International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 04 April-2026 | Impact Factor: 3.5

MEDI AI : A DISEASE PREDICTION SYSTEM

Riddhi Dubey
UG Student, Dept of CSE(Data Science) Vidya Jyothi Institute Of Technology , Hyderabad, Telangana, India
riddhi.prasaddubey@gmail.com

G.vyshnavi
UG Student, Dept of CSE(Data Science) Vidya Jyothi Institute Of Technology , Hyderabad, Telangana, India
gundaboinavaishnavivaishnavi@gmail.com

B.Bhargavi
UG Student, Dept of CSE(Data Science) Vidya Jyothi Institute Of Technology , Hyderabad, Telangana, India
bhargavibale2005@gmail.com

Dr. R.R.S. RAVI KUMARI
Asistant Professor
Dept of CSE(Data Science) Vidya Jyothi Institute Of Technology , Hyderabad, Telangana, India
rams. rrs@gmail.com

How to Cite this Article:
Dubey, R., G.vyshnavi, & B.Bhargavi, (2026). MEDI AI : A DISEASE PREDICTION SYSTEM. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.887

https://doi.org/ 10.55041/ijcope.v2i4.887

ABSTRACT --The Medi AI: A Disease Prediction System is a machine learning-based application designed to assist in the early identification of diseases based on user-provided symptoms. In many cases, individuals ignore early symptoms due to lack of awareness, busy lifestyles, or limited access to healthcare facilities, which leads to delayed diagnosis and severe health complications. To address these challenges, this study proposes an intelligent system that predicts diseases using symptom-based input.

The system utilizes a dataset containing various diseases and their corresponding symptoms. Data preprocessing techniques such as missing value handling, encoding, normalization, and structuring are applied to improve data quality and enhance model performance. Multiple machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Naive Bayes, Support Vector Machine, and K-Nearest Neighbors are implemented and evaluated using performance metrics like accuracy, precision, recall, F1-score, and confusion matrix.

Experimental results indicate that ensemble models such as Random Forest provide higher prediction accuracy compared to other algorithms. The best-performing model is integrated into a web-based application that allows users to input symptoms and receive real-time disease predictions. The proposed system helps in early diagnosis, reduces dependency on immediate medical consultation, and improves healthcare accessibility, especially in remote areas.

I. INTRODUCTION

In today's fast-paced world, many individuals tend to ignore early symptoms of diseases due to lack of time, awareness, or access to proper healthcare facilities. This often results in delayed diagnosis and severe health complications. In rural and remote areas, the shortage of medical professionals further increases the difficulty of obtaining timely medical attention.

With the advancement of machine learning technologies, intelligent systems can be developed to assist in early disease prediction. These systems analyze symptoms and identify possible diseases based on patterns learned from medical datasets. Machine learning algorithms are capable of handling large volumes of data and identifying complex relationships between symptoms and diseases.

a. In this project, a machine learning based disease prediction system is developed to automate the process of preliminary diagnosis. The system processes user-input symptoms, performs data preprocessing and feature extraction, and applies classification algorithms to predict possible diseases. A web-based interface is also implemented to provide real-time predictions.This approach helps improve healthcare accessibility, reduce diagnostic delays, and assist users in making informed decisions regarding their health.

II. PROBLEM DEFINITION

Financial institutions receive a large number of loan applications daily, making manual evaluation difficult and inefficient. Traditional loan approval systems rely on rulebased methods and human judgment, which may lead to inconsistent decisions and delays in loan processing.

Existing systems often fail to analyze large datasets effectively and cannot accurately identify patterns that influence loan approval. In addition, factors such as incomplete data, imbalanced datasets, and complex financial relationships make the prediction process challenging.

Therefore, there is a need for an intelligent and automated system that can analyze applicant data and accurately predict loan eligibility. The proposed system addresses these challenges by applying machine learning techniques to improve the efficiency, accuracy, and reliability of loan approval decisions.

1.2 PROJECT FEATURES

The proposed Medi AI system includes several features that enhance the efficiency and accuracy of disease prediction. The system utilizes multiple machine learning algorithms to analyze symptom data and predict diseases based on user input.

Data preprocessing techniques such as handling missing values, encoding categorical data, and scaling features are applied to improve data quality. Feature engineering techniques are used to derive meaningful insights from the dataset, which helps improve prediction accuracy.

Multiple machine learning models are trained and evaluated using performance metrics such as accuracy, precision, recall, and F1-score to identify the best-performing model. The system also includes a user-friendly web application that allows users to input symptoms and receive real-time predictions.

The proposed system reduces dependency on manual diagnosis, improves consistency in predictions, and provides a scalable solution for early disease detection.

Related Work

Several research studies have explored the use of machine learning techniques for disease prediction and healthcare analytics. Traditional diagnosis methods rely on clinical expertise and rule-based systems.

Machine learning algorithms such as Logistic Regression, Decision Trees, and Random Forest have been widely used for predicting diseases based on symptoms. These models can analyze large datasets and identify patterns that influence disease occurrence.

Recent studies have also explored ensemble learning techniques to improve prediction accuracy and reduce overfitting. These models combine multiple classifiers to produce more reliable predictions.

However, many existing systems lack integration of complete preprocessing, feature engineering, and deployment. This project addresses these limitations by implementing multiple machine learning models and deploying the best-performing model through a web-based application

III. METHODOLOGY

The proposed system follows a structured approach for disease prediction.

1. Data Collection

The dataset contains information about diseases and their associated symptoms collected from reliable sources.

2. Data Preprocessing

The dataset is cleaned by handling missing values and removing inconsistencies. Categorical variables are encoded, and numerical features are scaled.

3. Feature Engineering

New features are derived to improve model performance and capture meaningful relationships between symptoms.

4. Model Training

Multiple machine learning algorithms are trained to identify patterns between symptoms and diseases.

5. Model Evaluation

Models are evaluated using accuracy, precision, recall, F1-score, and confusion matrix.

IV. PROPOSED SYSTEM

The proposed system introduces a machine learning-based approach to predict diseases using symptom data. It analyzes symptoms such as fever, headache, fatigue, cough, and other health indicators to determine possible diseases.

The dataset undergoes preprocessing steps including data cleaning, encoding, and scaling. Feature engineering techniques enhance the predictive capability of the models.

Multiple machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Naive Bayes, Support Vector Machine, and K-Nearest Neighbors are trained and evaluated.

Among these models, ensemble techniques such as Random Forest demonstrate better performance due to their ability to handle complex patterns.The final model is integrated into a web application that allows users to input symptoms and obtain instant predictions.

This system improves efficiency, reduces manual effort, and provides a reliable solution for early disease detection.

V. IMPLEMENTATION DETAILS

The system is implemented using machine learning techniques and a web-based interface. The backend is developed using Python for data preprocessing, model training, and prediction.

Libraries such as Pandas and NumPy are used for data manipulation, while Matplotlib and Seaborn are used for visualization. Machine learning algorithms are implemented using Scikit-learn.

The dataset is preprocessed by handling missing values, encoding features, and scaling data. The best-performing model is integrated into a web application that allows users to input symptoms and receive predictions.

5.1 ALGORITHMS USED

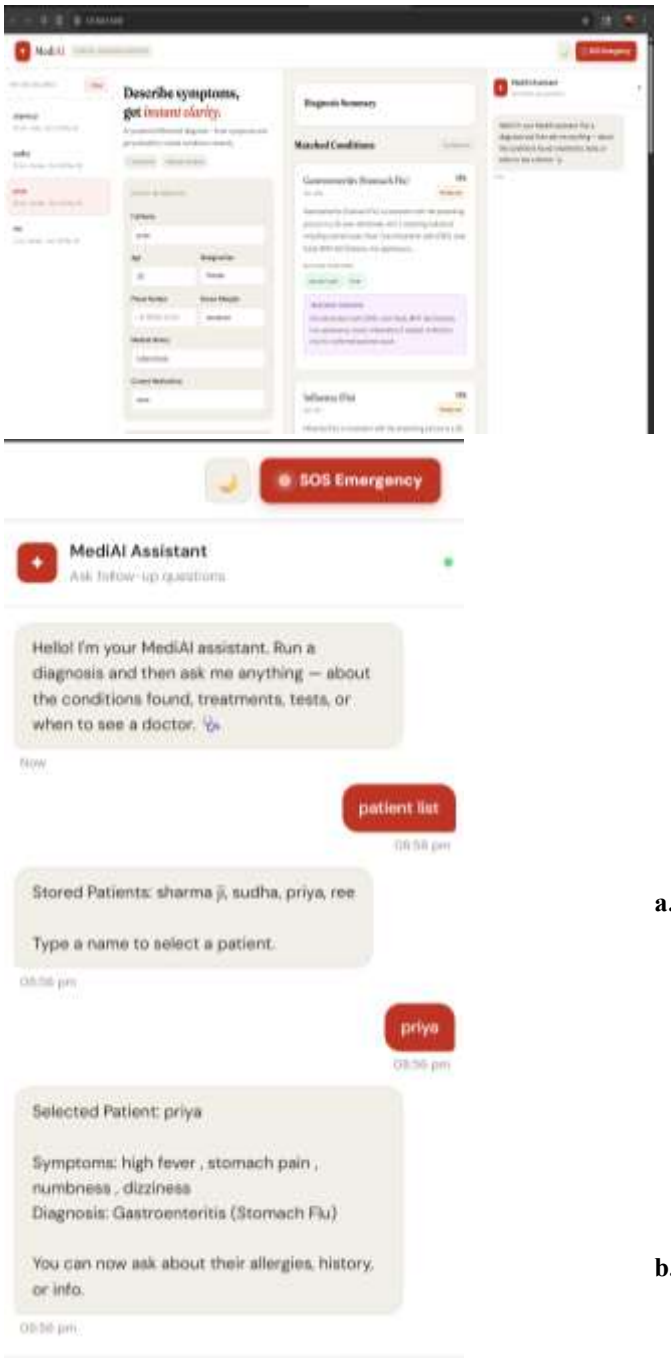
- Logistic Regression
- Used for binary classification of disease presence.
- Decision Tree
- Helps in understanding feature importance and classification.
- Random Forest
- Provides high accuracy using ensemble learning.
- Support Vector Machine
- Classifies diseases using optimal boundaries.
- K-Nearest Neighbors
- Predicts based on similarity with existing data.
- Naive Bayes
- Efficient for probabilistic classification of symptoms.

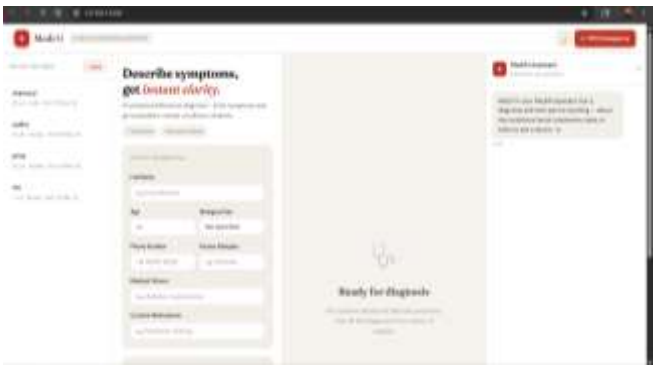
VI.EXPERIMENTAL RESULTS AND DISCUSSION

The system was evaluated using multiple machine learning algorithms. Data visualization techniques were used to understand relationships between symptoms.

Random Forest achieved the highest accuracy among all models. Other models such as Decision Tree and Logistic Regression also performed well but with slightly lower accuracy.

Future improvements include using larger datasets, integrating real-time healthcare data, and applying deep learning techniques for better accuracy.





VII. CONCLUSION

This project presents a machine learning-based system for predicting diseases using symptoms. Multiple algorithms were implemented and evaluated, with Random Forest achieving the best performance. The system helps automate early diagnosis, reduce manual effort, and improve healthcare accessibility. Integration with a web application allows real-time predictions through a user-friendly interface.

VIII. FUTURE SCOPE

The system can be enhanced by incorporating deep learning techniques such as Neural Networks. Real-time datasets from hospitals can improve model accuracy.

Additional features such as patient history and lifestyle factors can be included. Deployment on cloud platforms can enable large-scale usage.

IX.ACKNOWLEDGMENT

a. We would like to express our sincere gratitude to our project guide, Dr.R.R.S.RAVI KUMAR , Associate Professor, Department of Computer Science and Engineering (Data Science), Vidya Jyothi Institute of Technology, Hyderabad, for his valuable guidance, continuous support, and encouragement throughout the development of this project. His insightful suggestions and motivation greatly contributed to the successful completion of this work.

b. We would also like to thank the Head of the Department and faculty members of the CSE (Data Science) department for providing the necessary support and resources required for carrying out this project. We extend our sincere thanks to the Principal and management of Vidya Jyothi Institute of Technology for providing the infrastructure and academic environment that helped us complete this project successfully.

Finally, we express our heartfelt gratitude to our parents, friends, and well-wishers for their constant encouragement and support during the course of this work.

X. REFERENCES

[1] T. M. Mitchell, Machine Learning, New York, USA: McGraw-Hill, 1997.

[2] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[3] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[4] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2011.

[5] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2017. Available: https://archive.ics.uci.edu

[6] Kaggle, "Disease Prediction Using Symptoms Dataset,"Available: https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset

[7] Pandas Documentation, Available: https://pandas.pydata.org/docs/

[8] NumPy Documentation, Available: https://numpy.org/doc/

[9] Scikit-learn Documentation, Available: https://scikit-learn.org/stable/

[10] World Health Organization (WHO), "Global Health Observatory Data Repository," Available: https://www.who.int/data