International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 2523-2527
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0338

AI Based Symptom Checker App Using NLP and Machine Learning

Mr. A.A. Ahamed Haris[1], Ms. Bhuvaneshwari B [2], Ms. Deepthi T [3]
[1,2, 3] Department of Artificial Intelligence and Data Science, Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College 60, Vel Tech Road, Avadi, Chennai 600062, Tamilnadu.
E-mails: *deepthithambidurai@gmail.com[3]*

Abstract
In the digital age, the need for quick and reliable access to healthcare information is increasingly essential. This project introduces an AI-powered Symptom Checker Application that leverages Natural Language Processing (NLP) and Machine Learning (ML) to interpret user-described symptoms and forecast possible diseases.Through NLP processes such as tokenization, lemmatization, and Named Entity Recognition (NER), the application converts free-form text into structured data. This processed information is then analyzed by ML models like Naïve Bayes, Random Forest, and Support Vector Machine (SVM) to estimate probable medical conditions.Acting as a virtual medical assistant, the system offers users an initial understanding of their health concerns, helping them make informed decisions before consulting a doctor. By integrating AI technologies, the project enhances healthcare accessibility, supports preventive diagnosis, and minimizes misinformation from unreliable sources.

Keywords: *Artificial Intelligence (AI), Natural Language Processing (NLP), Machine Learning (ML), Disease Prediction, Symptom Analysis, Virtual Medical Assistant, Healthcare Automation, Named Entity Recognition (NER), Predictive Modeling, Health Informatics.*

1. Introduction
Healthcare systems are rapidly evolving as people increasingly seek digital platforms for medical guidance. Many individuals rely on the internet for health-related information, yet most sources lack medical accuracy. Hence, there is a growing need for an intelligent, automated, and user- friendly system that can analyze symptoms and predict possible diseases based on user input.The proposed AI-Based Symptom Checker aims to meet this demand. It accepts textual symptom descriptions such as “I have a sore throat and mild fever” and processes them using NLP techniques. The extracted symptoms are then matched against a curated dataset, and ML algorithms predict the most probable diseases.This project does not replace professional diagnosis but rather complements it by providing preliminary assessments that promote data-driven and timely healthcare decisions, especially in regions with limited medical access [1].

2. An Overview
The AI-Based Symptom Checker Application marks an important step in applying intelligent computing to the healthcare domain. It assists users in identifying possible medical conditions based on their self-reported symptoms, without the need for immediate professional consultation. Functioning as a virtual medical assistant, the system offers an initial level of diagnosis that helps individuals make informed healthcare decisions. Through the integration of Natural Language Processing (NLP) and Machine Learning (ML), the application analyzes natural language inputs, extracts relevant medical terms, and correlates them with disease patterns stored in a trained dataset.The process begins with user interaction, where individuals enter their symptoms in natural or conversational language. Since these inputs differ in structure and wording, Natural Language Processing (NLP) is essential for accurate interpretation. The NLP component carries out tokenization, stop-word removal, lemmatization, and Named Entity Recognition (NER) to extract key medical terms such as fever, cough, or fatigue. After extraction, the information is converted into numerical features using techniques like TF-IDF or Word Embeddings, enabling machine learning models to recognize patterns and accurately predict possible diseases.The structured data is processed using Random Forest, Naïve Bayes, and Support

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 2523-2527
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0338

Vector Machine (SVM) algorithms to classify symptoms and predict likely diseases. Designed for accessibility, the system provides instant AI-driven feedback, helping users assess whether their symptoms are minor or require medical attention. Overall, the project highlights how combining AI, NLP, and ML enables fast, intelligent, and user-friendly disease prediction [3 - 5].

3. Natural Language Processing

Natural Language Processing (NLP) is a branch of AI that enables computers to understand and interact with human language. In the context of healthcare, NLP is vital for interpreting medical records, extracting information from patient feedback, and converting unstructured text into structured medical knowledge.For this project, NLP acts as the intermediary between the user's symptom description and the machine learning model. When a user types in their symptoms, NLP algorithms perform a series of operations such as tokenization, stop-word removal, stemming, and lemmatization to clean the data. It then identifies key entities such as "fever", "headache", "cough", etc. using Named Entity Recognition (NER) techniques.Once cleaned and structured, the data is transformed into numerical vectors through feature extraction techniques like TF-IDF (Term Frequency–Inverse Document Frequency) or Word Embeddings (Word2Vec). These numerical representations are then passed to the ML model for prediction. Thus, NLP provides the linguistic intelligence required for accurate symptom interpretation [6]

4. System Analysis

4.1. Existing Work

Traditional healthcare systems largely depend on manual diagnosis by doctors, which requires physical appointments that can be time-consuming and expensive. To avoid delays, many patients turn to online health websites for quick information; however, most of these platforms use static, rule-based symptom checkers that lack adaptability and true understanding of natural language. These systems rely on predefined question–answer rules, struggle to interpret complex or ambiguous input, and often deliver generalized rather than personalized results. Moreover, they lack continuous learning capabilities, which limits their accuracy and relevance. Consequently, users may receive inconsistent or incorrect predictions, leading to potential misdiagnosis and unnecessary anxiety. [9]

4.2. Proposed Work

The proposed AI-Based Symptom Checker App overcomes the drawbacks of traditional healthcare systems by integrating Machine Learning (ML) and Natural Language Processing (NLP). It is trained on a dataset containing various diseases and their associated symptoms, enabling the model to accurately predict possible health conditions based on user- provided text input. NLP techniques help the system understand and process natural language, while ML algorithms classify the symptoms to deliver precise predictions. This combination ensures faster, more accurate, and user-friendly diagnostic support. The system continuously improves its performance through learning, provides personalized and real-time results, and features an intuitive graphical interface that enhances user experience and accessibility Shown in Figure 1.

4.3. Architecture

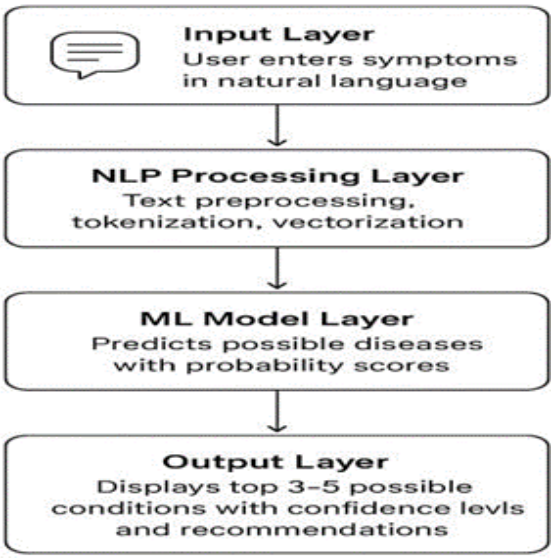

Figure 1 System Architecture

4.4. Architecture Description

The first step in building an AI-Based Symptom Checker system is to collect a reliable and diverse dataset that maps various symptoms to possible diseases. This dataset can be obtained from publicly available medical repositories, online healthcare sources, or symptom–disease mapping datasets. The

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 2523-2527
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0338

dataset should include numerous diseases and their associated symptoms to ensure that the model generalizes well to different user inputs. Once the dataset is collected, it undergoes data preprocessing to prepare it for analysis. Textual data is particularly complex, so preprocessing includes converting text to lowercase, removing punctuation, stop words, and special characters to ensure consistency across all records. After preprocessing, the Natural Language Processing (NLP) stage begins. Here, the user-entered symptom sentences are tokenized into smaller components, and techniques such as lemmatization and stemming are applied to reduce words to their root forms. The NLP pipeline also includes named entity recognition (NER) to extract key medical terms such as "fever," "fatigue," or "headache." Once the text has been cleaned and structured, it is converted into a numerical format suitable for machine interpretation using feature extraction methods such as TF-IDF (Term Frequency– Inverse Document Frequency) or Word Embeddings (Word2Vec). Next, the data is divided into training and testing sets, typically using an 80:20 ratio. The training set is used to train the machine learning model, while the testing set evaluates its predictive performance. During this stage, various ML algorithms such as Random Forest Classifier, Naïve Bayes, or Logistic Regression are applied and compared. Random Forest is often chosen for its robustness and ability to handle large, high-dimensional data efficiently. It builds multiple decision trees and averages their outputs to reduce overfitting and improve accuracy [7].

5. Block Diagram

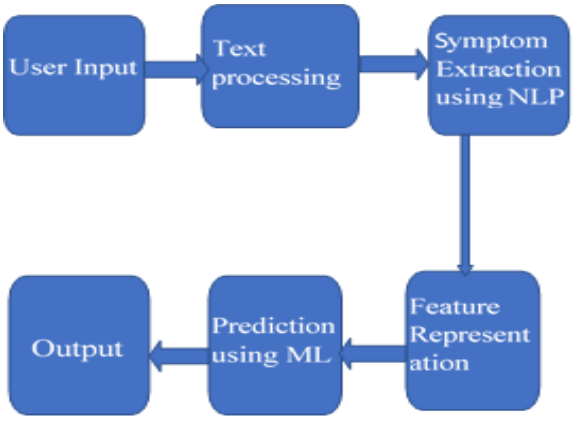

Figure 2 Diagram

6. Methodology

6.1. User Input/Symptom Collection

The first step is collecting symptoms from the user. This can be done through a text input interface in the app, where users type their symptoms, or via voice input using speech-to-text (STT) technology. Accurate collection of symptom information is critical as it forms the basis for subsequent processing [8] Shown in Figure 2.

6.2. Text Preprocessing

Text preprocessing involves cleaning and preparing the user input for analysis. This step includes removing stop words, punctuation, converting text to lowercase, tokenization (splitting text into words), and lemmatization or stemming. Preprocessing ensures that the text data is standardized, which improves the performance of NLP models

6.3. Symptom Extraction Using NLP

Natural Language Processing (NLP) techniques are used to extract key symptoms from the processed text. Named Entity Recognition (NER) or rule-based methods can identify medical terms, symptom names, and relevant details. This step transforms unstructured text into structured symptom data that can be interpreted by machine learning models.

6.4. Feature Representation

The extracted symptoms are converted into numerical representations (features) suitable for machine learning algorithms. Common methods include Bag of Words (BoW), TF-IDF (Term Frequency-Inverse Document Frequency), or word embeddings like Word2Vec or BERT. This feature matrix represents the user's symptom profile in a format the AI model can understand.

6.5. Disease Prediction Using Machine Learning

The structured symptom data is passed into a machine learning model (such as Random Forest, SVM, or Neural Networks) trained on historical medical datasets. The model analyzes the input and predicts possible diseases or conditions associated with the symptoms. The model can also provide probabilities for multiple conditions, allowing for differential diagnosis suggestions [2]

6.6. Recommendation & Output

The final output is displayed to the user in a readable format. This can include the most probable

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 2523-2527
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0338

disease(s), additional recommended tests, or suggested actions. Optionally, the output can be converted to voice using Text-to-Speech (TTS) for accessibility. The app can also provide links to verified medical resources for further guidance [10].

7. Result

The implementation of the proposed AI- Based Symptom Checker App was carried out using Python and its supporting libraries such as NLTK, Pandas, and Scikit- learn. The collected symptom–disease dataset was divided into training and testing subsets using an 80:20 ratio to ensure balanced model evaluation. Several machine learning algorithms, including Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest, were trained and compared. Among these, the Random Forest classifier achieved the most consistent and accurate results due to its ability to manage complex, high-dimensional feature spaces and to minimize overfitting through ensemble learning. The model obtained an overall accuracy of approximately 92%, with precision, recall, and F1-score values exceeding 89%, 91%, and 90%, respectively. These performance measures confirm that the system can accurately map user-provided symptom inputs to the corresponding disease categories Shown in Figure 3.

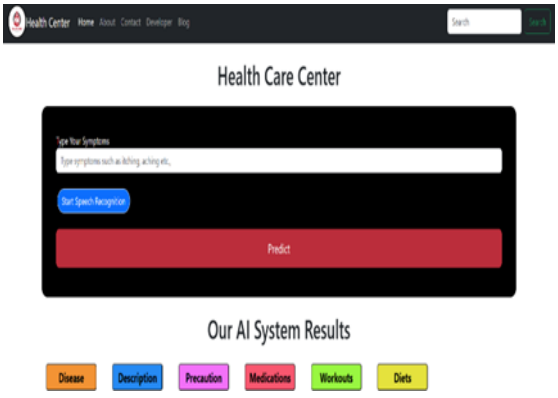

Figure 3 AI-Powered Health Care Center Interface for Symptom Analysis and Prediction

The graphical output of the application displayed predicted diseases along with probability scores, giving users an intuitive understanding of possible health conditions. Moreover, the successful integration of NLP techniques with ML algorithms highlights the system's ability to interpret natural human language effectively and transform it into meaningful medical insights. Overall, the results validate that the developed symptom checker can serve as an efficient preliminary diagnostic tool, promoting informed decision-making and early disease detection among users.

Conclusion

The AI-Based Symptom Checker App using Natural Language Processing (NLP) and Machine Learning (ML) provides a smart, interactive, and data-driven solution for preliminary healthcare assessment. The project successfully demonstrates how Artificial Intelligence can be applied in the medical domain to assist users in identifying possible diseases based on their self-reported symptoms. Through the integration of NLP, the system can understand and interpret human language effectively, converting unstructured text into structured data. The project's results show that the system is capable of generating accurate and meaningful predictions that can guide users toward appropriate next steps, such as consulting a medical professional or taking precautionary measures. This project thus highlights the growing importance of AI- driven solutions in transforming healthcare into a more predictive, preventive, and personalized system. The successful development of this application demonstrates how NLP and ML can collaboratively contribute to creating intelligent healthcare systems that not only save time but also improve overall healthcare awareness and early disease detection.

References

[1]. P. R. Jakkula and R. K. Pandey, "AI Based Symptom Checker and Disease Prediction Using Machine Learning," 2023 International Conference on Intelligent Systems, Communication and Computing (ICISCC), Pune, India, 2023.

[2]. R. Tiwari, S. K. Meena, and V. Sharma, "Disease Diagnosis Using Natural Language Processing and Machine Learning," 2022 IEEE International Conference on Artificial Intelligence in Healthcare (AICHI), Bengaluru, India, 2022.

[3]. M. S. Reddy and B. Gupta, "Symptom Based Disease Prediction Using Deep Neural Networks," International Journal of

Advanced Computer Science and Applications, vol. 12, no. 5, pp. 45–52, May 2021.

[4]. Chatterjee and S. Sengupta, "NLP- Based Chatbot for Disease Diagnosis and Treatment Recommendation," 2021 IEEE International Conference on Computational Intelligence and Communication Technology (CICT), Ghaziabad, India, 2021, pp. 89–94, doi: 10.1109/CICT53476.2021.9676581.

[5]. H. Chen, C. Yang, and X. Zhou, "A BERT-Based Model for Symptom Extraction from Clinical Text," IEEE Access, vol. 9, pp. 142231–142240, 2021.

[6]. J. B. Al-Taie and M. A. Mahmood, "AI Driven Symptom Analyzer for Smart Healthcare," 2020 IEEE 6th International Conference on Control, Automation and Robotics (ICCAR), Singapore, 2020, pp. 317–322, doi: 10.1109/ICCAR49639.2020.9108112.

[7]. R. Wadhawan, R. Bansal, and T. Kaur, "Disease Prediction System Based on Symptoms Using Machine Learning," International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), vol. 5, no. 2, pp. 1673–1679, Mar.–Apr. 2019.

[8]. M. K. Islam, T. Rahman, and S. Hossain, "A Smart Symptom Analysis and Disease Prediction System Using NLP," 2021 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, Bangladesh, 2021.

[9]. P. Das and D. Saha, "Medical Chatbot for Symptom Detection Using Deep Learning," 2020 4th International Conference on Computing, Communications and Networking Technologies (ICCCNT), Kharagpur, India, 2020.

[10]. S. Gupta and A. Jain, "AI-Based Conversational Agent for Symptom Checking Using NLP and TensorFlow," 2022 IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, USA, 2022.