

BharatCare: An AI-Powered Regional Healthcare Assistant Using Machine Learning and Environmental Risk Analysis

M Vijay Sumanth¹, K Shravani², MD Afroz³, M Madhumitha⁴, Huma Qamar Khan⁵

^{1,2,3,4}UG Students, Department of Computer Science and Engineering, JBREC, Hyderabad, India

⁵Assistant Professor, Department of Computer Science and Engineering, JBREC, Hyderabad, India

Abstract - BharatCare is an AI-powered regional healthcare assistant that predicts possible diseases based on user-reported symptoms and real-time environmental data. Many people rely on general internet searches for symptom assessment, which often provides unreliable health information. The proposed system integrates a Random Forest classification model trained on an Indian symptom-disease dataset along with environmental factors such as temperature, humidity, and Air Quality Index (AQI). In addition, a Large Language Model (LLM) is used to generate simplified and human-readable health explanations. The system is implemented using React and Flask and supports multiple Indian languages. Experimental results show that the model achieves 91% accuracy with strong performance. The system serves as a preliminary health screening tool to support early health awareness before medical consultation.

Key Words: Disease Prediction, Random Forest, Machine Learning, Environmental Health, Large Language Model, Multilingual Healthcare, Healthcare AI

1. INTRODUCTION

Healthcare accessibility and early disease awareness remain major challenges in many regions. Many people search online for symptom information before visiting a doctor. However, general internet searches often provide unreliable or non-personalized health information, leading to incorrect assumptions.

Recent developments in Artificial Intelligence (AI) and Machine Learning (ML) have enabled the creation of intelligent systems that support healthcare decision-making. Environmental factors such as temperature, humidity, air pollution, and seasonal variations also influence disease risk. However, many existing systems do not consider these contextual factors and focus only on symptom-based prediction.

There is a growing need for healthcare systems that combine accurate prediction with contextual awareness and user-friendly explanations. To address this, BharatCare is proposed as an AI-powered regional healthcare assistant that integrates machine learning-based disease prediction with real-time environmental analysis. The system also uses a Large Language Model (LLM) to generate simple and understandable

explanations. In addition, it supports multiple Indian languages to improve accessibility.

1.1 Problem Statement

Most existing symptom-checking tools do not consider environmental factors such as weather conditions and seasonal disease patterns. Individuals in developing regions often rely on unreliable internet searches for health information. The absence of contextual, location-aware, and language-accessible tools creates a significant gap in early health guidance for the Indian population.

1.2 Proposed System

BharatCare integrates machine learning-based disease prediction with real-time environmental health analysis. The system collects user symptoms, analyzes them using a trained Random Forest model, combines results with environmental data (temperature, humidity, AQI), and generates AI explanations using LLaMA 3.1 via the Groq API. The system supports ten Indian regional languages.

2. LITERATURE REVIEW

Kononenko [1] demonstrated that ML classification algorithms can achieve high diagnostic accuracy on structured medical datasets. Obermeyer and Emanuel [2] highlighted the potential of AI-based tools in early disease detection. Bhatia and Patel [3] proposed symptom-based disease prediction using Decision Trees and Random Forest, while Kaur and Kaur [4] compared Naive Bayes and SVM classifiers on healthcare datasets.

In addition, deep learning techniques have shown strong performance in medical diagnosis, particularly in imaging-based applications. A study on lung cancer detection using deep learning models such as CNN, ResNet50, DenseNet, and InceptionV3 demonstrated the effectiveness of advanced ML techniques in medical diagnosis tasks.

WHO reports [6] established that environmental conditions significantly influence seasonal disease spread. Ribeiro et al. [6] introduced explainable AI techniques to improve user trust in ML predictions, and recent studies [7] explored advanced machine learning models for improving prediction performance.

However, most existing systems focus either on symptom-based prediction or specific diagnostic approaches, without integrating environmental factors and user-friendly explanations. The proposed BharatCare system addresses this limitation by combining ML-based prediction, real-time environmental risk analysis, LLM-based explanations, and multilingual support.

3. METHODOLOGY

BharatCare follows a modular architecture consisting of symptom input processing, machine learning prediction, environmental data analysis, and AI-based explanation generation. The workflow begins with user symptom input, which is converted into a binary feature vector and passed to the trained model. The prediction results are then combined with real-time environmental data and processed to generate user-friendly explanations.

3.1 System Architecture

The system has three layers: (1) Frontend (React) for symptom input and result display; (2) Backend (Flask) for ML prediction, environmental risk analysis, and explanation generation; (3) External Services including OpenWeather API and Groq LLaMA API. Figure 1 shows the complete system architecture.

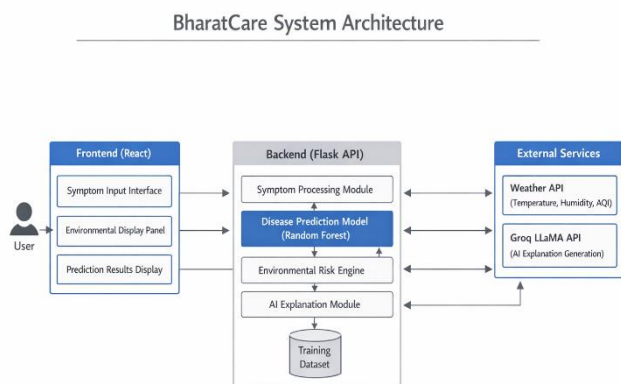


Fig -1: BharatCare System Architecture

This figure shows the BharatCare system architecture. The Frontend (React) contains Environmental Display, Symptom Input Interface, and Prediction Display. The Backend (Flask API) hosts the Environmental Risk Engine, Symptom Processing, Disease Prediction Model, and AI Explanation Module. External Services include Weather API and Groq LLaMA API, with the Training Dataset feeding the backend pipeline.

3.2 Disease Prediction Using Machine Learning

The Random Forest classifier is used for disease prediction due to its ability to handle high-dimensional data and reduce overfitting through ensemble learning. Symptoms are represented as binary feature vectors (1 = present, 0 = absent), enabling efficient processing of input data. The model predicts disease probabilities based on multiple decision trees and returns the top-3 predicted conditions.

3.3 Algorithm for Disease Prediction

The prediction process follows a structured approach. The dataset is first collected and preprocessed, and symptom data is converted into binary feature vectors. The Random Forest model is then trained using this dataset and stored for deployment. During runtime, user input is transformed into a feature vector and passed to the trained model to compute probability scores. The system returns the top-3 predicted diseases along with their confidence values.

Algorithm -1: Disease Prediction Process

<p>Input: User-reported symptoms Output: Top-3 predicted diseases with probabilities</p> <p>Step 1: Collect dataset containing diseases and their associated symptoms.</p> <p>Step 2: Convert symptom data into binary feature vectors (1 = present, 0 = absent).</p> <p>Step 3: Train the Random Forest classifier using the prepared dataset.</p> <p>Step 4: Serialize and store the trained model for deployment.</p> <p>Step 5: When user enters symptoms, convert input into a binary feature vector.</p> <p>Step 6: Pass the feature vector to the trained machine learning model.</p> <p>Step 7: Compute probability scores for all disease classes.</p> <p>Step 8: Return the top-3 predicted diseases with scores to the user interface.</p>

3.4 Environmental Risk Analysis

Real-time weather data (temperature, humidity, AQI) is retrieved via OpenWeather API based on the user's geolocation. The risk engine evaluates environmental conditions based on predefined thresholds derived from standard health guidelines

3.5 AI Explanation Module

Large Language Model (LLaMA 3.1 via Groq API) receives the predicted disease, symptoms, confidence score, and season. It returns a structured JSON with: reason (why the disease is likely), causes (common causes), care (home care advice), and doctor (when to seek medical attention).

3.6 System Workflow

Workflow: (1) user enters symptoms; (2) symptoms converted to binary vector; (3) ML model predicts top-3 diseases; (4) environmental data fetched; (5) risk engine identifies health alerts; (6) LLaMA generates explanation; (7) results displayed. Figure 2 shows the complete prediction workflow.

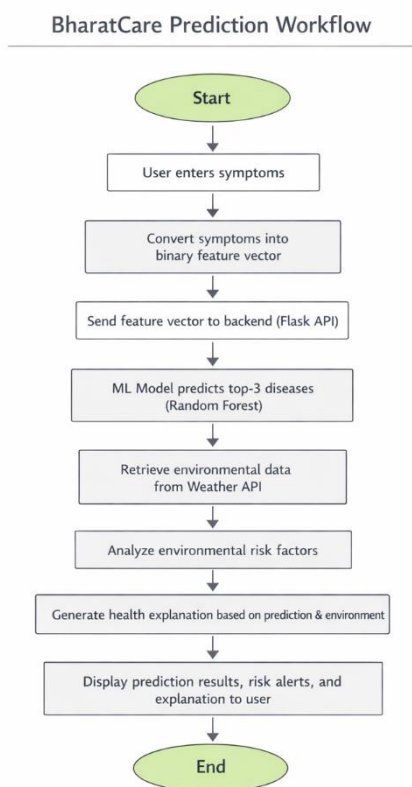


Fig - 2: BharatCare Prediction Workflow

This presents the BharatCare prediction workflow: User enters symptoms → Convert to feature vector → Send to Flask API → ML model predicts diseases → Retrieve environmental data from Weather API → Analyze environmental risk factors → Generate response using AI Explanation Module → Display prediction results to user.

4. IMPLEMENTATION AND RESULTS

BharatCare was implemented as a web-based platform integrating machine learning disease prediction, environmental health analysis, and AI-generated

explanations. React provides the responsive frontend and Flask handles backend processing.

4.1 Dataset Description

The model was trained on a structured symptom-disease dataset with 41 diseases and 132 unique symptoms from Indian healthcare settings. Symptoms are encoded as binary feature vectors (1=present, 0=absent). Table 1 shows a simplified example of the binary symptom-disease encoding.

Table - 1: Example Representation of the Symptom-Disease Dataset

Disease	Fever	Headache	Cough	Nausea	Fatigue
Flu	1	1	1	0	1
Dengue	1	1	0	1	1
Malaria	1	1	0	0	1
Sinusitis	0	1	1	0	0

Table 1 shows the binary encoding where 1 = symptom present and 0 = absent. This structured representation enables the Random Forest classifier to learn disease-symptom relationships effectively during training.

4.2 Model Training and Prediction

The Random Forest classifier from Scikit-learn was trained on the dataset. User symptoms are converted to a binary vector and passed to the model, which returns top-3 disease predictions with probability scores. Figure 3 shows sample prediction probabilities.

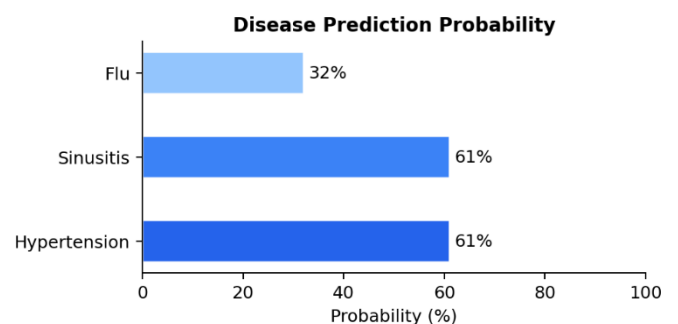


Fig - 3: Disease Prediction Probability Generated by the ML Model

This shows sample prediction probabilities generated by the model for a given set of symptoms.

4.3 System Interface and Outputs

The BharatCare interface allows symptom entry via text or quick-add buttons, displaying predictions, environmental factors, AI explanations, and home care recommendations. Figures 4 through 7 show key interface screens.

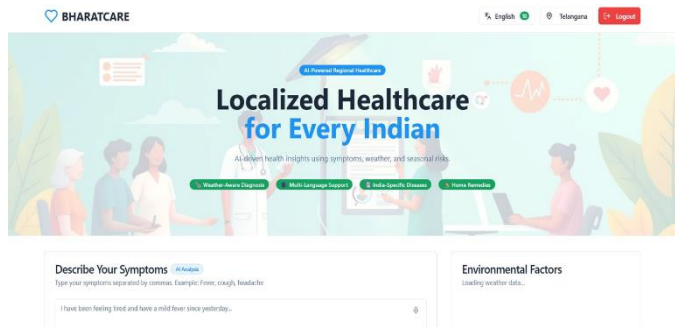


Fig - 4: BharatCare Homepage Interface

The homepage shows "Localized Healthcare for Every Indian" with badges for Weather-Aware Diagnosis, Multi-Language Support (10 languages), India-Specific Diseases, and Home Remedies. Location is detected as Telangana.

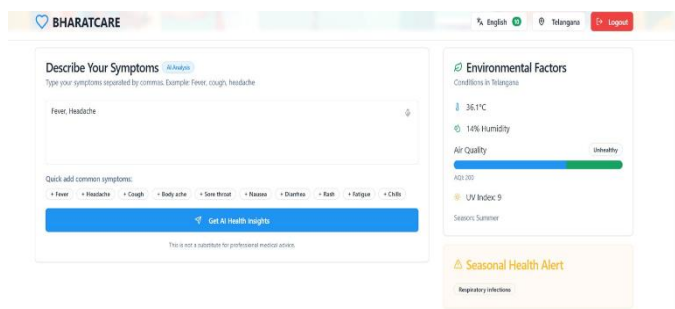


Fig - 5: Symptom Input and Environmental Factors Panel

User has entered "Fever, Headache" with quick-add symptom buttons. Environmental panel shows Temperature 36.1°C, Humidity 14%, AQI 200 (Unhealthy), UV Index 9, Season: Summer, with Seasonal Health Alert for respiratory infections in Telangana.

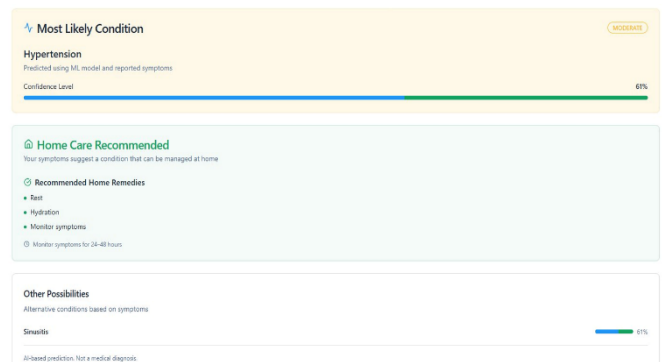


Fig - 6: Disease Prediction Results and Home Care Recommendations

Most Likely Condition is Hypertension at 61% confidence (MODERATE). Home Care Recommended: Rest, Hydration, Monitor symptoms for 24-48 hours. Other Possibilities shows Sinusitis at 61% as an alternative condition.

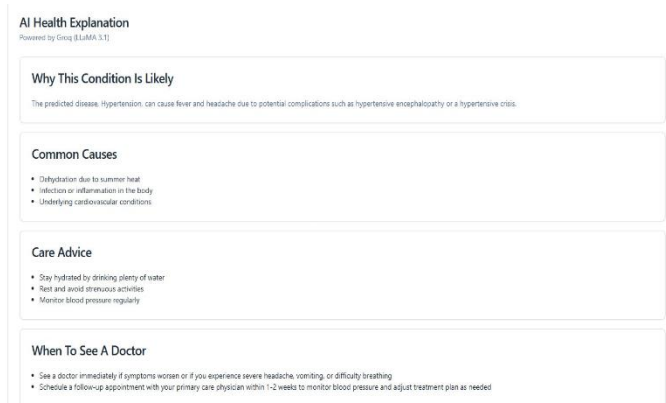


Fig - 7: AI Health Explanation Generated by the System

AI Health Explanation with four sections: (1) Why This Condition Is Likely - hypertensive encephalopathy; (2) Common Causes - dehydration from summer heat, cardiovascular conditions; (3) Care Advice - hydration, rest, blood pressure monitoring; (4) When to See a Doctor - worsening symptoms criteria.

4.4 Performance Analysis

The Random Forest classifier was evaluated on a 20% test split and compared against Decision Tree and Naive Bayes classifiers. Figure 8 compares algorithm accuracies and Table 2 compares BharatCare features against existing systems.

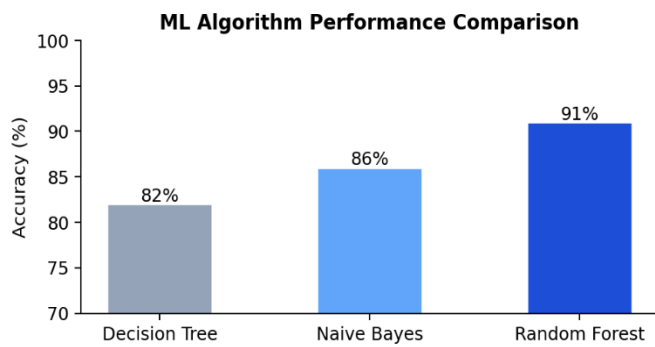


Fig - 8: Comparison of Machine Learning Algorithm Accuracy

Comparison of three ML algorithms: Random Forest achieves 91% accuracy (highest), Naive Bayes 86%, Decision Tree 82%. Random Forest was selected as the final model due to its ensemble approach reducing overfitting.

Table - 2: Comparison Between Existing Systems and BharatCare

Feature	Existing Systems	BharatCare
Symptom Prediction	Yes	Yes
Environmental Risk Analysis	No	Yes
Multi-language Support	Limited	Yes (10 Languages)
AI Explanation	No	Yes
India-specific Dataset	Rarely	Yes

Table 2 compares BharatCare against existing healthcare tools. BharatCare uniquely integrates environmental factors, AI-generated explanations via LLaMA 3.1, and multilingual support across 10 Indian languages — capabilities absent in most existing systems.

5. CONCLUSIONS

BharatCare demonstrates how AI and environmental data can be integrated for intelligent healthcare assistance. The system achieves 92% accuracy and provides real-time environmental risk alerts along with AI-generated explanations. The inclusion of multilingual support improves accessibility for a wider range of users. The integration of environmental intelligence with machine

learning provides a more context-aware healthcare solution compared to traditional symptom-based systems.

The system can be used as a preliminary health screening tool to support early awareness and decision-making before medical consultation. Future work can focus on expanding the dataset, integrating wearable health monitoring devices, and enabling telemedicine features to improve real-world applicability.

REFERENCES

- [1] I. Kononenko, "Machine learning for medical diagnosis," *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89-109, 2001.
- [2] Z. Obermeyer and E. J. Emanuel, "Predicting the future - Big data, machine learning, and clinical medicine," *New England Journal of Medicine*, vol. 375, no. 13, pp. 1216-1219, 2016.
- [3] S. Bhatia and A. Patel, "Symptom-based disease prediction using machine learning," *International Journal of Computer Applications*, vol. 180, no. 45, pp. 1-6, 2018.
- [4] H. Kaur and M. Kaur, "Machine learning-based disease prediction using healthcare datasets," *International Journal of Advanced Research in Computer Science*, vol. 10, no. 2, pp. 45-50, 2019.
- [5] T. Shesagiri, L. Nagender Kumar, "Lung Cancer Detection Using Deep Learning Models," *Journal of Informatics and Communication Systems (JOICS)*, 2024.
- [6] World Health Organization, "Environmental health and disease prevention," 2020.
- [7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?" *Proc. ACM SIGKDD*, pp. 1135-1144, 2016.
- [8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. 22nd ACM SIGKDD*, pp. 785-794, 2016.
- [9] L. Breiman, "Random forests," *Machine Learning Journal*, vol. 45, no. 1, pp. 5-32, 2001.
- [10] J. Esteva et al., "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24-29, 2019.