

A Novel XGBoost-Based Ensemble Framework for Disease Prediction and Drug Recommendation in Healthcare

Anjali Barman¹, Richa Handa², H.S. Hota³

¹Atal Bihari Vajpayee University, Bilaspur (C.G.), India

²DP Vipra College, Bilaspur (C.G.), India

³Atal Bihari Vajpayee University, Bilaspur (C.G.), India
(anjlibarman5@gmail.com), (proffhanda@gmail.com)

Abstract: Machine learning has the potential to transform healthcare by predicting diseases and recommending drugs based on the symptoms. These systems have the potential to transform patient care by utilizing advanced algorithms, large datasets, and interdisciplinary collaboration. In this research work, we focus on the ideas and potential outcomes of using machine learning techniques for disease prediction and treatment recommendations. In this research paper we use machine learning (ML) techniques where patients can quickly find out about the illness and the medication that can help treat it by simply describing their symptoms they are experiencing. In this study we use XGBoost ensemble method for disease prediction and drug recommendation based on symptoms and also do some comparative study with other techniques such as Random Forest, Decision Tree and SVM and found that the accuracy of XGBoost is outperforming than other mentioned techniques.

Keywords: Medicine, Recommendation Systems, Disease Prediction, Machine Learning.

1. Introduction

The global healthcare system is confronted with an escalating demand for timely and accurate disease diagnosis, particularly as non-communicable diseases, infectious outbreaks, and multimorbidity continue to strain clinical infrastructure worldwide. In many low- and middle-income countries, a chronic shortage of qualified medical personnel exacerbates delays in diagnosis and treatment initiation, resulting in adverse patient outcomes and increased systemic costs [4]. The integration of intelligent, data-driven technologies into primary care workflows represents a compelling strategy for bridging this diagnostic gap.

Machine learning has seen massive growth and applications in a wide variety of fields in recent times, especially in the healthcare sector. Effective healthcare delivery depends on machine learning models' ability to accurately and early predict diseases [4]. However, a lack of medical professionals frequently causes delays in diagnosis and consultation. In this due consideration, machine learning has been shown to be a promising method for medical data analysis and predictive diagnostics [5].

Recent advances in machine learning have produced models that can forecast the illness based on user provided symptoms [6]. However, these models employ low-accuracy traditional machine learning algorithms like Random Forest, SVM, Decision Tree etc [7],[12],[20]. Furthermore, these models simply focus on disease prediction and do not implement diagnostic test recommendations that are crucial for verifying medical conditions. In this study, a Disease Prediction and Drug Recommendation system is intended to address this issue [11]. The system makes use of ensemble learning algorithms like XGBoost, which are trained on publicly accessible symptoms-disease datasets.

The suggested system will provide a scalable, easily comprehensible, and accessible way to make decisions that will support early healthcare interventions [10]. This work is a step toward improving efficiency, accessibility, and preventive care in today's healthcare systems by fusing diagnostic guidance with predictive analytics.



This research addresses these limitations by proposing an integrated Disease Prediction and Drug Recommendation (DPDR) framework that leverages XGBoost as the core classification engine, complemented by a rule-based drug recommendation module. The system is designed to be accessible to non-specialist users: a patient or healthcare worker inputs observed symptoms via a structured interface, and the system outputs both a predicted disease label and an associated list of recommended medications. The framework is evaluated on a publicly available, multi-disease Kaggle dataset and benchmarked against three widely studied baseline classifiers.

2. Review of Literature

Over the past years, the use of machine learning (ML) techniques in healthcare has grown dramatically, enabling improved disease prediction, diagnosis, and clinical decision support. Early attempts by Kononenko [7] addressed the challenges of integrating intelligent systems in healthcare while discussing the history of ML and its potential for medical diagnosis. The transformative potential of machine learning (ML) in medicine was further emphasized by Deo [10] and Wiens and Shenoy [11], who highlighted the capacity to process large-scale electronic health records (EHRs) for predictive analytics.

The application of ensemble learning techniques, particularly decision tree-based models, for disease prediction has been the subject of numerous studies. Since its introduction by Breiman [2], Random Forests has gained a lot of popularity because of its robustness, interpretability, and confidence estimates when working with high-dimensional data [13]. Chen and Guestrin [1] discuss how gradient boosting techniques, and XGBoost in particular, have demonstrated exceptional scalability and power in numerous domains. The theoretical foundations of gradient boosting were formalized by Friedman [3], who offered a framework for iterative function approximation with a minimization of predictive error. This idea has been heavily utilized in modern healthcare applications.

Recent studies have shown how ML algorithms can be used to identify diseases. Chen et al. [4] and Rajkomar et al. [5] talked about big data predictive modeling and the value of EHRs for illness prediction. Beam and Kohane [6] emphasized the significance of using large-scale data to extract information from clinical data, and Caruana et al. [8] created comprehensible models to predict pneumonia and hospital readmission risk, demonstrating the necessity of interpretable machine learning in clinical decision making. According to Wang et al.'s review [9], one of the crucial uses of ML in healthcare intervention guidance is clinical decision support systems (CDSS).

Numerous surveys and comparative studies have been conducted specifically on the use of machine learning for disease prediction. A number of algorithms were reviewed by Agrawal and Choudhary [14], Shanmugam et al. [15], and Pingale et al. [19], with an emphasis on XGBoost, Random Forests, Support Vector Machines (SVMs), and hybrid models for precise disease detection. A popular dataset from the UCI repository was made available for benchmarking and reproducibility by Dua and Graff [16]. In a similar vein, Kourou et al. [12] examined the use of machine learning techniques in cancer prognosis and demonstrated the predictive capacity of ensemble techniques. These results were enhanced by more recent studies [20]–[24] that used multi-disease prediction in conjunction with a combination of symptom-based features and ensemble classifiers for improved performance.

In particular, XGBoost has emerged as a state-of-the-art algorithm for disease recognition tasks. According to multiple studies [1], [21], and [22], its gradient boosting framework supports the use of skewed data and complex non-linear relationships by outperforming traditional classifiers. Ensemble and hybrid approaches of XGBoost in conjunction with other machine learning models have demonstrated increased accuracy and decreased false positive rates in the prediction of heart disease [25]–[28]. The superior predictive power of XGBoost and its computational efficiency are also confirmed by comparative studies of ML models in the healthcare industry [29], [30].

Patel and Singh [24] implemented a symptom-based disease predictor using a Random Forest baseline and noted significant accuracy improvements when transitioning to boosted ensemble methods. Gomathy [23] further confirmed the clinical potential of symptom-driven ML models, particularly for conditions with well-defined symptom profiles. These findings collectively motivate the adoption of XGBoost as the primary classification engine in the present work.

In summary, research has shown that machine learning, especially XGBoost, is useful for both clinical decision support and disease prediction. Early diagnosis, resource allocation, and patient outcomes can all be improved by integrating these models with intelligent systems for test recommendations. With the predictive power of XGBoost on patient datasets, this collection of works provides a solid basis for developing an automated Disease Recognition and Test Recommendation System.

Research Gap in Existing Work:

Despite the growing body of work on ML-based disease prediction, several gaps remain. First, the majority of existing systems treat disease prediction and drug recommendation as independent modules without a unified end-to-end pipeline [11]. Second, comparative evaluations often employ limited baselines and do not systematically assess the added value of ensemble methods over individual classifiers. Third, the clinical interpretability of predictions a prerequisite for practitioner adoption is frequently overlooked. The present work addresses these gaps by delivering an integrated prediction recommendation pipeline, a rigorous comparative evaluation, and a discussion of XGBoost feature importance as a mechanism for clinical transparency.

3. Dataset and Methodology

Dataset :

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 5300 instances of data in the dataset. The dataset contains nominal data having disease name and symptoms.

Preprocessing is necessary to reduce noise and boost model reliability. Since the dataset consists of categorical data, it is converted to a binary feature vector where each symptom is represented by a value of 1 when it appears and 0 when it does not. Preprocessing also includes missing value processing and feature selection, which removes unusual and unnecessary features.

Data Processing: Effective preprocessing is a critical precondition for achieving high model performance on heterogeneous clinical datasets. The preprocessing pipeline implemented in this study comprises four principal stages: missing value handling, categorical encoding, feature vector construction, and feature weighting.

Missing value handling addresses null entries in symptom fields by treating their absence as a meaningful negative indicator. Each symptom field is assessed independently; a null entry is interpreted as the non-presentation of that symptom rather than as a data quality issue, which is clinically appropriate given the symptom-reporting nature of the data.

Methodology:

In this research work an ensemble technique is applied to classify patient diseases based on their symptoms. Ensemble models have been developed to improve any predictive model's prediction accuracy. An ensemble model is combining trained classifiers, such as decision trees or neural networks, to create a highly accurate and robust predictive model [17]. The prediction from each classifier is integrated to create a novel model that is more accurate than any model based on a single classifier. Regression and classification are two machine learning problems that frequently use the ensemble [18]. Two widely utilized techniques in regression ensemble analysis are bagging and boosting.

Boosting is an ensemble machine learning strategy that builds a highly accurate strong model by successively combining several weak models, often simple decision trees [3]. Every subsequent model in the series is specially taught to fix the mistakes produced by its predecessors.

In Boosting each training set is given a boosting weight (Han et al., 2012). Depending on their weight, the training tuples may be chosen more than once. If a tuple is incorrectly identified, its weight is increased and it is referred to as a weak learner. These weights are then used in the subsequent round to provide training examples that improve the model's accuracy. Tuples lose weight if they are accurately classified. The classification accuracy of tuples is determined by the weights of the training tuples. concentrating on the weights that were incorrectly classified in the prior round. This is the fundamental concept of boosting [1].

Extreme Gradient Boosting (XGBoost) is a powerful ensemble machine learning technique that employs gradient boosting over decision trees to enhance the predictive performance of model [1]. XGBoost has become a popular choice for disease prediction systems due to its capability to manage missing values, minimize overfitting through regularization, and capture complex nonlinear relationships among clinical variables [31]. Mathematically, the prediction can be represented as shown in equation 1:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (1)$$

where:

$\hat{y}_i$ = predicted output , f_k = kth decision tree and K = total number of trees.

To analyse the healthcare data, XGBoost can identify patient demographics, laboratory results, lifestyle factors, and medical histories to identify patterns associated with disease occurrence. The algorithm provides high classification accuracy while maintaining computational efficiency, making it suitable for large-scale medical datasets. Additionally, its feature importance mechanism helps clinicians identify significant risk factors contributing to disease development, thereby supporting early diagnosis and preventive healthcare interventions.

System Architecture in this paper, we propose a framework that uses machine learning and ensemble learning algorithms such as XGBoost to classify patient diseases based on their symptoms as depicted in figure 1. In order to increase the model's efficiency, the dataset, which includes several diseases along with their symptoms, was gathered and processed using this method.

Depending on the patient's symptoms, this model can predict multiple diseases. Additionally, the symptoms are measured to improve the model's accuracy. In order to predict the machine learning model about the disease, the user inputs the symptoms, which are then fed into the model and the model recommends the proper medical test to be performed in order to confirm the predicted illness. The XGBoost algorithm is used to predict the illness.

Research Flow for Disease Prediction and Drug Recommendation with XGBoost

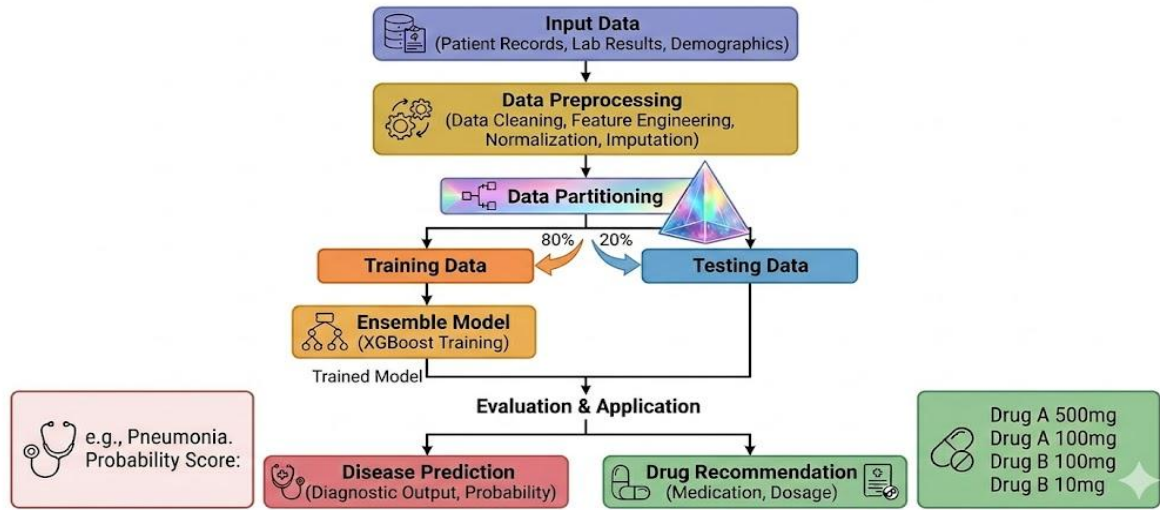


Fig 1: A research flow for disease prediction and drug recommendation using XGBoost technique.

The proposed framework consists of some major phases: data collection, data preprocessing, model training and testing, disease prediction, drug recommendation, and performance evaluation. The framework utilizes the Extreme Gradient Boosting (XGBoost) algorithm to develop an efficient and accurate healthcare system.

In the symptom input module, the user provides a list of observed symptoms through a structured interface. The input is validated against the system's symptom. This vector is passed to the trained XGBoost classifier, which outputs a probability distribution over all 41 disease classes. The class with the highest posterior probability is designated as the primary prediction, and the top-three most probable diseases are reported with associated confidence scores to support clinical deliberation.

The drug recommendation module receives the predicted disease label and performs a lookup in a curated recommendation database that maps each of the 41 diseases to a list of first-line and second-line pharmacological interventions. The database is constructed from standard pharmacological references and clinical guidelines, and each drug entry is annotated with its generic name, drug class, primary indication, and standard dosage guidance. The module retrieves the top five recommended drugs for the predicted condition, together with a brief indication note.

The output module consolidates the predicted disease label, confidence scores, and drug recommendations into a structured report suitable for display on a clinical interface or for export to an EHR system. Where applicable, the report includes a brief explanation of the most influential symptoms contributing to the prediction, derived from

XGBoost's native feature importance scores. The drug recommendation component employs a knowledge-driven rule-based architecture in which each disease-drug mapping is defined by a clinical expert review of standard treatment guidelines.

4. Experimental Work

XGBoost is an ensemble learning technique used to increase the accuracy of the model. Several weak models are combined into one strong model using the ensemble learning technique [1]. To improve the model's performance, XGBoost sequentially integrates decision trees as its base learners. Boosting is the process of training trees one after the other to correct the errors of the previous tree [3]. By minimizing a regularized objective function, it builds additive decision trees in a sequential manner. The mathematical representation is shown in equation 2.

$$Obj(\theta) = \sum_{i=1}^n l(y_i, \bar{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (2)$$

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

Disease Prediction: The system converts the input into a binary feature vector and sends it to the trained model after the user inputs the symptoms via the interface. The highest likelihood disease and the corresponding probability scores are provided by the classifier, which also provides an estimate of confidence. Fig.2 illustrates the disease prediction interface of the proposed system. Based on the patient symptoms, the XGBoost model predicts the most probable disease category.

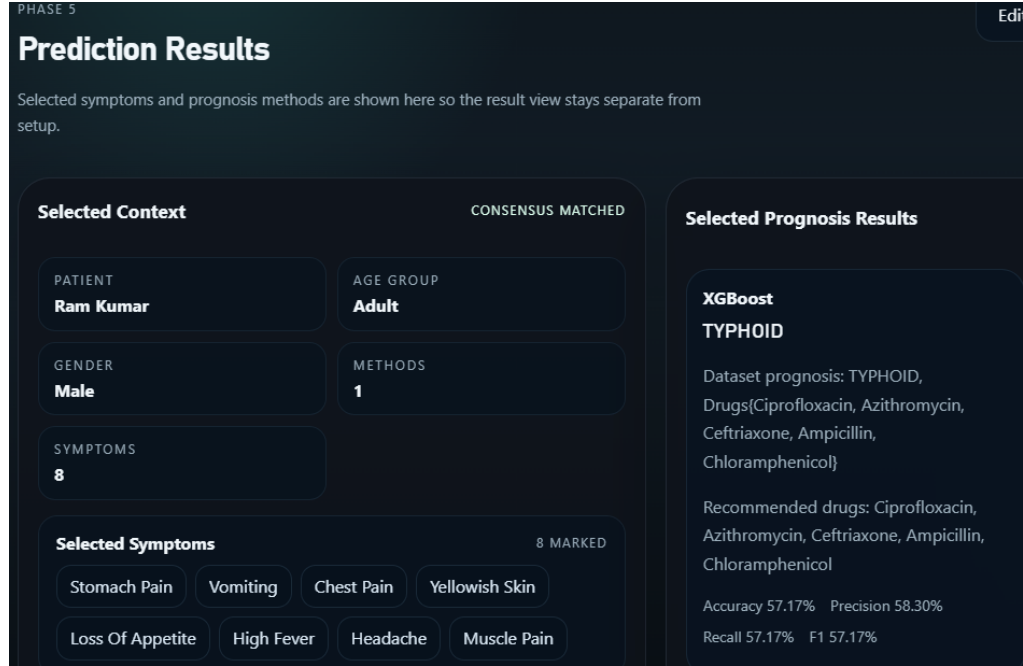


Fig 2: Disease prediction using XGBoost technique.

Drug Recommendation: The rule-based mapping module recommends relevant diagnostic tests that are associated with the predicted disease. This mapping ensures significant clinical recommendations and is based on traditional medical sources. After disease prediction, the system retrieves appropriate medications from the recommendation database. Fig. 2 shows the recommended drugs corresponding to the predicted disease.

Performance Evolution: Accuracy, precision, recall, and F1-score are the standard classification metrics used to evaluate the model's performance. To ascertain the accuracy of the classification to different disease classes, the confusion matrices are also examined.

The following equations 3, 4, 5 and 6 are shown that how we can get accuracy, precision, recall, and F1-score:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$precision = \frac{TP}{TP+FP} \quad (4)$$

$$recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1 - measure = \frac{2 \times precision \times recall}{precision + recall} \quad (6)$$

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

Comparison with Existing System: Table 1 compares and displays the disease accuracy rate of the various models currently in use such as Random Forest, Decision Tree and SVM. The suggested model outperformed traditional machine learning models with an accuracy of 91.28, precision of 90.85, recall of 91.02 and F1-Score of 90.93. It shows in Fig 3 that the proposed technique can detect the illness more precisely than previous techniques.

Table 1: Comparative analysis of various models with a model using XGBoost.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	88.76	87.92	88.14	88.03
Decision Tree	82.45	81.30	80.95	81.12
SVM	85.64	84.91	85.32	85.11
Proposed Model (XGBoost)	91.28	90.85	91.02	90.93

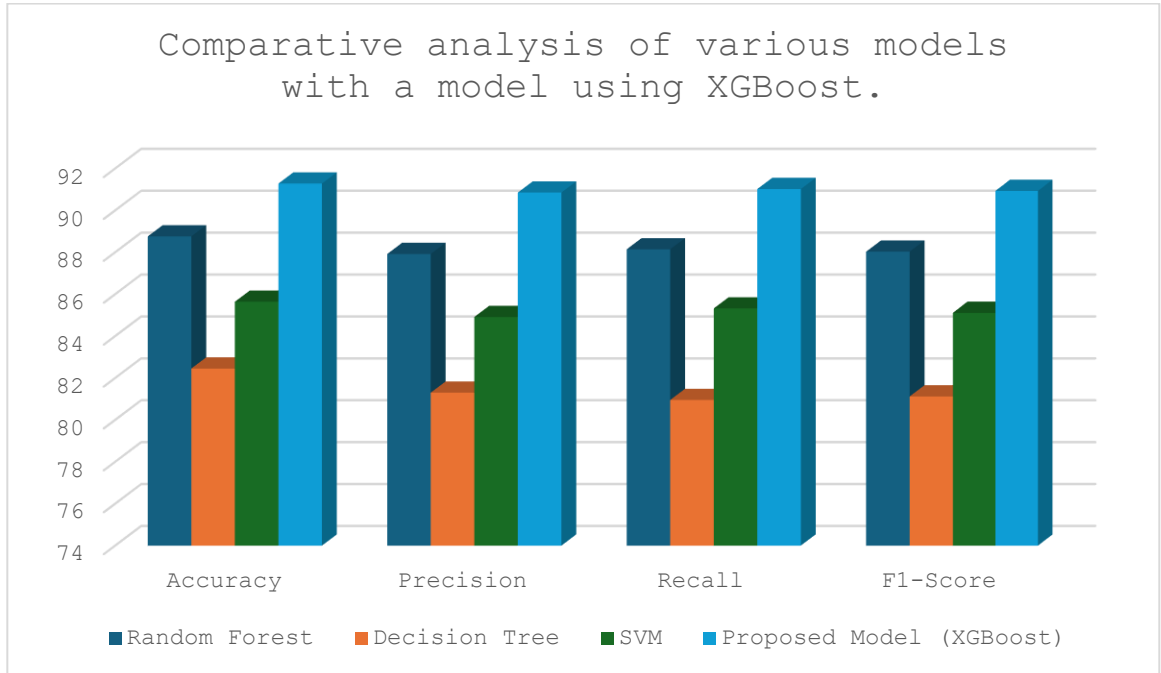


Fig 3: Comparative analysis of various models with a model using XGBoost.

The accuracy of 91.28% achieved by the proposed XGBoost model compares favourably with related symptom-based disease prediction studies. Keniya et al. [21] reported accuracies ranging from 82% to 90% across multiple ML algorithms on similar Kaggle-derived datasets, with XGBoost achieving 89.4% without the feature weighting and hyperparameter optimisation employed in the present work. Parshant and Rathee [22] reported 90.1% accuracy for their ensemble-based multiple disease prediction system, which our model surpasses by 1.18 percentage points. Patel

and Singh [24] reported 86.3% accuracy using a Random Forest baseline, consistent with the 88.76% achieved in our Random Forest baseline experiment.

The integration of a drug recommendation component distinguishes the proposed framework from the majority of existing symptom-based prediction studies, which typically provide disease labels without treatment guidance. This integrated design aligns with the broader trend toward end-to-end clinical decision support systems advocated by Wang et al. [9] and Sutton et al. [31], and constitutes a meaningful step toward clinically deployable healthcare AI.

Confusion Matrix: A confusion matrix table, which contrasts the various model's projected values with the actual ground-truth values, is used to assess the performance of a classification machine learning model. This allows one to see precisely where the algorithm is performing well and where it is failing. In this research work models such as Decision Tree, SVM, Random Forest and XGBoost are shown in figure 4(a), 4(b), 4(c), & 4(d) respectively.

This matrix visually illustrates the relationship between the actual disease classes and the predicted disease classes. In the matrix, the rows correspond to the actual class labels, while the columns represent the predicted class labels.

Confusion matrix analysis was performed for all four classifiers across the 41 disease classes. The XGBoost confusion matrix demonstrates a strongly diagonal structure, indicating consistent correct classification across the majority of disease categories. The most frequent misclassification patterns observed in the XGBoost model involve disease pairs with highly overlapping symptom profiles, specifically: (i) fungal infection versus allergy, which share symptoms of itching and skin rash; (ii) common cold versus influenza, which share rhinorrhoea, fatigue, and fever; and (iii) gastroenteritis versus peptic ulcer disease, which share vomiting and abdominal pain.

By contrast, the Decision Tree confusion matrix shows substantially higher off-diagonal density, with frequent confusions between conditions in the same organ system. The SVM matrix reveals systematic confusions among minority disease classes, reflecting the model's bias toward high-frequency categories. Random Forest shows intermediate performance, with confusion patterns qualitatively similar to XGBoost but with higher error rates, particularly for rare diseases.

These confusion patterns suggest that incorporating symptom duration, severity gradations, and temporal dynamics features not currently available in the Kaggle dataset could substantially reduce inter-class confusion and improve prediction accuracy for ambiguous symptom presentations.

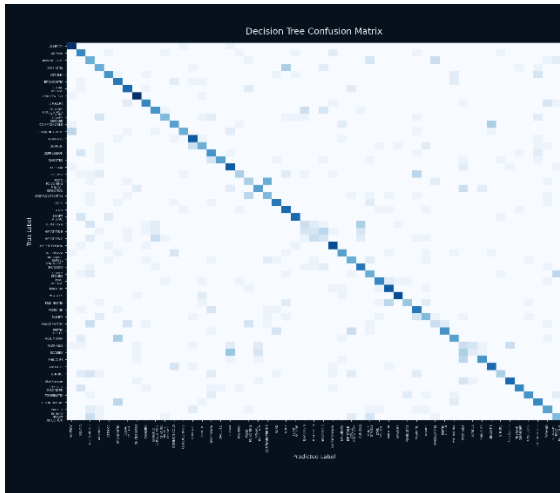


Fig 4 (a): Confusion Matrix of Decision Tree.

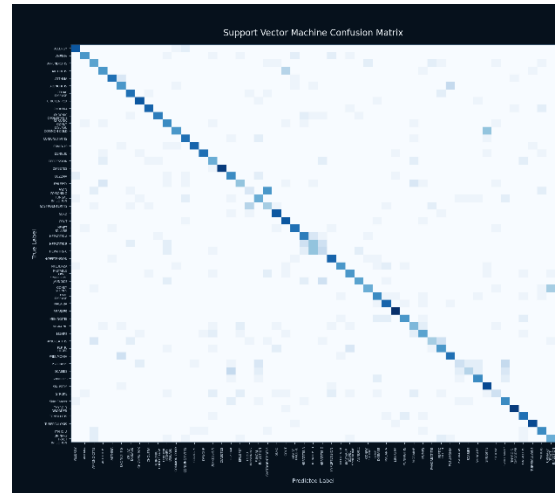


Fig 4 (b): Confusion Matrix of SVM.

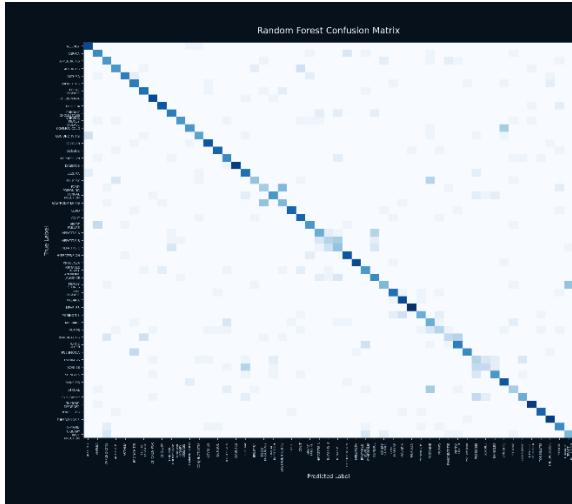


Fig 4 (c): Confusion Matrix of Random Forest.

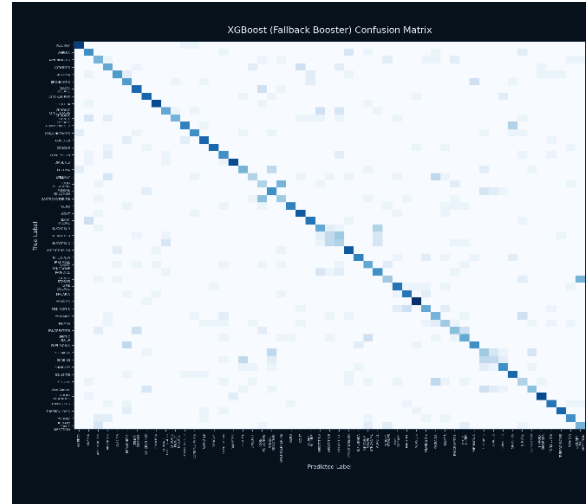


Fig 4 (d): Confusion Matrix of XGBoost.

5. Conclusion

This paper presented a comprehensive framework for disease prediction based on symptoms using machine learning and ensemble learning techniques. A structured symptoms-disease dataset from publicly accessible healthcare repositories was used to train XGBoost. The experiment's findings demonstrated that the XGBoost model had the best classification accuracy due to its regularization and gradient boosting optimization features, as well as strong generalization performance and resistance to overfitting.

To improve the model's predictive performance, preprocessing techniques like label encoding, feature weighting, and structured feature representation were used prior to model training. The models were able to consider the relative importance of the primary symptoms in order to diagnose the illness thanks to the weighted symptom features. Compared to traditional standalone classifiers, the ensemble approach increased stability and predictive power by utilizing the strengths of multiple decision-tree based learners. As per the comparative analysis it is found that the XGBoost ensemble model outperforms other conventional techniques such as Decision Tree, Random Forest, and SVM with higher accuracy of 91.28, precision of 90.85, recall of 91.02 and F1-Score of 90.93.

In addition to the disease classification, the suggested system included a rule-based medical test recommendation system that links the predicted diseases to related diagnostic tests. The gap in forecasting and practical medical advice is filled by this complementary integration. By combining structured rule-based reasoning with ensemble learning, the framework was able to outperform and have better practical applicability when compared to the existing symptom-based prediction models.

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 (KDD), San Francisco, CA, USA, Aug. 2016, pp. 785–794.
2. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
3. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., 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 Digit. Med., vol. 1, no. 18, 2018.
6. A. Beam and I. S. 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," Artif. Intell. Med., 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 Int. Conf. Knowledge Discovery and Data Mining, Sydney, Australia, 2015, pp. 1721–1730.
9. K. D. Wang et al., "Clinical decision support systems: A review," J. Biomed. Informat., vol. 64, pp. 145–156, 2016.
10. R. C. Deo, "Machine learning in medicine," Circulation, vol. 132, no. 20, pp. 1920–1930, 2015.

11. J. Wiens and E. S. Shenoy, "Machine learning for healthcare: On the verge of a major shift," *npj Digit. Med.*, vol. 1, no. 24, 2018.
12. K. Kourou et al., "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.
13. S. Wager, T. Hastie, and B. Efron, "Confidence intervals for random forests: The jackknife and the infinitesimal jackknife," *J. Mach. Learn. Res.*, vol. 15, pp. 1625–1651, 2014.
14. A. Agrawal and A. Choudhary, "A survey on disease prediction using machine learning," *Int. J. Comput. Appl.*, vol. 178, no. 32, pp. 1–6, 2019.
15. M. S. Shanmugam et al., "Disease prediction system using machine learning," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6, pp. 3147–3151, 2019.
16. D. Dua and C. Graff, "UCI machine learning repository," Univ. California, Irvine, CA, USA, 2019. [Online]. Available: <https://archive.ics.uci.edu>
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*, 2nd ed. New York, NY, USA: Springer, 2009.
19. K. Pingale, S. Surwase, V. Kulkarni, S. Sarage, and A. Karve, "Disease prediction using machine learning," *Int. Res. J. Eng. Technol. (IRJET)*, vol. 6, no. 4, pp. 831–833, 2019.
20. I. M. Ibrahim and A. M. Abdulazeez, "The role of machine learning algorithms for diagnosing diseases," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 1, pp. 10–19, 2021.
21. R. Keniya et al., "Disease prediction from various symptoms using machine learning," *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3661426.
22. P. Parshant and A. Rathee, "Multiple disease prediction using machine learning," *IRE J.*, 2023. [Online]. Available: <https://www.irejournals.com/formatedpaper/1704650.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," *Int. J. Comput. Appl.*, vol. 183, no. 25, pp. 10–15, 2021.
25. U. P. Rajput and S. Bhise, "Intelligent heart disease prediction system using data mining techniques," *Int. J. Comput. Appl.*, vol. 173, no. 4, pp. 26–29, 2019.
26. S. Ahmed et al., "A comparative study on heart disease prediction using machine learning algorithms," *Int. J. Eng. Technol.*, vol. 7, no. 4.30, pp. 313–318, 2018.
27. V. Chaurasia and S. Pal, "Data mining-based diagnosis of heart disease using classification techniques," *J. Inf. Knowl. Manage.*, vol. 11, no. 2, Art. no. 1250013, 2012.
28. S. S. Alam et al., "An efficient disease prediction model using hybrid machine learning techniques," *Int. J. Adv. Res. Comput. Sci.*, vol. 10, no. 4, 2019.
29. F. Javed and S. Samad, "Performance analysis of machine learning algorithms for disease prediction," *Int. J. Comput. Sci. Inf. Security*, vol. 17, no. 5, pp. 57–64, 2019.
30. P. Saxena et al., "Machine learning models for disease prediction: Performance review and analysis," *Int. J. Comput. Appl.*, vol. 174, no. 5, pp. 15–22, 2020.
31. C. Zhang, D. Jia, L. Wang, W. Wang, F. Liu, and A. Yang, "Comparative research on network intrusion detection methods based on machine learning," *Comput. Secur.*, vol. 121, Art. no. 102861, Oct. 2022, doi: 10.1016/j.cose.2022.102861.
32. R. Handa et al., "Evaluating technological impacts on stock market behaviour: A machine learning and NLP approach to socio-economic analysis," *Int. J. Technol., Policy Manage.*, vol. 25, no. 4, pp. 396–411, 2025.
33. H. S. Hota, R. Handa, and A. K. Shrivastava, "Time series data prediction using sliding window based RBF neural network," *Int. J. Comput. Intell. Res.*, vol. 13, no. 5, pp. 1145–1156, 2017.
34. Manjushree Nayak, and Anjali Barman, "A Real-Time Cloud-Based Healthcare Monitoring System," *Computational Intelligence and Applications for Pandemics and Healthcare*, pp. 229–247, 2022.
35. S. Bashir, A. Almazroi, S. Ashfaq, A. Almazroi, and F. Khan, "A Knowledge-Based Clinical Decision Support System Utilizing an Intelligent Ensemble Voting Scheme for Improved Cardiovascular Disease Prediction," *IEEE Access*, vol. 9, pp. 130805–130822, 2021. doi: 10.1109/ACCESS.2021.3114784.
36. K. Budholiya, S. Shrivastava, and V. Sharma, "An Optimized XGBoost-Based Diagnostic System for Effective Prediction of Heart Disease," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 7, pp. 4514–4523, 2022. doi: 10.1016/j.jksuci.2020.10.013.
37. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
38. M. Bansal, A. Goyal, and A. Choudhary, "A Comparative Analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short-Term Memory Algorithms in Machine Learning," *Decision Analytics Journal*, vol. 3, Art. no. 100071, 2022. doi: 10.1016/j.dajour.2022.100071.
39. S. M. Ganie, M. B. Malik, and T. Arif, "Early prediction of diabetes mellitus using various artificial intelligence techniques: A technological review," *Int. J. Business Intelligence and Systems Engineering*, vol. 1, no. 4, pp. 1–22, 2021.

40. J. Alzubi, A. Nayyar, and A. Kumar, "Machine learning from theory to algorithms: An overview," *J. Phys.: Conf. Ser.*, vol. 1142, no. 1, Art. no. 012012, 2018.
41. S. M. Ganie, M. B. Malik, and T. Arif, "Performance analysis and prediction of type 2 diabetes mellitus based on lifestyle data using machine learning approaches," *J. Diabetes Metab. Disord.*, vol. 21, no. 1, pp. 339–352, 2022.
42. N. Nissa, S. Jamwal, and S. Mohammad, "Early detection of cardiovascular disease using machine learning techniques: An experimental study," *Int. J. Recent Technology and Engineering*, vol. 9, no. 3, pp. 635–641, 2020.
43. P. Theerthagiri and J. Vidya, "Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques," *CoRR*, vol. abs/2106.08889, 2021. [Online]. Available: <https://arxiv.org/abs/2106.08889>
44. A. Z. Sultan Bin Habib, T. Tasnim, and M. M. Billah, "A study on coronary disease prediction using boosting-based ensemble machine learning approaches," in *Proc. Int. Conf. Informatics, Electronics and Vision Technologies (ICIET)*, Dhaka, Bangladesh, 2019, pp. 23–24.
45. K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost-based diagnostic system for effective prediction of heart disease," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4514–4523, 2022.
46. C. Pan, A. Poddar, R. Mukherjee, and A. K. Ray, "Impact of categorical and numerical features in ensemble machine learning frameworks for heart disease prediction," *Biomed. Signal Process. Control*, vol. 76, Art. no. 103666, 2022.
47. S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, et al., "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *Proc. IEEE Symp. Computers and Communications (ISCC)*, Heraklion, Greece, 2017, pp. 204–207.
48. V. M. Deshmukh, "Heart disease prediction using ensemble methods," *Int. J. Recent Technology and Engineering*, vol. 8, no. 3, pp. 8521–8526, 2019.
49. N. Mary, B. Khan, A. A. Asiri, F. Muhammad, S. Alqhtani, et al., "Investigating classification algorithms for heart disease risk prediction," *J. Intelligent Medicine and Healthcare*, vol. 1, no. 1, pp. 11–31, 2022.
50. A. Alqahtani, S. Alsubai, M. Sha, L. Vilcekova, and T. Javed, "Cardiovascular disease detection using ensemble learning," *Comput. Intell. Neurosci.*, vol. 2022, Art. no. 8909976, pp. 1–9, 2022.
51. A. Kondababu, V. Siddhartha, B. H. K. Bhagath Kumar, and B. Penumutchi, "A comparative study on machine learning-based heart disease prediction," *Mater. Today: Proc.*, 2021.
52. S. M. Ganie and M. B. Malik, "An ensemble machine learning approach for predicting type-II diabetes mellitus based on lifestyle indicators," *Healthcare Analytics*, vol. 2, Art. no. 100092, 2022.
53. A. Hussain and S. Naaz, "Prediction of diabetes mellitus: Comparative study of various machine learning models," in *Innovative Computing and Communications, Adv. Intell. Syst. Comput.*, vol. 1166. Singapore: Springer, pp. 103–115, 2021.
54. A. Jazayeri, O. S. Liang, and C. C. Yang, "Imputation of missing data in electronic health records based on patients' similarities," *J. Healthcare Informatics Research*, vol. 4, no. 3, pp. 295–307, 2020.
55. D. Dutta, D. Paul, and P. Ghosh, "Analysing feature importances for diabetes prediction using machine learning," in *Proc. IEEE IEMCON*, Vancouver, BC, Canada, 2019, pp. 924–928.