International Journal of Computer Techniques–IJCT Volume 13 Issue 3, May-June - 2026

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

An AI-Based Disease Prediction System for Early Healthcare Diagnosis

Sachin Kumar[1], Kajal Singh[2], Harshita Sharma[3], Divyanshu Upadhyay[4], Rupendra Kaushik[5]

Noida Institute of Engineering & Technology, Greater Noida

(Affiliated To Dr. APJ Abdul Kalam Technical University, Lucknow)

***Abstract* — Artificial intelligence is being utilized in healthcare to support decision-making and achieve early disease detection. This article proposes an artificial intelligence-based disease prediction system that determines diseases based on symptoms input by the user. The Random Forest classifier was selected for the proposed system due to its speed and accuracy on medical data. The dataset is preprocessed by cleaning, encoding, and feature selection to increase the prediction accuracy. A basic interface is designed to let the user input the symptoms and receive the prediction result instantly. It should be noted that the system does not replace a medical practitioner but can provide early insight for the user to take appropriate action. The experimental results indicate that the model achieves high accuracy and generalizes well over other cases.-**

Keywords - ***Crisis Detection, Social Media Analysis, Natural Language Processing, Multi-Layer Perceptron, TF-IDF, Geospatial visualization, Emergency Management.***

I. INTRODUCTION

The transformations in healthcare industry in the recent years have seen enormous adoption of Artificial Intelligence in medical systems. In the past disease diagnoses was based solely on clinical observation, laboratory testing and medical professionals' experience. Although those methods are still valuable, due to the time-consuming, costly and inaccessible nature of each technique they can be difficult for people in economically deprived remote areas.

With the progress in computational technologies and accessibility of digital health records, AI-based systems can be seen as assistance tools for faster and accurate predictions of diseases. In contemporary life style erratic meals stress pollution and lack of exercise etc. have caused an exponential rise in chronic diseases like diabetes, heart, respiratory diseases, liver disorders etc. Most of the time a patient is so forgetful that he does not notice the early symptoms of a disease or he only realizes it when the disease has come to advanced stages.

Early diagnosis and treatment can significantly improve and reduce the treatment cost. This increasing demand for early detection has motivated researchers to look into intelligent systems that can examine medical data and discover disease patterns. Artificial Intelligence, specifically its subset known as Machine Learning Algorithm, has shown great promise and effectiveness in healthcare implementations based on historical data-driven prediction by learning underlying patterns of symptoms, personal history, and clinical and biochemical parameters. Different from rule-based systems, ML models improve their results over time as they incorporate more data for learning. Decision Tree, Support Vector Machine, Naïve Bayes, and Random Forests are some of the most common ML Algorithms utilized for disease predictions due to its ability to process big datasets with complex feature associations.

This paper proposes an AI based Disease Prediction System for predicting possible diseases based on the symptoms given by users. The system is intended to be an intelligent assistance system to provide initial medical advice professional consultation. A structured preprocessing pipeline is employed to clean and transform raw medical data into a suitable format for training. The processed data is then used to train a Random Forest classifier, chosen for its high accuracy, robustness, and ability to reduce overfitting.

The proposed system is developed with practical usability in mind. It operates on standard computing hardware, requires minimal technical knowledge to use, and provides predictions through a simple and user-friendly interface. The major contributions of this work are as follows:

1- A comprehensive preprocessing pipeline for handling medical symptom datasets, including data cleaning, encoding, normalization, and feature selection.

2- A computationally efficient machine learning model capable of accurately predicting diseases while maintaining low hardware requirements.

3- A user-friendly prediction platform that allows individuals to input symptoms and receive instant disease predictions in real time.

II. RELATED WORK

Predicting diseases in health care system has been researched using different computational techniques for a long time. Before the era of Artificial Intelligence, the early health care

prediction systems were mainly developed using rule-based expert systems based on predefined medical rules that were manually coded to support doctors in diagnosis. However, these approaches were only very basic and could not adapt to new relationships among symptoms, medical history and patient conditions. Besides, with the increase of medical data complexity and volume, rule-based health care prediction systems could not make accurate predictions in large-scale diseases.

The next step in the development of medical prediction systems was statistical machine learning-based methods, such as Naive Bayes, Decision Trees and Support Vector Machines (SVM) which have shown better performance in identifying disease patterns from structured data. To increase prediction performance, researchers combined these algorithms with feature extraction and data preprocessing techniques. For example, SVM-based systems performed well in predicting heart disease and diabetes because of their strong capability in classifying high-dimensional medical data.

Nevertheless, many of these methods were still based on handcrafted features and could not perform well when the datasets had missing, noisy or imbalanced data .The advancement of ensemble learning techniques improved the predictive capabilities of healthcare applications even further. Random Forest and Gradient Boosting models were popular because they handle overfitting and complex feature interactions more efficiently than standalone classifiers. Several studies found that the prediction accuracy of Random Forest models were high in predicting liver disorders, cardiovascular diseases, and chronic kidney disease. These methods are also more interpretable than deep neural networks.

More recently, there has been interest in deep learning architectures such as Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), or Recurrent Neural Networks (RNNs) for disease prediction and medical diagnosis. These models can learn highly complex patterns from large datasets, and have shown promise in areas such as cancer detection, medical imaging, and patient risk assessment.

Recent advancements in transformer-based architectures and deep learning frameworks have also contributed to enhancing the predictive capabilities of modern healthcare systems. However, these models require large training datasets, and powerful GPU hardware and computational resources. In this research, we propose a practical middle-ground by proposing a Random Forest-based disease prediction system that balances prediction accuracy, computational resources, and ease of deployment. Instead of using complex deep learning architectures, we propose a lightweight model that can effectively run on standard hardware after careful preprocessing of medical datasets. The proposed system is a suitable candidate for healthcare assistance systems in real-world applications.

III. PROBLEM FORMULATION

Disease prediction is a supervised multi-class classification problem. Suppose the data set is
$D = \{d_1, d_2, \ldots, d_n\}$, where each instance d_i is a record of patients' symptoms and other Medical Attributes. The label y_i is from the set of disease $C = \{c_1, c_2, \ldots, c_k\}$, where diseases are such as diabetes, heart disease, liver disease etc.

The goal of our system is to learn a mapping function:

$$f : \mathbb{R}^n \rightarrow C,$$

where the n is the medical features extracted. By studying the relationship between symptoms disease, the model can predict the best results for unseen patients.
In the preprocessing step, after data cleaning, data encoding, data normalization, feature selection to enhance the quality of data and reduce redundancy to the data set. In the final step, Random Forest is taken as the prediction model because of high accuracy and robustness.

Accuracy, Precision, Recall and F1-Score are used to evaluate the performance of the model. Our aim is to maximize the prediction accuracy with high computational efficiency and usability.

IV. SYSTEM DESIGN AND METHODOLOGY

A. Architectural Overview

The proposed system is structured using a modular architecture to ensure extensibility, maintainability and efficient disease prediction. The entire pipeline in the proposed framework is composed of four major modules:

1) Data Collection Module

As the data collection module, structured datasets that include symptoms, disease labels and patient specific attributes are gathered from publicly available datasets in CSV format for training and evaluation of the models. The modular ingestion layer to the system is designed to make it extensible to future integration with Electronic Health Records and hospital databases.

2) Preprocessing Pipeline

The raw medical datasets are usually not in standard form for training and have missing values, inconsistent formatting, duplicate values, categorical attributes etc. that cannot be directly fed to the machine learning models. The preprocessing pipeline transforms the raw data into a structured, numerical format for training.

3) Prediction Engine

The prediction engine loads the trained Random Forest model and performs inference by predicting the possible disease if user provided symptoms as input. The stateless design used for the architecture ensures quick inference and multiple requests for predictions can be served efficiently without the need for training.

B. Preprocessing Pipeline

Medical Data Preprocessing for Predicting Diseases.

One of the problems of medical data is that the information can be incomplete, inconsistent and noisy, which directly influences the prediction results. Hence, the preprocessing process is needed to transform the raw medical data into a clean and structured one for machine learning. The preprocessing pipeline includes the following steps:

- Data cleaning

Missing values or duplicate data are detected and handled. The duplicate data is deleted, and the missing values are filled using the suitable methods. All these actions can increase the reliability and quality of the data.

- Encoding

Medical data usually contains many categorical attributes, like symptoms and diseases, that cannot be directly processed by the learning models. Encoding methods are used to transform these textual values into numbers for model training.

- Normalization

Different medical features have different number ranges. Thus, the features are normalized into a common range to prevent those with larger values from dominating the learning process.

- Feature selection

Some attributes of the data set are not related or are redundant for disease prediction. These irrelevant or less relevant attributes can be removed after selecting the most important feature attributes.

- Data balancing

Imbalanced data is a common issue of medical data sets. The number of the records of some diseases can be much larger than the others. Some data balancing techniques such as oversampling and under sampling can be used to solve this problem.

- Data splitting

The final data set is split into training and test sets. The training data is used to train the model, and the test data is used to validate the model's performance on unseen data.

C. FEATURE *EXTRACTION -TF-IDF*

Feature extraction is essential to mapping the most relevant medical attributes contributing to disease prediction. Symptoms, medical history and diagnostic indicators are encoded as feature vectors that can be used as inputs for prediction models.

The feature space aims to remove irrelevant attributes and find the most effective ones for disease prediction, in order to help prediction models to perform well while keeping computational requirement low.

4) User Interface Layer

The presentation layer includes a simple graphical interface to make predictions instantly by providing the symptoms. The UI is designed to be user-friendly for non-technical users to perform well while keeping computational requirement low

D. MLP CLASSIFIER ARCHITECTURE.

The backbone of the prediction model is trained by using Scikit-Learn's implementation of the Random Forest algorithm. Random Forest is an ensemble learning method that builds multiple decision trees to combine predictions for better classification and reduce overfitting.

The architecture is described as follows:

- Several decision trees are built using random subsets of the dataset.
- Decision trees are used to predict disease category independently.
- Majority vote is used to determine the final prediction.
- Random Forest was chosen due to its high predictive accuracy, ability to handle medical data sets with many features, and low computational cost compared to deep learning models.

The model is trained using a balanced data set that contains categories with multiple diseases. Hyperparameters such as number of trees and maximum depth are optimized to balance performance with low computational cost..

E. GEOSPATIAL EXTRACTION AND VISUALIZATION

The system we propose features a user-friendly and simple interface that easily allows patients to input symptoms and determine disease predictions instantaneously. The interface is designed on Flask to create the interaction between the user and the model smoothly. After the input of symptoms, the trained Random Forest classifier will predict the disease for the patient and return the prediction result instantly.

Furthermore, to demonstrate interpretability, we present basic visualizations, including disease distribution, prediction accuracy, and feature importance, using Matplotlib, and help patients understand the prediction process better and better contribute to the practical applications in healthcare.

F. IMPLEMENTATION DETAILS

The proposed AI-based Disease Prediction System is built using Python due to its simplicity, flexibility, and strong support for machine learning and healthcare applications. The entire system takes a modular approach, which includes data preprocessing, model training, prediction generation, visualization, and user interface integration. Because the project uses only open-source technologies, the development cost is low. The system can be easily deployed on standard computing devices without the need for expensive infrastructure.

1. Programming Environment

The project is developed with Python 3.8, using Jupyter Notebook and Visual Studio Code. Python is selected for its wide range of libraries for machine learning, data analysis, and web development. The implementation is broken into separate modules, which makes it easier to maintain, update, and scale in the future.

2. Libraries and Tools Used

Several libraries and tools support different functions within the system:

- Pandas: Used for loading, organizing, and cleaning the dataset
- NumPy: Used for numerical calculations and array operations
- Scikit-Learn: Used for preprocessing, feature selection, and implementing the Random Forest classifier
- Matplotlib: Used for generating graphs and visualizing results
- Flask: Used for developing the web-based user interface
- Pickle: Used for saving and loading the trained machine learning model

These libraries make the implementation process smoother and improve the system's efficiency.

3. Dataset and Preprocessing

The medical dataset is stored in CSV format and contains symptoms along with their corresponding disease labels. Before training, the data goes through preprocessing steps such as cleaning missing values, encoding categorical attributes, normalization, feature selection, and balancing. These steps improve data quality and help increase prediction accuracy.

V. RESULT

Results of this research show that an AI System can accurately predict disease from medical symptoms using a structured medical set of data. The method evaluated that the AI predicted claims based upon the symptoms and disease claims with a minimum margin for error based upon the symptoms shown. As a result, multiple ML algorithms performed comparably with respect to the accuracy of each method.

. TABLE I. MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.82	0.80	0.79	0.79
Random Forest	0.86	0.84	0.83	0.83
Random Forest (Proposed)	0.912	0.90	0.88	0.89

4. Model Training

The Random Forest classifier is trained on the preprocessed dataset. The model builds multiple decision trees during training, and the final prediction comes from majority voting. Hyperparameters like the number of trees and maximum depth are adjusted carefully to boost performance and minimize overfitting.

5. Prediction Process

The prediction system follows this sequence:

- The user enters symptoms through the web interface.
- The symptoms are converted into numerical values.
- The processed data is passed to the trained Random Forest model.
- The model analyzes the input using patterns learned during training.
- The system predicts the most likely disease.
- The prediction result is shown instantly to the user.

This workflow ensures quick and real-time disease prediction.

6. User Interface Integration

A simple and interactive web interface is developed with Flask. The interface allows users to enter symptoms easily and receive prediction results right away. The design is user-friendly, so even non-technical users can operate the system without trouble.

7. Hardware Requirements

The system is lightweight and does not need special hardware to operate. Both training and prediction can run efficiently on a standard laptop with an Intel Core i5 processor and 8 GB of RAM. The lightweight design ensures faster response times, scalability, and easy deployment in educational and healthcare settings.

Baseline models evaluated with identical preprocessing and train/test splits

The results from the experimental study indicate that the Random Forest classifier performed best out of all the models that were evaluated; therefore, the Random Forest classifier is superior to Logistic Regression and Decision Tree classifiers. The Random Forest classifier had an overall classification accuracy of 91.2%, meaning that it is very capable of identifying diseases from symptoms reported by users.

Furthermore, Random Forest showed better generalization as well as greater stability across the various classes of diseases, than either Logistic Regression or Decision Tree classifiers.

To evaluate the reliability of Random Forest predictions, both precision (90%) and recall (88%) of predictions were computed. A precision score of 90% means that most of the diseases predicted by Random Forest were indeed correct; a recall score of 88% means that most of the actual disease cases contained in the dataset were successfully identified by the Random Forest classifier. Therefore, this information is especially critical for healthcare applications because misdiagnosis or missed diagnosis could have a significant negative effect on the early diagnosis and treatment of disease.

With an F1-Score of 0.89, the model demonstrates a good balance between precision & recall. The random forest algorithm is an ensemble model, helping to decrease the chance of overfitting and allowing for better prediction of complex relationships between symptoms than an individual classifier can. To determine the robustness of the proposed disease prediction system, additional testing was performed using different subsets of the clinical database for testing purposes. The model produced consistent results, with only small deviations in the prediction accuracy, indicating that it has excellent generalization abilities. The prediction model also produced rapid responses during real time prediction, because the results were produced almost immediately following inputting of symptoms.

Classification and prediction of disease were evaluated using visualization techniques to assist with understanding disease distribution, feature importance, and measuring model performance. The graphs create visualizations confirming that some symptoms more strongly contribute to the prediction of disease than others, thus creating a better understanding of, and more interpretability into, how the model produces predictions.

Thus, these predictive modeling results show the capability of this proposed AI-based disease prediction model to provide accurate and dependable, as well as computationally efficient, predictions. The lightweight implementation and high predictive performance of the proposed disease prediction model present excellent options for use in real-world health care applications as well as for education related to health care.

VI. DISCUSSION

One of the major problems with disease prediction systems is the fact that many diseases have similar symptoms. An accurate diagnosis is difficult, because symptoms such as fever, fatigue, headache and chest pain are common to a number of medical conditions. In these cases, conventional rule-based systems often fail, as they rely on fixed medical rules rather than learning complex relations between symptoms.

This limitation can be mitigated by machine learning models that extract hidden patterns from medical data.

The proposed system performed well with the ensemble learning by using Random Forest classifier. Instead of building a single decision tree , the model takes the prediction from multiple trees . This makes the model more accurate and less likely to overfit . It enables the system to be more capable of dealing with complex combinations of symptoms, which in turn results in stable predictions for different cases of patients.

Another significant aspect that affects predicted accuracy is dataset quality. Medical datasets often contain errors such as incomplete records, duplicate records, errors in their data, or classification imbalance of diseases. If these errors are not processed correctly, this could lead to poor model performance.

In this project, the preprocessing pipeline for cleaning, normalizing, encoding, and feature selection helped to provide reliability and efficiency in the resulting system.

While the proposed model is highly accurate, it also has some drawbacks. One major limitation of the model is that its outputs are based entirely on symptom-only inputs, and do not take other forms of patient information, such as lab results, images or medical history into account in any considerable way. As such, the outputs of this model could vary significantly even for diseases that share many of the same symptoms. In addition to the above considerations, the accuracy of the system will be dependent on the quality and diversity of the training dataset that has been used to develop the model.

Overall, the proposed AI Disease Prediction System offers a practical balance of accuracy, computational efficiency and ease of use. Its lightweight implementation means that the system can run on ordinary computer systems without requiring expensive computing resources, and with future enhancements such as the addition of real-time healthcare information, support for multiple languages and the application of advanced deep learning techniques, the system will be able to serve as a highly effective tool for early disease identification and the provision of healthcare support.

The proposed AI-based Disease Prediction System shows that machine learning can effectively assist in early disease detection using user symptoms. The Random Forest classifier achieved high accuracy and provided fast predictions with low computational cost. The system is lightweight, user-friendly, and suitable for real-world healthcare applications, especially in areas with limited medical facilities.

Advantages of the system:

- High prediction accuracy
- Fast and real-time results
- Low hardware requirements
- User-friendly web interface
- Easy future scalability

The proposed system can also help increase healthcare awareness among users by encouraging early medical consultation based on symptom analysis.
However, the system mainly depends on symptom-based input and does not include laboratory reports or medical imaging data. Future improvements may include deep learning integration, real-time healthcare data, and multilingual support.

VII. CONCLUSION

The research presented here is an AI-based Disease Prediction System that employs machine learning techniques to predict diseases based on user symptoms.

The methodology used in this implementation consists of data preprocessing, feature selection, and using a Random Forest classifier to provide accurate and efficient predictions of diseases.

Further, the proposed model can produce reliable predictions by analyzing the patterns of medical symptoms. Additionally, it is able to predict with a low level of computational complexity.

Through the experimental results of this project, it was found that the Random Forest classifier was superior to the other evaluated models (in terms of accuracy, precision, recall, and F1-score). In addition, the proposed implementation provided high prediction accuracy with low weight and thus could be deployed on standard hardware easily.

Furthermore, the implementation included a web-based user interface, which allowed users to submit their symptoms and receive their evaluated predictions almost instantaneously; therefore, making the system more accessible to the user. The implementation of this research demonstrates how machine learning can be effectively used to provide early detection of disease and subsequently provide assistance in the health care systems. The system works efficiently but there is a future scope for improvement. The existing model is highly dependent on symptom based input and does not take into account advanced medical information like laboratory reports or medical imaging.

In summary, the proposed system highlights the increasing ability of Artificial Intelligence to improve healthcare access, early diagnosis and decision-making support.

The Random Forest classifier produced accurate and reliable results while maintaining low computational complexity. The system provides quick disease prediction through a simple web interface and can support early diagnosis in healthcare applications.

Key contributions:

- Accurate disease prediction
- Lightweight and low-cost implementation
- Easy deployment on standard systems
- Real-time prediction capability

In conclusion, the proposed system demonstrates the potential of Artificial Intelligence in improving healthcare accessibility, early diagnosis, and decision-making support.

VII REFRENCES

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

[2] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, Cambridge, MA, USA: MIT Press, 2016.

[5] World Health Organization, “Artificial intelligence in healthcare,” WHO Report, 2021.

[6] D. Dua and C. Graff, “UCI Machine Learning Repository,” University of California, Irvine, 2019.

[7] S. Kotsiantis, “Supervised machine learning: A review of classification techniques,” *Informatica*, vol. 31, no. 3, pp. 249–268, 2007.

[8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[9] A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine,” *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[10] K. Sharma and R. Gupta, “Disease prediction using machine learning algorithms,” *International Journal of Advanced Research in Computer Science*, vol. 10, no. 2, pp. 45–50, 2020.