions.

4.6. Performance Evaluation

The standard classification measures applied in the assessment of the model performance are accuracy, precision, recall, and F1-score. The confusion matrices are also studied to determine the reliability of the classification to various disease classes. We can obtain accuracy, precision, recall and F1-score as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 - Measure = \frac{2 \times precision \times recall}{precision + recall}$$

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

4.7. Comparison with Existing System

The accuracy of the different existing models for disease recognition is compared and shown in Table 1. The proposed model achieved higher accuracy as compared to traditional machine learning models with the accuracy of 98.5%. It shows that the proposed approach can identify the disease more accurately than earlier methods.

Table 1 Results with applied models

Model	Accuracy
KNN [22]	88%
Weighted KNN [21]	93.5%
Naïve Bayes [23]	89.4%
Proposed System	98.5%

5. Conclusion

In this paper, a complete framework of the symptom-based prediction of diseases was presented through machine learning and ensemble learning methods. XG Boost was trained on a structured symptoms-disease dataset that was obtained in publicly available healthcare repositories. The results of the experiment showed that the XGBoost model was the most accurate in classifications because of its gradient boosting optimization and regularization attributes and also presented good generalization performance and immunity to overfitting.

Preprocessing methods including label encoding, feature weighting, and feature representation in a structured manner were applied before training the model to increase the model's predictive performance. The weighted symptom features enabled the models to reflect on the relative significance of the key symptoms to identify the disease. The ensemble approach made use of the strengths of several decision-tree based learners thereby enhancing stability and predictive power as opposed to conventional standalone classifiers.

Besides the classification of the disease, the proposed system also incorporated a rule based medical test recommendation system which relates the predicted diseases to diagnostic tests associated with the diseases. This complementary integration gets this gap in prediction and actionable healthcare guidance. The framework was able to outperform and have better practical applicability in comparison with the current symptom-based prediction models by applying ensemble learning along with structured rule-based reasoning.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

References

[1] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785– 794.

[2] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[3] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” Annals of Statistics, vol. 29, no. 5, pp. 1189– 1232, 2001.

[4] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from healthcare communities,” IEEE Access, vol. 5, pp. 8869– 8879, 2017.

[5] S. Rajkomar et al., “Scalable and accurate deep learning with electronic health records,” npj Digital Medicine, vol. 1, 2018.

[6] A. Beam and I. Kohane, “Big data and machine learning in health care,” JAMA, vol. 319, no. 13, pp. 1317–1318, 2018.

[7] I. Kononenko, “Machine learning for medical diagnosis: History, state of the art and perspective,” Artificial Intelligence in Medicine, vol. 23, no. 1, pp. 89–109, 2001.

[8] R. Caruana et al., “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission,” in Proc. 21st ACM SIGKDD, 2015.

[9] K. D. Wang et al., “Clinical decision support systems: A review,” Journal of Biomedical Informatics, vol. 64, pp. 145–156, 2016.

[10] A. Deo, “Machine learning in medicine,” Circulation, vol. 132, no. 20, pp. 1920–1930, 2015.

[11] J. Wiens and E. Shenoy, “Machine learning for healthcare: On the verge of a major shift,” npj Digital Medicine, vol. 1, 2018.

[12] P. Kourou et al., “Machine learning applications in cancer prognosis and prediction,” Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, 2015.

[13] S. Wager, T. Hastie, and B. Efron, “Confidence intervals for random forests,” Journal of Machine Learning Research, vol. 15, pp. 1625–1651, 2014.

[14] A. Agrawal and A. Choudhary, “A survey on disease prediction using machine learning,” International Journal of Computer Applications, vol. 178, no. 32, 2019.

[15] M. S. Shanmugam et al., “Disease prediction system using machine learning,” International Journal of Engineering and Advanced Technology, vol. 8, no. 6, 2019.

[16] D. Dua and C. Graff, “UCI machine learning repository,” University of California, Irvine, 2019.

[17] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning. New York, NY, USA: Springer, 2013.

[18] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. New York, NY, USA: Springer, 2009.

[19] K. Pingale, S. Surwase, V. Kulkarni, S. Sarage, and A. Karve, “Disease prediction using machine learning,” International Research Journal of Engineering and Technology (IRJET), vol. 6, pp. 831–833, 2019.

[20] I. Ibrahim and A. Abdulazeez, “The role of machine learning algorithms for diagnosing diseases,” Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 10–19, 2021.

[21] R. Keniya et al., “Disease prediction from various symptoms using machine learning,” SSRN, 2020. doi: 10.2139/ssrn.3661426.

[22] P. Parshant and A. Rathee, “Multiple Disease Prediction Using Machine Learning,” IRE Journals, 2023. [Online]. Available: https://www.irejournals.com/formatedpaper/17 04650.pdf

[23] C. K. Gomathy, “The prediction of disease using machine learning,” 2021.

[24] R. Patel and A. Singh, “Disease Prediction Using Machine Learning Algorithms,” International Journal of Computer Applications, vol. 183, no. 25, pp. 10–, 2021.

[25] U. P. Rajput and S. Bhise, “Intelligent heart disease prediction system using data mining techniques,” International Journal of Computer Applications, vol. 173, no. 4, pp. 26–29, 2019.

[26] S. Ahmed et al., “A comparative study on heart disease prediction using machine learning algorithms,” International Journal of Engineering and Technology, vol. 7, no. 4.30, pp. 313–318, 2018.

[27] V. Chaurasia and S. Pal, “Data mining– based diagnosis of heart disease using classification,” Journal of Information and Knowledge Management, vol. 11, no. 2, pp.1450027, 2012.

[28] S. S. Alam et al., “An efficient disease prediction model using hybrid machine learning techniques,” International Journal of Advanced Research in Computer Science, vol. 10, no. 4, 2019.

[29] F. Javed and S. Samad, “Performance analysis of machine learning algorithms for disease prediction,” International Journal of Computer Science and Information Security (IJCSIS), vol. 17, no. 5, pp. 57–64, 2019.

[30] P. Saxena et al., “Machine learning models for disease prediction: Performance review and analysis,” International Journal of Computer Applications, vol. 174, no. 5, pp. 15–22, 2020.