INTERNATIONAL RESEARCH JOURNAL IN
ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET)
E - ISSN: 2454-4752 P - ISSN: 2454-4744 (www.irjaet.com)
VOL 11 ISSUE 6 (2025) PAGES 32-37
RECEIVED:25.10.2025 PUBLISHED:20.11.2025

Ai-Powered Chatbot for Symptom-Based Disease Prediction

S. Mahesh
2nd Year – M.S. in Data Science
EdTech Division, Exafluence Inc.
Sri Venkateswara University, Tirupati, India
Email: suryamaheshs7@gmail.com

Vijaya Lakshmi Kumba
Professor, Department of Computer Science
SVU College of CM & CS
Sri Venkateswara University, Tirupati, India
Email: vijayalakshmik4@gmail.com

Abstract
This project presents an AI-powered chatbot designed to assist users in identifying likely diseases based on a list of symptoms they provide. Developed in Python, the system leverages libraries such as Pandas and NumPy for data handling, while a Random Forest classifier from Scikit-learn is used to generate predictions with high accuracy. A user-friendly Gradio web interface enables seamless interaction and real-time feedback. For convenient access and deployment, the complete application is hosted on Hugging Face Spaces, allowing users to interact with the model online without requiring a local setup. This approach integrates machine-learning-based prediction with an intuitive interface, offering a helpful tool for preliminary guidance while emphasising that it is not a substitute for professional medical diagnosis.
Keywords: AI chatbot, disease prediction, Random Forest, machine learning, Python, Gradio, healthcare AI, Hugging Face.

1. INTRODUCTION

Healthcare systems generate vast volumes of symptom and diagnosis data, yet making this information understandable and actionable for general users remains challenging. This project aims to bridge that gap by developing an AI-powered chatbot capable of predicting likely diseases from user-provided symptoms. The system is implemented using Python, employing Pandas and NumPy for data preprocessing and a Random Forest classifier for disease prediction due to its balance of accuracy, robustness, and interpretability. Users interact with the system through a Gradio-based web interface, enabling real-time predictions. Deployment on Hugging Face Spaces ensures accessibility without the need for local installation. The primary motivation is to provide quick preliminary insights for individuals seeking to understand potential causes of their symptoms while clearly stating that the system cannot replace professional medical evaluation. The project demonstrates a complete, reproducible pipeline from raw symptom input to ranked disease predictions, highlighting the interpretability and effectiveness of Random Forest for this task.

INTERNATIONAL RESEARCH JOURNAL IN
ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET)
E - ISSN: 2454-4752 P - ISSN: 2454-4744 (www.irjaet.com)
VOL 11 ISSUE 6 (2025) PAGES 32-37
RECEIVED:25.10.2025 PUBLISHED:20.11.2025

2. LITERATURE REVIEW

Advancements in artificial intelligence and machine learning have significantly enhanced healthcare applications, especially in symptom-based disease prediction. Early systems used rule-based expert models that relied on manually crafted if-then rules. While interpretable, these systems required extensive maintenance and lacked adaptability. Machine learning approaches, including decision trees, support vector machines, and ensemble methods such as Random Forests, brought improved accuracy and generalisation. Random Forest, in particular, is widely used because it provides strong performance and interpretable feature importance metrics. Recent work has integrated natural language processing (NLP) to interpret free-text symptom descriptions.

Additionally, tools like Gradio have enabled user-friendly interfaces, and cloud platforms such as Hugging Face have made deployment accessible globally. Despite progress, many existing systems focus on single-disease predictions or lack an end-to-end pipeline integrating preprocessing, modelling, and deployment. This project addresses these limitations by combining a robust Random Forest model with an intuitive Gradio interface and cloud deployment.

3. METHODOLOGY

The methodology consists of four primary stages:

3.1 Data Collection and Preprocessing

A symptom–disease dataset was curated from reliable medical sources. Preprocessing included:

- Handling missing values
- Standardising symptom names
- Removing duplicates
- Encoding symptoms using multi-hot encoding

This ensured consistent and clean input for model training.

3.2 Feature Engineering

Multi-hot encoded vectors represent symptom presence for each record. If available, demographic features such as age or gender can be added to improve predictive performance. Feature ordering was standardised to maintain reproducibility.

3.3 Model Training and Evaluation

A Random Forest classifier was selected for its:

- Robustness
- Ability to handle high-dimensional data
- Interpretability
- The dataset was split into training, validation, and test sets. Hyperparameters (e.g., number of trees, max depth) were optimised using cross-validation. Metrics used included:
- Accuracy
- Precision
- Recall

INTERNATIONAL RESEARCH JOURNAL IN
ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET)
E - ISSN: 2454-4752 P - ISSN: 2454-4744 (www.irjaet.com)
VOL 11 ISSUE 6 (2025) PAGES 32-37
RECEIVED:25.10.2025 PUBLISHED:20.11.2025

- F1-Score
- Confusion matrix
- ROC curve

3.4 User Interface and Deployment

A Gradio interface enables users to input symptoms and receive disease predictions with confidence scores. The application was deployed on Hugging Face Spaces for global accessibility and ease of use.

4. DEMAND FORECASTING MODELS

(This section seems unrelated to the main project, but I corrected it while keeping your content.)
Demand forecasting models help organisations predict future product or service demand. They support inventory optimisation, resource allocation, and improved operational efficiency. Techniques include:

- Qualitative approaches: expert opinions, market research
- Quantitative approaches: time-series models (ARIMA, exponential smoothing), causal models (regression), and machine learning models (Random Forest, XGBoost, LSTM)

These models are evaluated using MAE, RMSE, and MAPE. AI-based forecasting has improved adaptability to dynamic markets.

5. ORDER OPTIMISATION MODELS

Order optimisation models guide businesses in deciding how much and when to order. Methods include:

- Economic Order Quantity (EOQ)
- Reorder Point (ROP)
- Stochastic models with demand uncertainty
- Mathematical programming (linear / mixed-integer programming)
- AI-based optimisation integrated with real-time systems

These models minimise costs, avoid stockouts, and improve inventory efficiency.

6. VISUALISATION AND DECISION SUPPORT

Visualisation tools enable users to explore data patterns through charts, dashboards, and heatmaps. Decision Support Systems (DSS) integrate visualisation with analytical models to help evaluate scenarios, optimise resources, and respond effectively to operational challenges.

7. MODEL EVALUATION

The Random Forest classifier was evaluated using multiple metrics.

The Random Forest classifier was evaluated using multiple metrics.

- Accuracy:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

INTERNATIONAL RESEARCH JOURNAL IN
ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET)
E - ISSN: 2454-4752 P - ISSN: 2454-4744 (www.irjaet.com)
VOL 11 ISSUE 6 (2025) PAGES 32-37
RECEIVED:25.10.2025 PUBLISHED:20.11.2025

- Precision:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- Recall:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- F1-Score:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Confusion matrices and ROC curves were used to interpret performance visually. Feature importance analysis enhanced model interpretability.

8. System Architecture

The system architecture integrates data processing, machine learning, and interactive user communication.

User Interface Layer:
Gradio interface for symptom input and prediction display.
Data Processing Layer:
Cleaning, encoding, and transforming symptoms into numerical vectors.
Machine Learning Layer:
Random Forest model trained on symptom–disease mappings.
Prediction Layer:
Displays ranked disease predictions with confidence scores and optional visualisations.
Deployment Layer:
Hosted on Hugging Face Spaces for public access.

9. RESULTS AND ANALYSIS

The model achieved:

- Accuracy: ~95%
- Precision: 94%
- Recall: 93%
- F1-Score: 94%

INTERNATIONAL RESEARCH JOURNAL IN
ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET)
E - ISSN: 2454-4752 P - ISSN: 2454-4744 (www.irjaet.com)
VOL 11 ISSUE 6 (2025) PAGES 32-37
RECEIVED:25.10.2025 PUBLISHED:20.11.2025

A confusion matrix showed minimal misclassification. ROC-AUC scores demonstrated strong discriminative ability. Feature importance analysis identified key symptoms contributing to predictions. The Gradio interface provided smooth interaction, and deployment on Hugging Face ensured accessibility and scalability.

10. ACKNOWLEDGMENT
I express sincere gratitude to my mentors, colleagues, and instructors for their support and guidance throughout this project. I also acknowledge the open-source tools and libraries that made development possible.

11. LIMITATIONS
Despite strong performance, the system has limitations:

- Accuracy depends on dataset quality.
- Common symptoms across diseases reduce precision.
- Lacks contextual understanding (medical history, lifestyle).
- Not a replacement for professional medical advice.
- Scalability may require optimisation for large deployments.

12. FUTURE WORK
Future enhancements may include:

- Integration with Electronic Health Records (EHRs)
- NLP models (BERT, GPT) for free-text symptom description
- Deep learning-based diagnostic models
- Adaptive learning from user feedback
- Multilingual support
- Mobile and IoT-based health tracking integration

13. CONCLUSION
This project successfully developed an AI-powered chatbot capable of predicting probable diseases based on user-provided symptoms. Using the Random Forest classifier and Python's powerful data-processing libraries, the system offers high accuracy and interpretable predictions. The Gradio interface enhances usability, and Hugging Face deployment ensures accessibility.
Though subject to certain limitations, the project demonstrates the practical potential of AI-driven healthcare support tools and lays the groundwork for future advancements in intelligent medical assistance systems.

REFERENCES

[1]. J. Brownlee, Machine Learning Mastery with Python, 3rd ed., Machine Learning Mastery, 2020.
[2]. S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed., Packt Publishing, 2019.
[3]. F. Chollet, Deep Learning with Python, 2nd ed., Manning Publications, 2021.
[4]. C. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press,

INTERNATIONAL RESEARCH JOURNAL IN
ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET)
E - ISSN: 2454-4752 P - ISSN: 2454-4744 (www.irjaet.com)
VOL 11 ISSUE 6 (2025) PAGES 32-37
RECEIVED:25.10.2025 PUBLISHED:20.11.2025

1999.
[5]. Gradio Documentation, “Build Machine Learning and Data Science Demos in Python,” 2025.
[6]. Scikit-learn Documentation, “Machine Learning in Python,” 2025.
[7]. WHO, “International Classification of Diseases (ICD-10),” 2019.