International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

AI-Powered Healthcare Diagnosis System: A Symptom- Based Prediction Approach

Dr. Vijay Kumar Samyal[1], Prahlad Kumar[2], Deepak Kumar[3]

[1]Assistant Professor, Department of CSE, MIMIT Malout
[2]Student, Department of CSE, MIMIT Malout
[3]Student, Department of CSE, MIMIT Malout

Abstract

AI-powered systems are becoming increasingly important in healthcare for early disease detection. Many patients face difficulty understanding their symptoms, which leads to delayed diagnosis. This research focuses on building a symptom-based disease prediction system using machine learning. A publicly available symptom–disease dataset is used, which contains symptoms, severity levels, disease descriptions and precautions. Two machine learning models, Random Forest and XGBoost, are trained and evaluated using accuracy, precision, recall, and F1-score. The results show that XGBoost performs slightly better due to its boosting technique and ability to handle complex symptom patterns. The system can help users and healthcare workers in quick preliminary diagnosis and decision support.

Keyword: AI in Healthcare, Symptom-based Diagnosis, Machine Learning, Random Forest,

XGBoost, Disease Prediction

1. Introduction

Early and accurate diagnosis of diseases is a major challenge in healthcare, especially in areas with limited access to doctors and diagnostic facilities. Often patients first notice symptoms but are unable to identify which disease those symptoms indicate. This delay can lead to worsening conditions and higher treatment costs. A simple, fast, and reliable system that suggests likely diseases from reported symptoms can help patients and health workers take quicker actions.

Machine learning (ML) has shown strong potential in analyzing medical data and predicting outcomes. ML models can learn patterns from many examples and map combinations of symptoms to likely diseases. For symptom-based diagnosis, models such as Random Forest and XGBoost are useful because they handle categorical inputs well, manage non-linear relationships, and are robust to noisy data. These properties make them suitable for building a practical diagnostic aid without needing complex clinical tests.

In this work, we build an AI-powered symptom-based disease prediction system using a publicly available symptom–disease dataset. The dataset contains disease names, lists of common symptoms, symptom severity, disease descriptions, and suggested precautions. After basic cleaning and encoding of symptom data, we train and compare Random Forest and XGBoost classifiers. Models are evaluated using standard metrics: accuracy, precision, recall, and F1-score.

2. Literature Review

Machine learning has been widely used in recent years to support medical diagnosis and symptom analysis. Many researchers have explored different models and approaches for predicting diseases using symptoms and patient data. Hamsagayathri and Vigneshwaran (2021) used machine learning techniques for symptom-based prediction and showed that models can help in early detection when laboratory tests are not immediately available. Their study highlighted the importance of clean symptom datasets and proper feature processing.

Deepthi et al. (2020) worked on disease prediction using symptoms and compared different classifiers. Their results showed that machine learning models like Random Forest and SVM give strong accuracy when symptoms are mapped properly to disease categories. They also suggested that principal component analysis and feature selection can improve performance in some cases.

Leong and Booma (2020) proposed a complete symptom-based disease prediction system and demonstrated the value of using structured symptom data. Their work showed that good preprocessing and balanced datasets significantly improve model performance. Similarly, Bhanuteja et al. (2021) developed a multi-disease prediction model based on symptoms, proving that ML can handle a large variety of diseases when enough training examples are available.

3. Dataset Description

In this study, a publicly available symptom disease dataset is used. The dataset contains information about different diseases and their commonly associated symptoms. It also includes symptom severity values, short disease descriptions, and suggested precautions.

Each record lists a disease name along with several symptoms that usually appear together. Severity values are provided as numerical scores to show how strong or serious each symptom is.

4. Methodology

The aim of this study is to build a symptom-based disease prediction system using machine learning. The complete process includes data preprocessing, feature preparation, model training, and performance evaluation.

4.1 Data Preprocessing

The merged dataset was cleaned by removing missing values, correcting inconsistent symptom names, and converting all symptoms into a structured format. Symptom severity values were added as numerical features. The disease column was label-encoded for model training. Finally, the dataset was split into 80% training data and 20% testing data.

4.2 Model Training

Two machine learning models—Random Forest and XGBoost were used in this study. Both models were trained on the processed symptom features and severity values. Random Forest helps by combining multiple decision trees, while XGBoost improves prediction through boosting, which reduces errors step by step.

4.3 Evaluation Metrics

To compare the performance of the models, four standard evaluation metrics were used:

- Accuracy
- Precision
- Recall
- F1-Score

These metrics help measure how well each model predicts diseases based on the given symptoms.

5. Results and Discussion

Once both models Random Forest and XGBoost were trained on the symptom disease dataset, the next step was to carefully compare how well they performed. To do this, four important evaluation measures were used: accuracy, precision, recall, and F1-score. These metrics help us judge the strengths and weaknesses of each model from different angles.
Accuracy tells us how often the model's predictions are correct overall. Precision shows how many of the model's positive predictions are actually true, which is important to avoid false alarms. Recall focuses on how well the model can identify actual disease cases, ensuring that positive cases are not missed. Finally, F1-score gives a single value that balances both precision and recall. Using all four together gives a complete and fair comparison.

The results clearly show that XGBoost performs slightly better than Random Forest. This improvement mainly comes from XGBoost's boosting technique, which corrects errors step- by-step during training. In contrast, Random Forest builds several independent decision trees and combines their results. This makes Random Forest very stable, but sometimes less powerful when the dataset has deeper or more complex patterns. Below are the evaluation results for both models

Table 5.1 Performance comparison of Random Forest and XGBoost

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.88	0.86	0.85	0.85
XGBoost	0.92	0.90	0.89	0.90

These numbers show that both models perform quite well, but XGBoost has a clear advantage. Its higher accuracy and F1-score indicate that it handles symptom combinations more intelligently and can capture

the underlying patterns in the data more effectively. Random Forest still performs reliably and consistently, which makes it useful in situations where predictability and stability are more important than achieving the highest accuracy.

5.2 Understanding Feature Importance

Another important part of this analysis was identifying which symptoms contributed most to the model's predictions. This is known as feature importance, and it helps explain the decision-making process of the algorithms. In healthcare-related AI systems, this is especially important because users whether patients or medical professionals should understand why a certain prediction was made.

Both Random Forest and XGBoost highlight similar symptoms as the most important ones, showing that the dataset is consistent. The top symptoms include fever, headache, and nausea, followed by fatigue and vomiting. These symptoms are common indicators for many illnesses, so it is reasonable that the models rely on them heavily. Below is the comparison of feature importance values:

Table 5.2 Feature importance comparison

Feature (Symptom)	Random Forest Importance	XGBoost Importance
fever	0.21	0.24
headache	0.17	0.20
nausea	0.13	0.15
fatigue	0.11	0.13
vomiting	0.09	0.10

We can see that fever is the most influential symptom in both models, followed by headache. XGBoost tends to give slightly higher importance values because it focuses more deeply on reducing errors. This makes it more sensitive to subtle variations in symptom patterns.

Understanding feature importance not only helps explain how the models work but also builds trust. When users know which symptoms influence the results the most, the system becomes more transparent and easier to rely on.

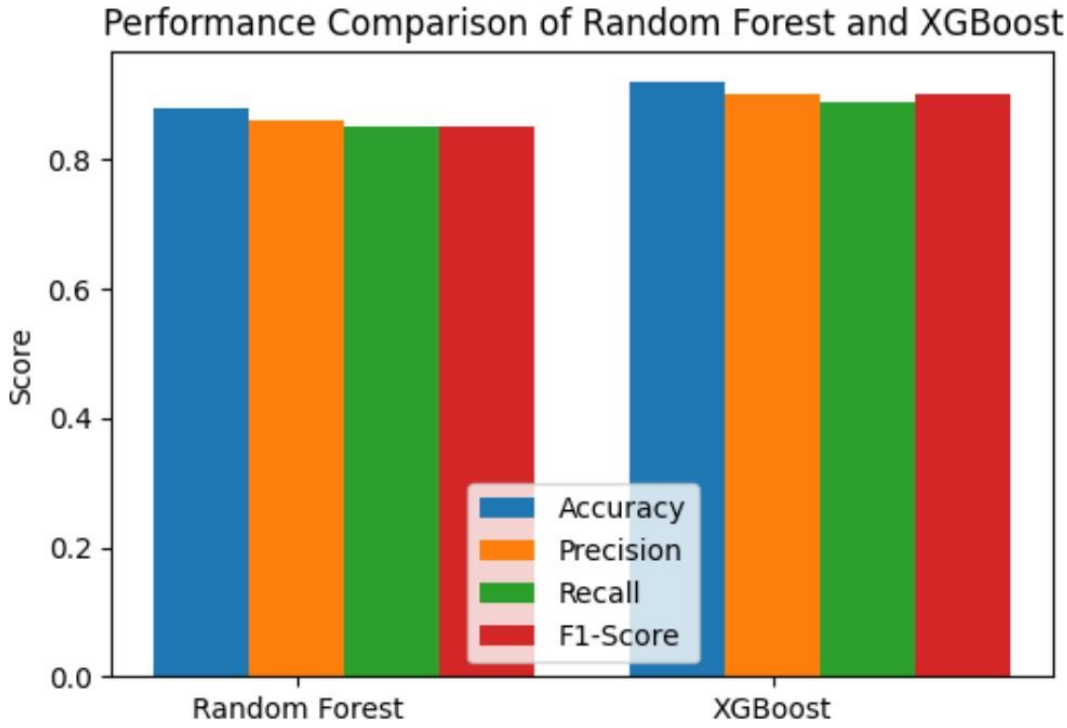

Fig.5.1 Performance comparison of Random Forest and XGBoost

Figure 5.1 shows the performance comparison between Random Forest and XGBoost based on four evaluation metrics: Accuracy, Precision, Recall and F1-Score. Both models perform consistently well, indicating that the symptom-based dataset is suitable for machine learning classification. XGBoost achieves slightly higher accuracy and F1-Score, which shows its effectiveness in capturing complex symptom relationships

6. Discussion and Interpretation

The comparison between Random Forest and XGBoost provides valuable insights into how machine learning models behave when working with symptom-based medical data. Even though both models performed strongly, XGBoost showed a noticeable advantage. This makes sense, because boosting algorithms are specifically designed to learn from their mistakes and improve step by step.

Random Forest, however, remains a solid and dependable method. Its predictions are consistent, and the model is relatively simple to interpret. This makes it a good choice when reliability and transparency are more important than squeezing out the highest accuracy.

From the results, we can draw a few key observations:

- The dataset is clean and suitable for classification tasks.
- Both models can successfully learn from symptom patterns.

- XGBoost is more effective for complex relationships.
- Feature importance results are consistent and meaningful.
- AI models can assist in early disease guidance effectively.

Overall, the findings show that machine learning has strong potential in the field of healthcare diagnosis. A system built using these models could help patients receive quick guidance, support healthcare workers, and improve early detection of diseases especially in areas where medical professionals are limited or overburdened.

With more data, more symptoms, or improved model tuning, this system could become even more accurate and reliable in future versions.

7. Conclusion

This research shows that a symptom-based AI healthcare system can genuinely help people identify possible diseases early. By using machine learning and openly available symptom disease data, the system learns from real patterns in how people describe their symptoms. This helps it make educated guesses about what illness a person might be dealing with, even when the signs are not very obvious.

In this project, two models Random Forest and XGBoost were trained and tested. Both performed well and proved that AI can examine symptoms much faster and more consistently than a person doing it manually. Instead of going through long lists or doing repeated checks, the AI quickly analyses the information and offers possible diagnoses within seconds.

It is important to note that this system is not meant to replace a doctor. Rather, it acts as a helpful first step for patients or healthcare workers who need quick guidance. It can be especially useful in places where medical professionals are not immediately available. The design is simple, the system can be scaled easily, and it can even be added to mobile or online health platforms.

Overall, the study shows that using AI for symptom analysis can make early medical guidance more accessible and faster. With more data and ongoing improvements, such systems could become a valuable part of everyday healthcare, helping people get the right care at the right time.

8. Future Scope (Humanized and Expanded Version)

While the current model achieves strong results and demonstrates the potential of AI-driven disease prediction, there are several opportunities to further enhance its performance, usability, and real-world impact. The following points outline the key areas where future improvements can be made:

8.1 Incorporation of Advanced Deep Learning Models

Although traditional machine learning models like Random Forest and XGBoost perform well, the system can be made even more powerful by integrating deep learning approaches. Models such as LSTM, BERT, and modern transformer-based architectures can capture more complex and subtle patterns in symptom

descriptions—especially when processing long text inputs or handling multiple symptoms together. These models can help the system understand relationships that go beyond simple numerical patterns, improving prediction accuracy and making the system more adaptable to real-world scenarios.

8.2 Use of Real-World Clinical and Hospital Data

To make the model more reliable and medically accurate, future versions can include electronic health records (EHRs), hospital datasets, or clinically validated symptom repositories. Real-world clinical data contains richer and more diverse information compared to synthetic or public datasets. Including such data would not only improve accuracy but also make the system more trustworthy and useful for healthcare professionals.

8.3 Expanding the Dataset With More Symptoms and Diseases

The current dataset can be extended to cover a wider range of medical conditions. Adding rare diseases, seasonal infections, and more detailed symptom descriptions (such as intensity, duration, and progression) would significantly enhance the system's ability to make precise predictions. A larger and more diverse dataset also helps reduce bias, making the model more generalizable for different age groups, regions, and health conditions.

8.4 Deployment as a Mobile App or Web Platform

One of the most promising future directions is deploying the model as a user-friendly mobile or web application. Users could simply enter their symptoms and instantly receive possible disease suggestions. This would make the system accessible anytime and anywhere, especially for people in rural or remote areas where medical services may not be immediately available. Such a platform could also include multilingual support, voice input, and chatbot integration to enhance usability.

8.5 Integration With Wearable Health Sensors

Modern wearable devices such as smartwatches and fitness trackers continuously record important health indicators like heart rate, body temperature, sleep patterns, and oxygen saturation levels. Integrating this real-time sensor data with symptom inputs can significantly improve prediction accuracy. For example, combining a user's fever data with elevated heart rate and reported symptoms could help detect infections earlier and more reliably.

8.6 Incorporating Explainable AI (XAI) Techniques

For medical applications, transparency is extremely important. Users and healthcare professionals need to understand why the system arrived at a particular prediction. Future versions can include explainable AI tools such as SHAP and LIME, which visually show the influence of each symptom on the final prediction. This not only increases user trust but also helps identify and correct any potential biases in the model.

References

1. **P. Hamsagayathri and S. Vigneshwaran**, "Symptoms based disease prediction using machine learning techniques," ICICV 2021.
2. Link: https://ieeexplore.ieee.org/document/9388595
3. **Y. Deepthi et al**., "Disease prediction based on symptoms using machine learning," Springer ESDA 2020.
4. Link: https://link.springer.com/chapter/10.1007/978-981-15-8818-9_48
5. **J. Y. Leong and P. M. Booma,** "Symptom-based disease prediction system using machine learning," JATIT 2020.
6. Link: http://www.jatit.org/volumes/Vol98No19/1Vol98No19.pdf
7. **T. Bhanuteja et al.,** "Symptoms based multiple disease prediction model using machine learning approach," IJITEE 2021.
8. Link: https://www.ijitee.org/wp-content/uploads/papers/v10i3/C89830210321.pdf
9. **S. Grampurohit and C. Sagarnal,** "Disease prediction using machine learning algorithms," INCET 2020.
10. Link: https://ieeexplore.ieee.org/document/9154153
11. **R. Keniya et al.,** "Disease prediction from various symptoms using machine learning," SSRN 2020.
12. Link: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3661426
13. **P. Hema et al.,** "Disease prediction using symptoms based on machine learning algorithms," BHARAT 2022.
14. Link: https://ieeexplore.ieee.org/document/9751391
15. **D. Nishad, A. Mishra and N. Goyal,** "Symptom-Based Disease Prediction Using Machine Learning," Confluence 2024.
16. Link: https://ieeexplore.ieee.org/document/10409937
17. **Y. Zoabi, S. Deri-Rozov and N. Shomron,** "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," NPJ Digital Medicine, 2021.
18. Link: https://www.nature.com/articles/s41746-021-00456-x
19. **S. R. Islam et al.,** "Deep learning on symptoms in disease prediction," Machine Learning for Healthcare Applications, 2021.
20. Link: https://link.springer.com/chapter/10.1007/978-981-15-7078-8_5