International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. 11 Special Issue 02 | 2026

ArogyaBot: AI Health Assistant

Omkar Chaudhari[1], Himanshu Chhajed[2], Anisha Dhone[3], Devyani Jamdar[4], Amruta Kadam[5], Dr.A.B.Pawar[6]

[1,2,3,4,5] B.Tech Final year,Department of Computer Engineering, Sanjivani College Of Engineering,Kopargaon,SPPU Pune, Maharashtra, India

[6]Professor, Department of Computer Engineering, Sanjivani College Of Engineering,Kopargaon,SPPU Pune, Maharashtra, India

DOI: 10.5281/zenodo.19287703

ABSTRACT

With the rapid growth of healthcare data, accurate disease prediction has become an essential challenge in medical science. Early detection of diseases can significantly reduce treatment costs and improve patient outcomes. This work proposes an intelligent disease prediction and health assistance system using machine learning models (XGBoost) combined with medical history and symptom analysis. The system accepts symptoms provided by the user, processes them through an XGBoost classifier, and predicts the most probable disease. It further classifies the severity of the disease into low, moderate, high, or extreme levels. Based on the prediction and severity, the system recommends suitable medicines, dietary plans, exercise suggestions, and necessary precautions, while advising consultation with a doctor for severe cases. By leveraging structured healthcare datasets and machine learning techniques, the proposed system not only enhances disease prediction accuracy but also offers personalized health assistance. This approach contributes to preventive healthcare by enabling early diagnosis, reducing dependency on direct physician consultation, and improving accessibility for patients in remote areas.

Keywords:- XGBoost, Disease Prediction, Machine Learning, Symptom Analysis, Medical Recommendation System, Patient Management.

I. INTRODUCTION

Healthcare is one of the most critical domains where technological advancements can directly impact human life. With the exponential growth of healthcare data, accurate analysis and prediction of diseases have become increasingly important. Traditional medical diagnosis requires expert consultation, multiple tests, and considerable time, which may not always be accessible to patients, especially in remote areas. In this context, artificial intelligence (AI) and machine learning (ML) have emerged as powerful tools for disease prediction, risk assessment, and personalized health recommendations.

In recent years, machine learning models such as XGBoost have shown remarkable performance in classification and prediction tasks. Leveraging these capabilities in the healthcare sector allows for the development of intelligent systems that can predict diseases based on patient symptoms and medical history. Such systems not only reduce diagnostic time and cost but also provide early warnings, enabling preventive healthcare measures.

The proposed system takes user symptoms as input and applies XGBoost-based predictive modelling to identify the most likely disease. It further classifies the severity of the predicted disease into categories such as low, moderate, high, or extreme based on symptom weights. Based on the outcome, the system suggests appropriate medicines, dietary plans, exercise routines, and precautions, thereby serving as a comprehensive health assistant. In critical cases, it also advises the user to consult a medical professional immediately.

By combining structured healthcare data analysis with machine learning techniques, this system aims to assist both patients and healthcare providers. Patients gain accessibility to reliable, low-cost, and instant health predictions, while medical professionals can use it as a decision-support tool to improve diagnostic efficiency. The system also includes patient management features, digital health records, and doctor panels for continuous care. Overall, the proposed approach contributes toward building a smart, preventive, and personalized healthcare solution.

II. LITERATURE SURVEY

The prediction of disease at earlier stage becomes important task. But the accurate prediction on the basis of symptoms becomes too difficult for doctor. There is a need to study and make a system which will make it easy for end users to predict the chronic diseases without visiting physician or doctor for diagnosis. Table 1 shows literature survey about disease prediction systems proposed in different literatures.

The study by Wenxing et al. (IEEE Access) proposes a deep learning-based hybrid recommendation algorithm that predicts potential diseases based on a patient's medical history. It considers both high-order and low-order relationships among disease features, thereby improving comprehensiveness compared to earlier systems. Similarly, Dahiwade et al. (IEEE Xplore) developed a general disease prediction model that incorporates lifestyle habits and medical check-up data to enhance prediction accuracy, achieving 84.5% accuracy with low time consumption and minimal cost.

Xu et al. introduced an explainable learning approach using comorbidity networks for disease risk prediction, integrating these networks into a Bayesian framework and demonstrating superior predictive performance using real-world medical data. Repaka et al. focused on heart disease prediction using a Naive Bayes approach, achieving an accuracy of 89.77% while reducing the number of attributes and ensuring improved security through advanced encryption techniques.

Gao et al. proposed a method for predicting disease similarity using heterogeneous disease information networks and node representation learning, outperforming traditional approaches limited to chemical-disease data. Mathew et al. developed a machine learning-based chatbot system capable of diagnosing diseases and recommending treatments, reducing the need for routine check-ups, lowering costs, and providing a personalized user experience without requiring a physician.

Maurya et al. introduced a system for chronic kidney disease (CKD) prediction along with diet recommendations using classification algorithms. The system identifies critical features from medical records and classifies CKD into stages, assisting both doctors and patients in treatment planning. Another work by Dahiwade et al. compares CNN and KNN algorithms for disease prediction using UCI datasets, concluding that CNN provides higher accuracy and faster processing for large datasets.

Pandey and Prabha proposed an IoT-based smart health monitoring system that collects real-time patient data using sensors and predicts heart disease using machine learning techniques. This system is particularly useful for early detection and for deployment in rural areas with limited healthcare facilities. Finally, VijiyaKumar et al. presented a Random Forest-based system for early diabetes prediction, which demonstrates higher accuracy and efficient, real-time disease prediction compared to other algorithms.

III.PROPOSED SYSTEM

The system analyzes the symptoms provided by the user as input and gives the predicted disease as an output. Disease prediction is done by implementing the XGBoost Classifier. The XGBoost Classifier calculates the probability of the disease and identifies the most likely condition. Along with disease prediction, the system also calculates the severity of the disease and as per severity level suggests appropriate medicines, dietary recommendations, exercise plans, and necessary precautions..

Architecture

The correct prediction of disease is the most challenging task in healthcare informatics. To overcome this problem, machine learning plays an important role in predicting diseases. Medical science has a large amount of data growth per year. Due to the increased amount of data growth in the medical and healthcare field, accurate analysis of medical data benefits early patient care. This system is used to predict diseases according to symptoms. As shown in the figure below, databases containing symptoms of different diseases, symptom severity weights, and disease recommendations are fed as input to the system along with current symptoms of the user and medical history of the patient (when the patient observed the same type of symptoms before). The Python-based system uses the XGBoost algorithm to predict the disease the patient is suffering from. After predicting the disease, the system classifies it into low, moderate, high, or extreme severity conditions.

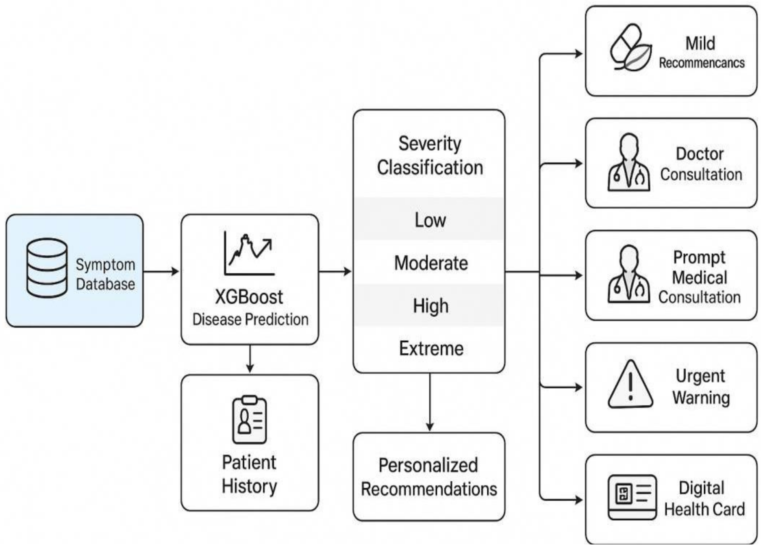

Fig 1 architecture of proposed system

If the disease is low severity, it suggests some medicine and lifestyle changes. In case of moderate severity, along with medicines, the system suggests the user visit a doctor if symptoms don't fade away. When it's a high or extreme severity case, the system warns the user to immediately visit a doctor. The system also suggests personalized diet plans and exercises as per the predicted disease.

XGBoost Algorithm

Over the last decade, tremendous progress has been made in the field of machine learning algorithms. Extreme Gradient Boosting (XGBoost) has demonstrated state-of-the-art results on many classification problems, especially in healthcare prediction tasks.

XGBoost is an ensemble learning method based on gradient boosted decision trees. The algorithm creates multiple decision trees sequentially, where each subsequent tree learns from the errors of the previous trees. The distinctive architecture of XGBoost makes it particularly effective for structured data classification problems like symptom-based disease prediction.

The mathematical formulation of XGBoost can be represented as:

$\hat{y}_i = \varphi(x_i) = \Sigma_k f_k(x_i), f_k \in F$ where:

- $\hat{y}_i$ is the predicted output for sample i
- x_i is the feature vector (symptoms)
- f_k represents independent tree structures
- F is the space of all possible trees

The objective function consists of both training loss and regularization: $Obj(\theta) = \Sigma_i l(\hat{y}_i, y_i) + \Sigma_k \Omega(f_k)$

where:

- $l(\hat{y}_i, y_i)$ is the differentiable convex loss function
- $\Omega(f_k) = \gamma T + ½\lambda\|w\|^2$ is the regularization term

For multi-class classification problems with K diseases, the softmax function is used to obtain probabilistic outputs:

$P(y_j = k|x_j) = e^{\hat{y}_k} / \Sigma_{m=1}^K e^{\hat{y}_m}$

This allows XGBoost to act as a probability estimator for disease classification problems, providing the likelihood of each potential disease given the input symptoms.

Key Features of XGBoost in Disease Prediction:

- Regularization: Helps prevent overfitting through L1 and L2 regularization
- Handling Missing Values: Automatically learns the best direction to handle missing symptom data
- Tree Pruning: Uses max_depth parameter to prevent overfitting
- Cross-Validation: Built-in capability for performance evaluation
- Parallel Processing: Efficient handling of large symptom datasets

Implementation Steps for XGBoost Training:

1. Data Preprocessing: Convert symptoms into multi-hot encoded feature vectors using MultiLabelBinarizer
2. Label Encoding: Encode disease labels using LabelEncoder for multi-class classification
3. Model Configuration: Set hyperparameters including:

- max_depth: 3
- learning_rate: 0.13
- n_estimators: 350
- subsample: 0.8
- colsample_bytree: 0.9
- reg_lambda: 1.2

4. Model Training: Train the classifier on symptom-disease mapping data
5. Model Evaluation: Assess performance using accuracy score and cross-validation
6. Model Persistence: Save trained model and encoders using joblib for deployment

Critical Components for XGBoost Implementation:

- Feature Engineering: Transform symptom lists into binary feature vectors
- Hyperparameter Tuning: Optimize parameters for maximum prediction accuracy
- Multi-class Classification: Handle multiple disease categories simultaneously
- Probability Calibration: Ensure predicted probabilities reflect true likelihoods

The XGBoost model in this system processes symptom inputs through multiple decision trees, combines their predictions, and outputs the most probable disease along with confidence scores. This approach enables accurate disease prediction while providing interpretable results based on symptom patterns learned from historical medical data.

IV.CONCLUSION

In this work, a comprehensive disease prediction and patient management system based on machine learning algorithms has been presented. By utilizing XGBoost classifier and symptom analysis, patient data can be effectively analyzed to predict diseases based on symptoms. Experimental results indicate that XGBoost consistently outperforms traditional algorithms such as Naïve Bayes, Decision Trees, and Multilayer Perceptron in terms of both accuracy and computational efficiency for structured medical data. This demonstrates the suitability of XGBoost for healthcare datasets where timely and accurate predictions are critical.

The proposed system not only predicts the most probable disease but also classifies its severity into low, moderate, high, and extreme levels, thereby guiding patients toward appropriate actions. Additionally, it suggests suitable medicines, diet plans, exercise routines, and necessary precautions, making the system more user-centric and practically useful. The integration of patient history tracking and doctor panels ensures continuity of care and enhances the system's practical utility in real healthcare scenarios.

The system effectively reduces diagnostic costs and time while improving healthcare accessibility, particularly for patients in remote areas. However, it should be noted that its scope is limited to non-emergency conditions, and professional medical consultation is still necessary for complex cases and severe symptoms. The digital QR codes provide emergency-ready health information, adding significant value to patient safety.

Future work may focus on integrating the system with telemedicine platforms, enabling automated appointment scheduling with specialized doctors based on predicted diseases, as well as generating detailed health reports that can be directly shared with medical professionals. Moreover, incorporating real-time health monitoring through IoT devices, expanding the symptom and disease database, enhancing data security protocols, and adding multi-language support can further improve reliability, accessibility, and trust in such predictive healthcare systems. The integration of more advanced ensemble methods and deep learning approaches could also be explored for improved prediction accuracy across a wider range of medical conditions.

V.REFERENCES

[1] Wenxing Hong, Ziang Xiong, Nannan Zheng, Yang Weng, “A Medical-History-Based Potential Disease Prediction Algorithm”, A Medical-History-Based Potential Disease Prediction Algorithm IEEE Access VOLUME 7, 2019, doi 10.1109/ACCESS.2019.2940644

[2] Dahiwade, D., Patle, G., & Meshram, E. (2019). Designing Disease Prediction Model Using Machine Learning Approach. 2019 Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019) IEEE Xplore doi:10.1109/iccmc.2019.8819782

[3] Xu, Z., Zhang, J., Zhang, Q., & Yip, P. S. F. (2019). Explainable Learning for Disease Risk Prediction Based on Comorbidity Networks. 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). doi:10.1109/smc.2019.8914644

[4] Repaka, A. N., Ravikanti, S. D., & Franklin, R. G. (2019). Design And Implementing Heart Disease Prediction Using Naives Bayesian. 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). doi:10.1109/icoei.2019.8862604

[5] Gao, J., Tian, L., Wang, J., Chen, Y., Song, B., & Hu, X. (2020). Similar Disease Prediction with Heterogeneous Disease Information Networks. IEEE Transactions on NanoBioscience,

1–
1. doi:10.1109/tnb.2020.2994983

[6] Mathew, R. B., Varghese, S., Joy, S. E., & Alex, S. S. (2019). Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning. 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). doi:10.1109/icoei.2019.8862707

[7] Maurya, A., Wable, R., Shinde, R., John, S., Jadhav, R., & Dakshayani, R. (2019). Chronic Kidney Disease Prediction and Recommendation of Suitable Diet Plan by using Machine Learning. 2019 International Conference on Nascent Technologies in Engineering (ICNTE). doi:10.1109/icnte44896.2019.8946029

[8] Dahiwade, D., Patle, G., & Meshram, E. (2019). Designing Disease Prediction Model Using Machine Learning Approach. 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). doi:10.1109/iccmc.2019.8819782

[9] Pandey, H., & Prabha, S. (2020). Smart Health Monitoring System using IOT and Machine Learning Techniques. 2020 Sixth International Conference on Bio Signals, Images, and Instrumentation (ICBSII). doi:10.1109/icbsii49132.2020.9167660

VijiyaKumar, K., Lavanya, B., Nirmala, I., & Caroline, S. S. (2019). Random Forest Algorithm for the Prediction of Diabetes. 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN). doi:10.1109/icscan.2019.8878802