

Gen AI for Disease Prediction

M V V Krishna^{1*}; G Sri Jaya Sairam²; P Karthik³;
M Shakeer⁴; G Arjun⁵; SD Basheer Babu⁶

¹Assistant Professor, CSE Dept, Sri Vasavi Engineering College, Tadepalligudem.
^{2,3,4,5,6}Student, CSE Dept, Sri Vasavi Engineering College, Tadepalligudem, A.P., India

Corresponding Author: M V V Krishna^{1*}

Publication Date: 2025/04/23

Abstract: The project "Gen AI for Disease Prediction", utilizes advanced machine learning methodologies to forecast diseases such as diabetes, heart disease, and cancer based on user-input symptoms. It employs the Random Forest algorithm, a powerful and flexible machine learning model, ensuring accurate predictions while reducing the likelihood of overfitting. To enhance prediction reliability, the system incorporates data preprocessing techniques such as feature selection, data cleaning, and encoding. Developed using Scikit-learn, Python, and Django, the project integrates sophisticated machine learning functions with an intuitive web interface. Users can conveniently select symptoms from dropdown menus, which are then processed by the backend system. The machine learning model, trained on a well-structured dataset covering various medical conditions and their symptoms, analyzes the input to generate predictions. Ultimately, this project delivers a scalable and efficient disease prediction system that aids in the early detection of potential health issues.

Keywords: Random Forest Algorithm, Medical Diagnosis, Scikit-Learn, Symptom Analysis, Early Disease Detection.

How to Cite: M V V Krishna; G Sri Jaya Sairam; P Karthik; M Shakeer; G Arjun; SD Basheer Babu (2025). Gen AI for Disease Prediction. *International Journal of Innovative Science and Research Technology*, 10(4), 1067-1074.
<https://doi.org/10.38124/ijisrt/25apr760>

I. INTRODUCTION

Healthcare systems across the globe struggle with the challenge of ensuring timely and precise disease diagnosis. Conventional diagnostic techniques often depend on manual assessments, which can result in delays and potential inaccuracies. The integration of machine learning (ML) with artificial intelligence (AI) offers a ground-breaking technique for improving the accuracy of disease prediction. Gen AI for Disease Prediction leverages the Random Forest algorithm to evaluate symptoms provided by users and identify potential diseases such as diabetes, heart disease, and cancer. Developed using Python, Scikit-learn, and Django, the system features an intuitive web-based interface, allowing users to input symptoms and receive accurate predictions, thereby supporting early diagnosis and timely medical intervention.

II. LITERATURE SURVEY

A comprehensive analysis of existing research is essential to understand the advancements in machine learning-based disease prediction systems. Various studies have explored different machine learning algorithms to enhance diagnostic accuracy and early detection of diseases.

- Dr. C. K. Gomathy and Mr. A. Rohith Naidu (2021) conducted research on machine learning algorithms for

disease prediction, focusing on models like Naïve Bayes, K-Nearest Neighbor (KNN), Logistic Regression, and Decision Tree. Their findings showed that the Random Forest algorithm achieved the highest accuracy of 98.95%, surpassing Support Vector Machine (96.49%) and Naïve Bayes (89.4%), making it a reliable choice for predictive healthcare applications.

- In 2020, Marouane Ferjani explored various machine learning techniques, dataset processing, and performance metrics for disease prediction. His research highlighted that Support Vector Machine (SVM) was highly effective for predicting kidney diseases and Parkinson's disease, whereas Logistic Regression (LR) yielded the best results for heart disease detection, demonstrating the adaptability of ML models in medical diagnosis.
- Pooja Panapana et al. (2024) developed a disease prediction and medication recommendation system using Naïve Bayes, Random Forest, and Gaussian Naïve Bayes algorithms. Their study evaluated performance metrics such as accuracy, precision, recall, and F1-score. Results showed that Support Vector Machine (SVM) achieved the highest accuracy of 99.63%, proving its effectiveness in disease classification and predictive analytics.
- Mr. Sharan L. Pais et al. (2023) investigated Decision Tree and Random Forest classifiers for disease prediction. Their research focused on the development of an ML-based system that utilizes ensemble learning techniques to

enhance predictive performance. The findings emphasized the efficiency of Decision Tree and Random Forest models in analyzing symptoms and diagnosing diseases accurately.

- In 2022, Md Manjurul Ahsan conducted a study on Logistic Regression, Decision Tree, and Random Forest in disease diagnosis. His review detailed how machine learning assists in the early detection of various diseases, reinforcing the importance of data-driven diagnostic approaches. The study concluded that ML algorithms significantly contribute to improving healthcare decision-making.
- Christina Zhuang and Ramin Ramezani (2024) examined Decision Trees, Logistic Regression, and Support Vector Machines (SVMs) in disease prediction. Their study addressed challenges related to multiclass classification and unbalanced datasets, highlighting the effectiveness of machine learning methods in enhancing diagnostic accuracy. The authors suggested further improvements through advanced data balancing techniques and hyperparameter tuning.

These studies collectively demonstrate the effectiveness of machine learning in disease prediction, emphasizing the role of Random Forest, SVM, and Decision Tree algorithms in achieving high accuracy. The findings reinforce the significance of artificial intelligence-driven healthcare solutions in facilitating early disease detection and medical decision-making.

III. PROBLEMS IN EXISTING SYSTEM

➤ *Manual and Time-Intensive Diagnosis*

The current healthcare system relies on traditional medical consultations, where doctors manually assess symptoms to diagnose diseases. This time-consuming process often results in delays in treatment and increases the risk of late-stage disease detection. Additionally, medical expertise varies among professionals, leading to subjectivity and inconsistencies in diagnosis.

➤ *Limited Accuracy in Symptom-Based Checkers*

Some online platforms provide basic symptom-checking tools, but these systems operate on predefined rule-based algorithms rather than intelligent machine learning models. As a result, they struggle to analyze complex symptom patterns, often delivering generalized and unreliable predictions that do not consider individual health variations.

➤ *Delayed Disease Detection and Preventive Care*

Most conventional diagnostic methods focus on reactive treatment rather than proactive prevention. This leads to late-stage diagnosis, making treatment more challenging and costly. Additionally, patients do not receive automated insights on potential health risks based on their symptoms, limiting their ability to take preventive actions.

➤ *Dependency on Medical Expertise*

Disease diagnosis depends heavily on medical professionals' experience and judgment, which introduces the possibility of human error and misdiagnosis. Patients in

remote areas or regions with limited healthcare access face even greater difficulties in obtaining timely consultations, increasing health risks due to delayed diagnosis.

➤ *Lack of Real-Time AI Integration*

Traditional diagnosis systems do not integrate real-time AI and machine learning models that can improve with continuous data input. Without automated learning mechanisms, these systems remain static and do not adapt to emerging diseases or evolving medical knowledge, making them less effective over time.

IV. PROPOSED SYSTEM

➤ *Data Collection and Processing:*

- **Data Sources:** Utilize a curated medical dataset containing symptoms and corresponding diseases to train the prediction model. The dataset should be diverse and comprehensive to enhance prediction accuracy.
- **Data Cleaning:** Preprocess the data by handling missing values, removing inconsistencies, and normalizing features to ensure better model performance. Tools like Pandas and NumPy can be used for efficient data processing.
- **Feature Selection:** Implement feature engineering techniques to identify the most relevant symptoms that contribute to accurate disease prediction.

➤ *Machine Learning Model:*

- **Algorithm Selection:** Deploy the Random Forest algorithm, known for its high accuracy and robustness, to predict diseases based on user-inputted symptoms.
- **Training and Testing:** The model is trained using historical patient data, validated with test datasets, and fine-tuned to improve prediction precision.
- **Performance Evaluation:** Assess model accuracy using metrics such as precision, recall, F1-score, and confusion matrix to ensure reliable predictions.

➤ *User Interface Module:*

- **Web Interface:** A Django-based web application provides an intuitive and interactive platform where users can select symptoms from dropdown menus.
- **User Input Handling:** The system efficiently processes selected symptoms and sends them to the ML model for disease prediction.
- **Personalization:** Users receive customized predictions along with relevant precautionary measures for the diagnosed condition.

➤ *AI-Powered Insights:*

- **Precautionary Measures:** The system integrates OpenAI's API to offer personalized healthcare advice and preventive suggestions for predicted diseases.
- **Recommendation System:** Based on the predicted disease, the system suggests next steps, including seeking medical consultation or lifestyle changes.

➤ *Automated Disease Prediction:*

- **Real-Time Processing:** The model instantly processes user symptoms and returns results within seconds, making the system fast and efficient.
- **Scalability:** The system is designed to be easily expandable, allowing for the addition of new diseases and symptoms as more data becomes available.

V. OBJECTIVES

➤ *Early Disease Detection –*

The system predicts diseases such as diabetes, heart disease, and cancer based on symptoms provided by users, enabling timely medical intervention and preventive care.

➤ *Accurate and Efficient Predictions –*

By utilizing the Random Forest algorithm, the system ensures high prediction accuracy while minimizing overfitting, leading to reliable diagnostic outcomes.

➤ *Automated Symptom Analysis –*

The model processes user-inputted symptoms efficiently through feature selection, data cleaning, and encoding techniques, ensuring fast and precise disease predictions.

➤ *AI-Generated Health Precautions –*

By integrating OpenAI's API, the system provides personalized precautionary measures and recommendations, helping users take proactive steps toward better health management.

➤ *Improved Healthcare Accessibility –*

The Django-based web interface ensures an easy-to-use platform, allowing users to effortlessly input symptoms and

receive AI-powered predictions, making disease diagnosis more accessible and convenient.

VI. METHODOLOGY

The methodology of this project follows a structured approach, including data preprocessing, model training using the Random Forest algorithm, web application development and database management. Each stage ensures the system effectively predicts diseases based on user-input symptoms and provides precautionary suggestions for better health management.

➤ *Data Preprocessing:*

Data preprocessing is a fundamental step in preparing medical datasets for disease classification. The dataset contains various symptoms associated with multiple diseases, requiring cleaning, transformation, and encoding to ensure consistency and reliability.

- **Handling Missing Data** – Missing values in the dataset are identified and managed through imputation techniques or removal to maintain data integrity.
- **Feature Selection** – Redundant or irrelevant features are eliminated using statistical methods to improve model efficiency.
- **Data Normalization** – Feature scaling techniques such as Min-Max Scaling or Standardization are applied to optimize model performance by ensuring uniformity in data distribution.
- **Encoding Categorical Data** – Since symptom data often contains categorical values, encoding techniques (e.g., One-Hot Encoding or Label Encoding) are employed to convert them into numerical form for machine learning compatibility.

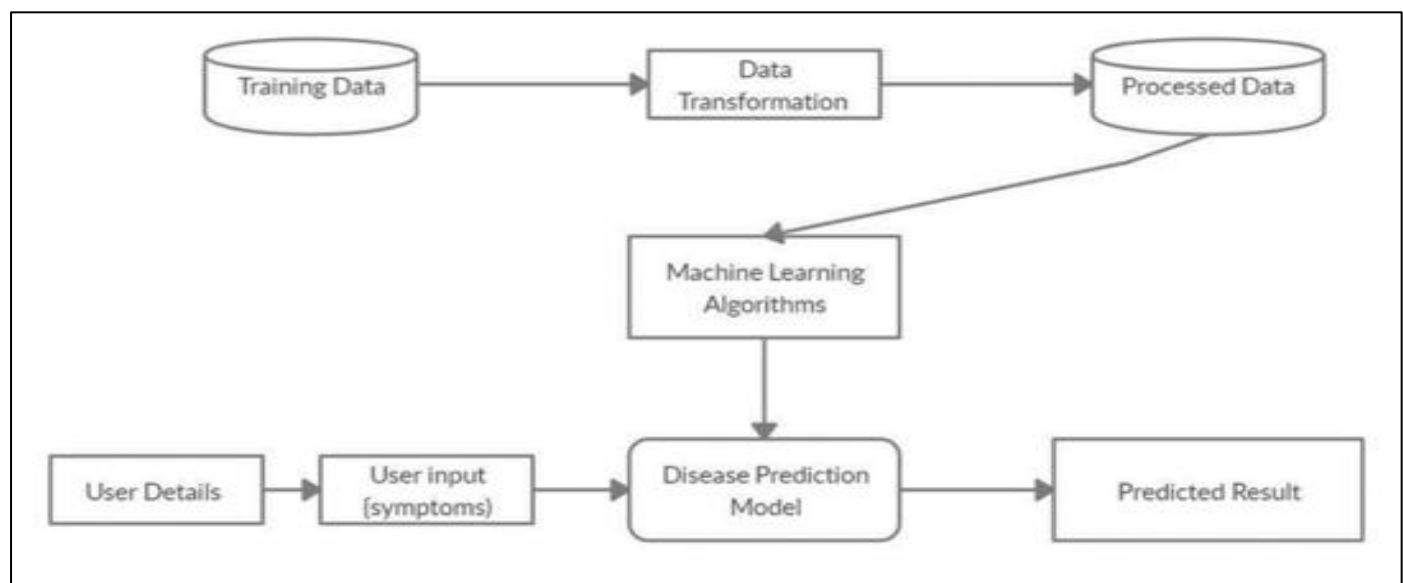


Fig 1 Architecture of Disease Prediction Based on Symptoms

➤ *Machine Learning Model – Random Forest:*

The Random Forest algorithm is used for disease classification due to its high accuracy and ability to handle large datasets efficiently. It is an ensemble learning

method that combines multiple decision trees to improve prediction performance. This approach reduces the risk of overfitting and enhances the model's reliability.

- *Working of the Random Forest Algorithm:*

- ✓ **Random Sampling** – A subset of data is randomly selected from the original dataset through the Bootstrap Aggregation (Bagging) technique.
- ✓ **Decision Tree Construction** – Each subset is used to train an independent decision tree, where it learns to classify diseases based on symptoms.

- ✓ **Feature Randomness** – During tree construction, only a random subset of features (symptoms) is considered at each decision node to enhance model diversity.
- ✓ **Voting Mechanism** – Each decision tree makes a separate disease prediction, and the final output is determined based on a majority voting approach (for classification tasks).
- ✓ **Final Prediction** – The disease with the highest votes from the ensemble of trees is selected as the final predicted outcome.

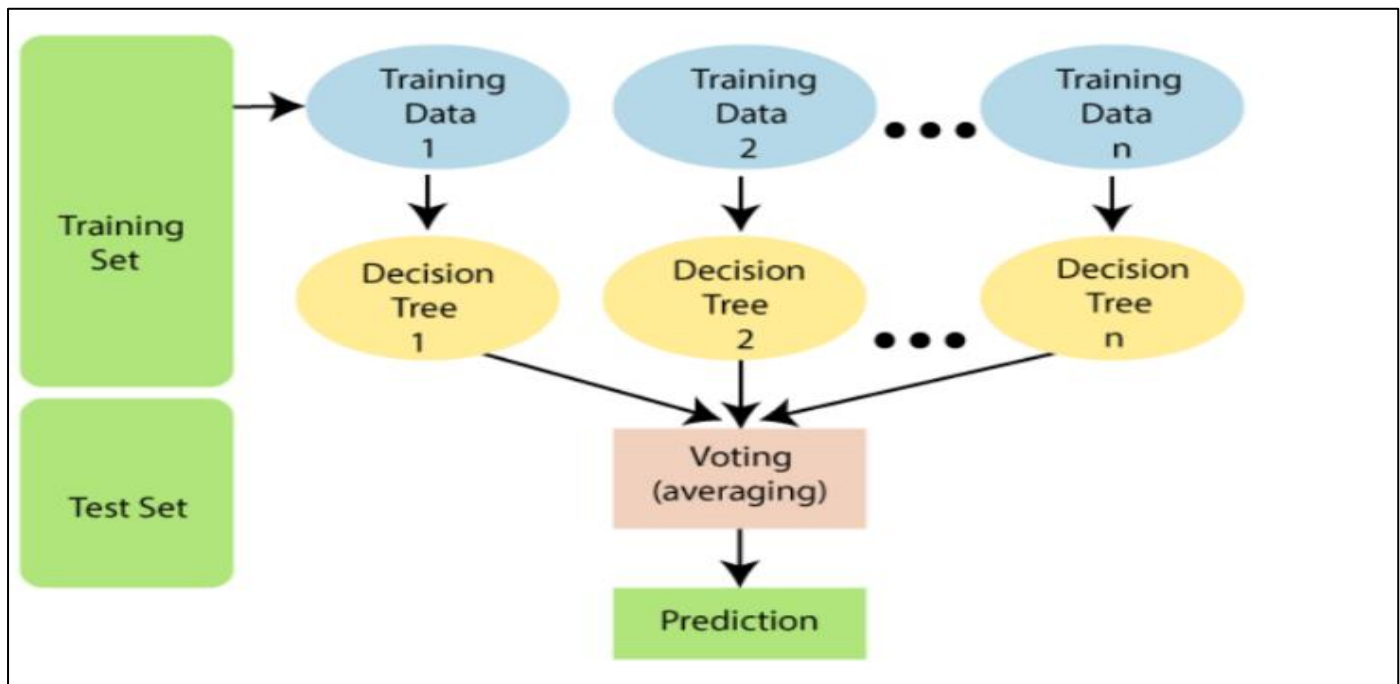


Fig 2 Working Flow of Random Forest Model

The bootstrap aggregation technique ensures that different subsets of data contribute to diverse decision trees, improving model accuracy. Unlike relying on a single decision tree, the Random Forest model aggregates multiple predictions, leading to a more robust and generalized classification.

➤ *Web Application Development :*

The web-based platform provides an interactive interface for users to input symptoms and receive real-time disease predictions.

- **User Authentication** – A secure authentication system enables users to sign up, log in, and manage their profiles.
- **Symptom Selection** – A dynamic dropdown menu allows users to select symptoms, which are then processed for prediction.
- **Prediction Display** – The predicted disease and associated precautionary measures (retrieved using OpenAI's API) are displayed.
- **User History Management** – The system stores past predictions, enabling users to track their health trends.

➤ *Database Management:*

Efficient database management ensures smooth functioning and structured storage of user data.

- **User Information Storage** – Securely stores user details, login credentials, and past predictions.
- **Medical Data Handling** – Stores symptom-disease relationships and model-generated insights.
- **Prediction History** – Logs previous disease predictions for future reference and analysis.

VII. RESULT AND ANALYSIS

The Gen AI for Disease Prediction system has been successfully implemented and tested to evaluate its accuracy, efficiency, and usability. The primary objective of this evaluation is to determine how well the system predicts diseases based on user-input symptoms and provides relevant precautionary measures. The performance of the Random Forest model was assessed using various machine learning evaluation metrics. Additionally, the web application's functionality and user experience were examined to ensure seamless interaction and accessibility.

➤ *Model Performance Evaluation :*

To analyze the effectiveness of the Random Forest classifier, key evaluation metrics were used, including:

- **Accuracy** – The percentage of correctly predicted disease cases among total cases.

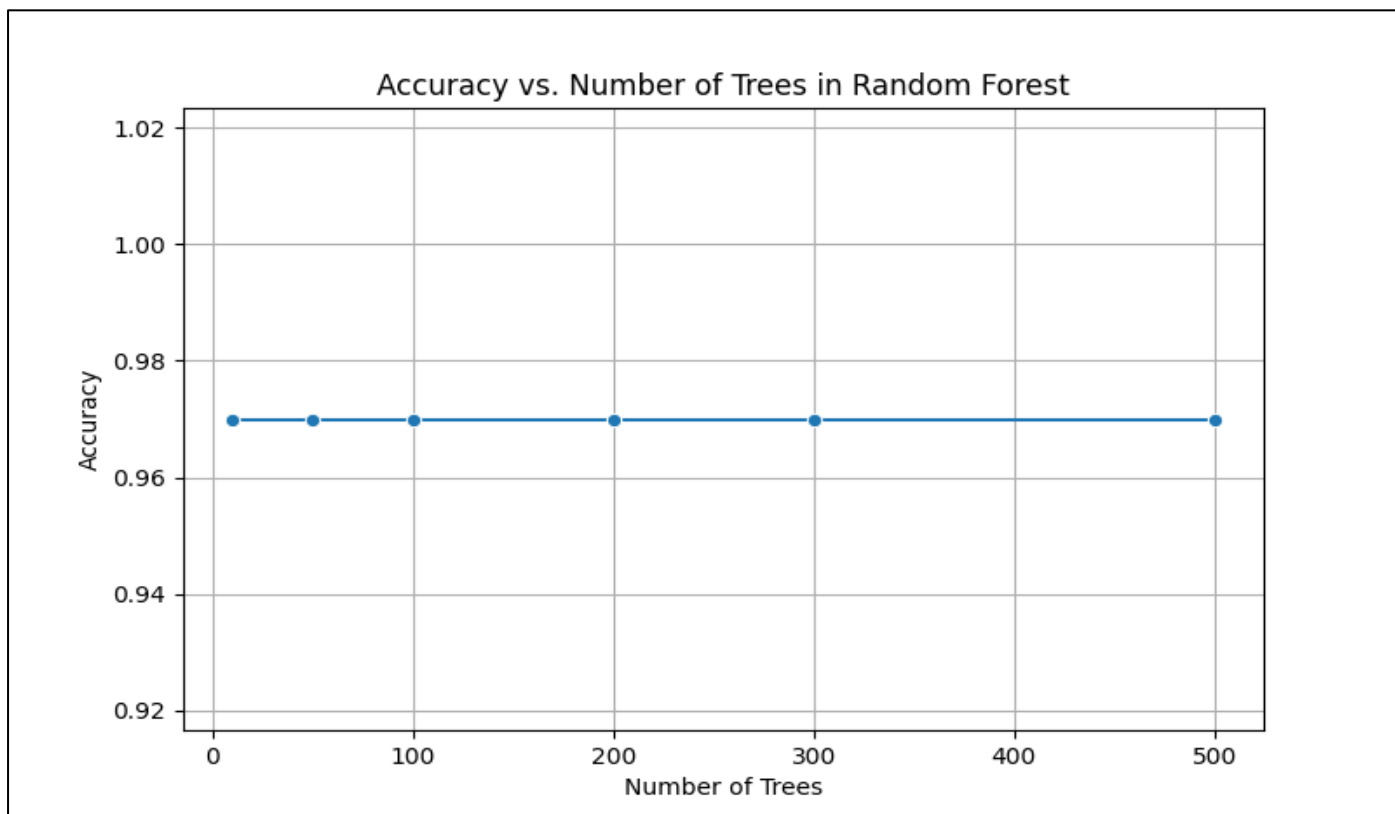


Fig 3 Accuracy vs no of Trees

- **Confusion Matrix** – A tabular representation of correct and incorrect predictions, categorized as true positives, false positives, true negatives, and false negatives.

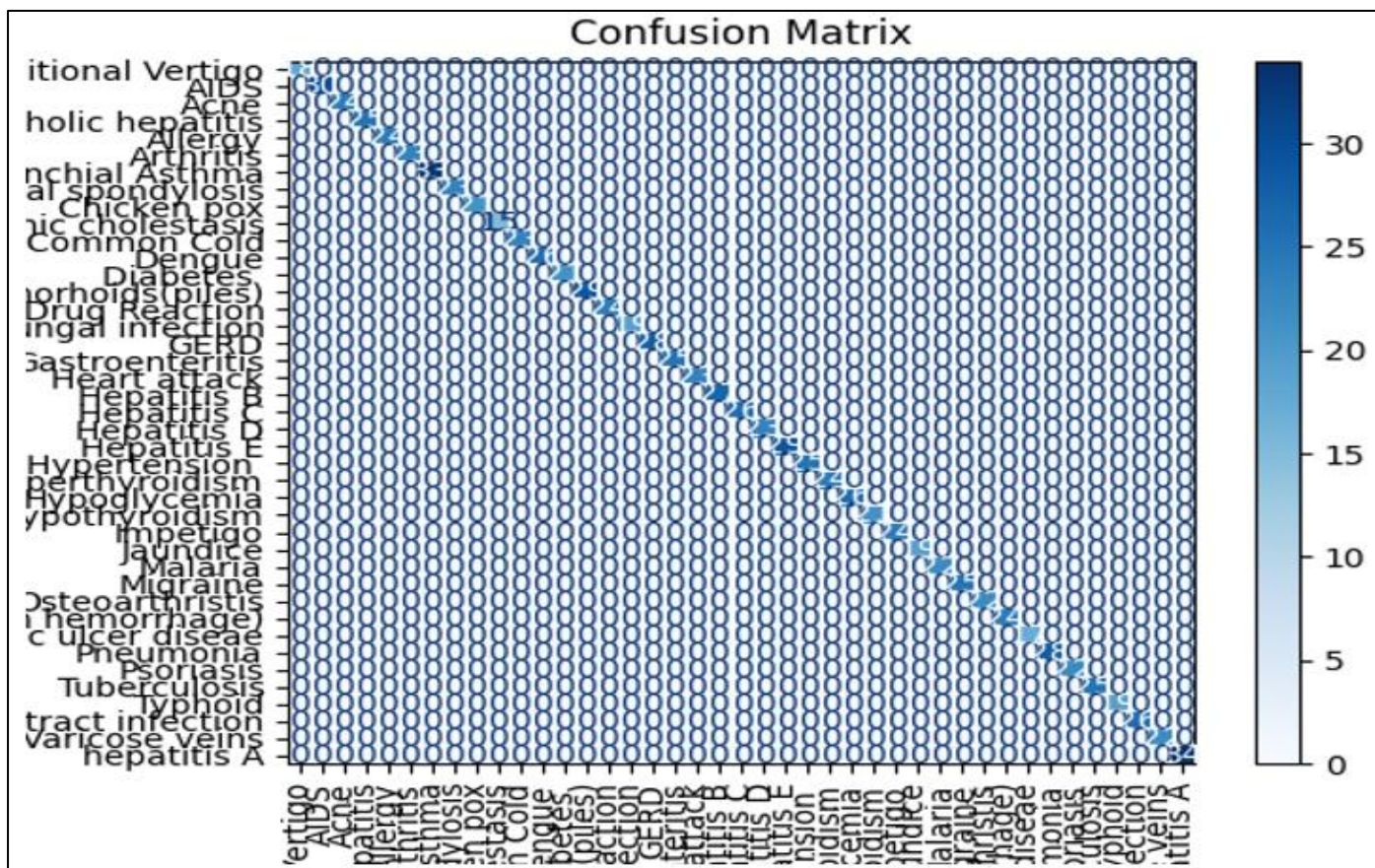
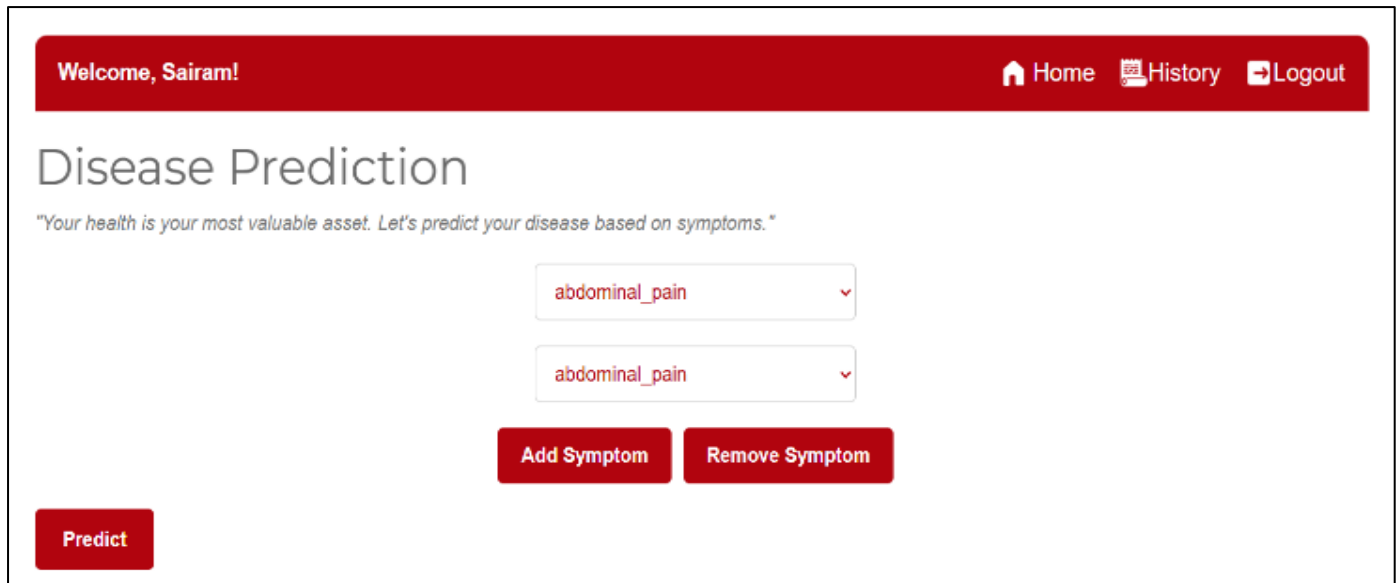


Fig 4 Confusion Matrix of Trained Model

➤ **Web Application Testing and user Experience :**

The web-based application was tested for functionality, responsiveness, and user satisfaction. Several test cases were executed to analyze how well the system handled symptom selection, disease prediction, and precautionary measure retrieval. Key observations included:

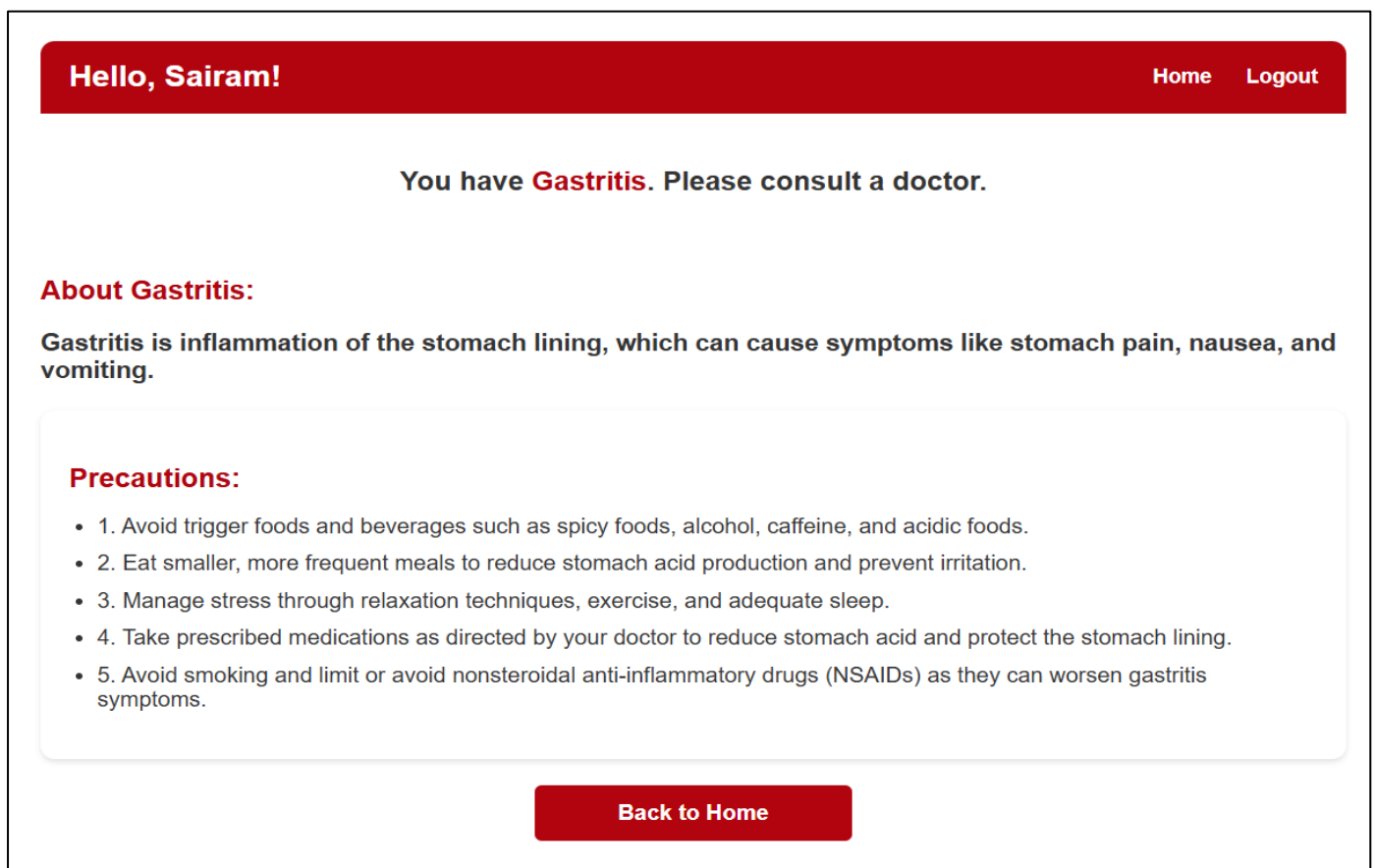
- **Fast and Accurate Predictions:** The model quickly processed symptom inputs and generated accurate disease predictions.
- **Interactive User Interface:** The web application was easy to navigate, allowing users to seamlessly input symptoms and receive results.



The screenshot shows the home page of a web application for disease prediction. At the top, a red header bar contains the text "Welcome, Sairam!" on the left and navigation links "Home", "History", and "Logout" on the right. Below the header, the main heading "Disease Prediction" is displayed, followed by a quote: "Your health is your most valuable asset. Let's predict your disease based on symptoms." Two dropdown menus are shown, both containing the text "abdominal_pain". Below these are two red buttons: "Add Symptom" and "Remove Symptom". At the bottom left, there is a red button labeled "Predict".

Fig 5 Displays the Home Page of Disease Prediction

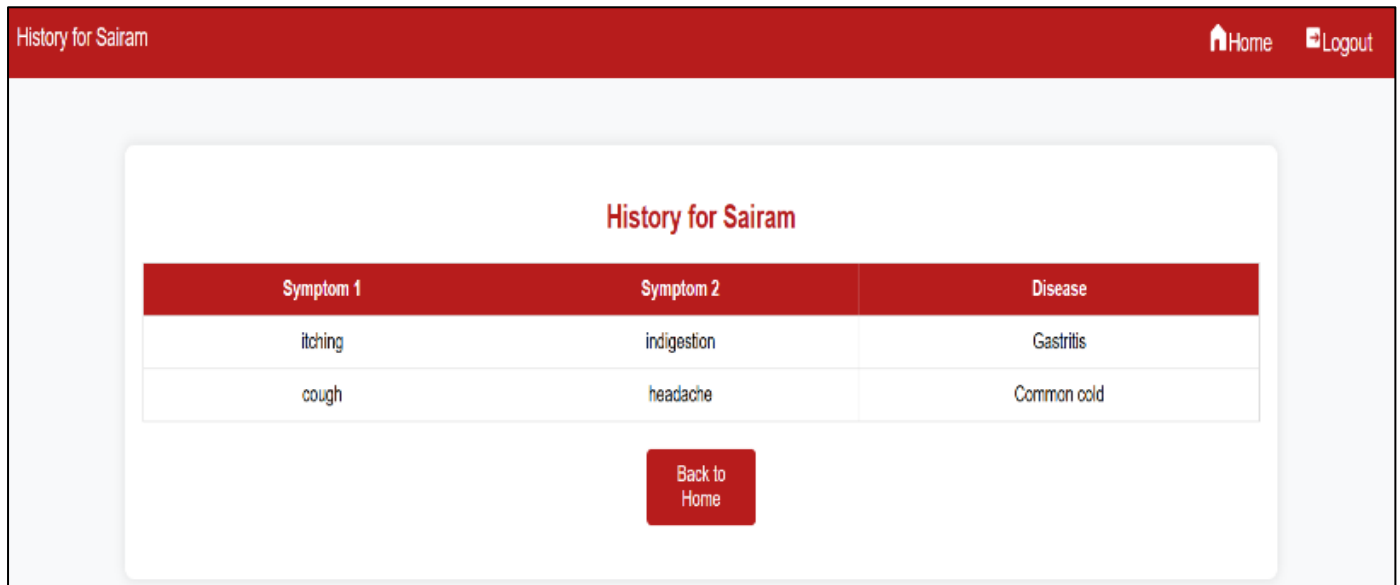
- **Precautionary Suggestions:** The system effectively fetched health recommendations using the OpenAI API, providing users with valuable guidance.



The screenshot shows a page displaying the predicted disease and precautions. At the top, a red header bar contains the text "Hello, Sairam!" on the left and navigation links "Home" and "Logout" on the right. Below the header, the main heading "You have Gastritis. Please consult a doctor." is displayed. Underneath, the section "About Gastritis:" is followed by a paragraph: "Gastritis is inflammation of the stomach lining, which can cause symptoms like stomach pain, nausea, and vomiting." Below this, a box titled "Precautions:" contains a list of five items: 1. Avoid trigger foods and beverages such as spicy foods, alcohol, caffeine, and acidic foods. 2. Eat smaller, more frequent meals to reduce stomach acid production and prevent irritation. 3. Manage stress through relaxation techniques, exercise, and adequate sleep. 4. Take prescribed medications as directed by your doctor to reduce stomach acid and protect the stomach lining. 5. Avoid smoking and limit or avoid nonsteroidal anti-inflammatory drugs (NSAIDs) as they can worsen gastritis symptoms. At the bottom, there is a red button labeled "Back to Home".

Fig 6 Displays the Predicted Disease and along with Precautions.

- **History Page:** The system securely maintains a record of user predictions, enabling easy access to previous diagnoses and corresponding symptoms.



Symptom 1	Symptom 2	Disease
itching	indigestion	Gastritis
cough	headache	Common cold

Back to Home

Fig 7 Displays the History of the user

- **User Feedback:** The application received positive reviews for its simplicity, accuracy, and usefulness.

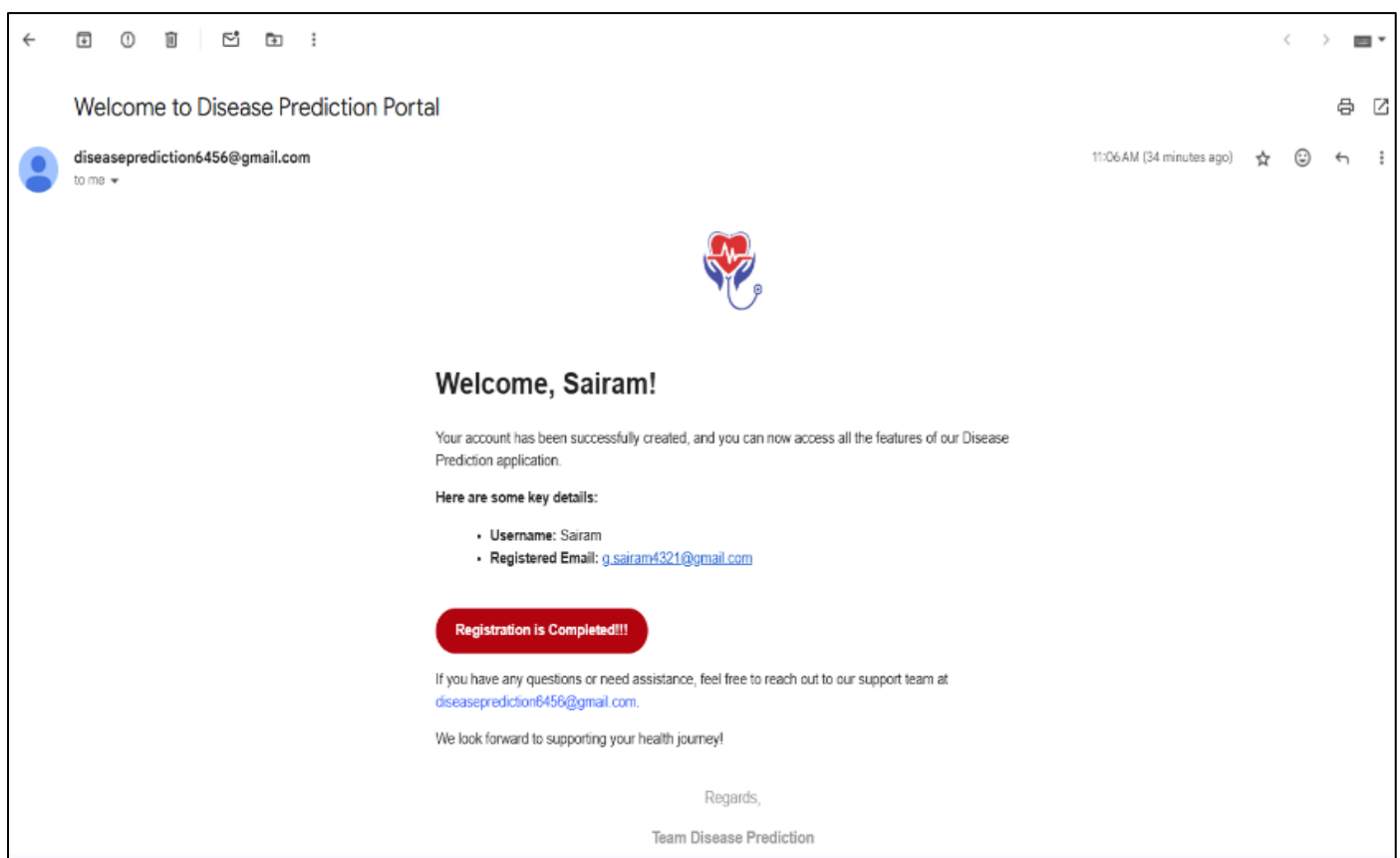


Fig 8 Mail Received by the user when user is Successfully Registered.

VIII. CONCLUSION

The Gen AI for Disease Prediction system enhances healthcare by leveraging Machine Learning (ML) and AI for accurate disease prediction. Using the Random Forest

algorithm, it automates diagnosis based on user-input symptoms, ensuring faster and more reliable results. The web-based interface allows users to select symptoms and receive instant predictions with precautionary suggestions. Integration with OpenAI's API provides personalized health insights,

making the system both predictive and advisory. Built with Django and Scikit-learn, it ensures security, scalability, and user privacy. This AI-driven approach improves early disease detection, reducing errors and supporting data-driven medical decision-making.

FUTURE SCOPE

Future enhancements of the Gen AI for Disease Prediction system will focus on improving machine learning accuracy by integrating deep learning techniques for better disease classification. The system will also incorporate Natural Language Processing (NLP) to allow users to input symptoms in natural language or through voice-based interaction for greater accessibility. Additionally, integration with wearable and IoT devices will enable real-time health monitoring, offering early warnings for potential health risks. Expanding the database will allow the system to predict a wider range of diseases, including rare conditions. Enhanced security measures, such as end-to-end encryption and compliance with data protection laws, will ensure user data privacy and safety.

REFERENCES

- [1]. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: <https://doi.org/10.1023/A:1010933404324>
- [2]. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.
- [3]. A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019. [Online]. Available: <https://doi.org/10.1056/NEJMra1814259>.
- [4]. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [5]. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: <https://doi.org/10.1162/neco.1997.9.8.173>
- [6]. Django Software Foundation, "Django Web Framework," [Online]. Available: <https://www.djangoproject.com/>
- [7]. Scikit-learn Developers, "Scikit-learn: Machine Learning in Python," [Online]. Available: <https://scikit-learn.org/>.
- [8]. McKinsey & Company. (2021). *AI in Healthcare: Transforming Diagnosis and Treatment*.
- [9]. Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118.
- [10]. Topol, E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books.