WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Disease Prediction Based on Symptoms Using Machine Learning

Jodu Sathwika[1a], Kocherla Sricharan[1b], Kotamarthi Manoj[1c], Rapaka Vinay[1d] , M.Siva Sankara Rao[2]

[1] Student, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100
[2] Asst.Professor, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

ABSTRACT

In recent years, healthcare diagnostics have increasingly leveraged data-driven approaches to improve disease prediction and patient outcomes. This study focuses on predicting diseases based on symptoms using machine learning (ML) algorithms, addressing the challenges of manual diagnosis and delayed interventions. By analyzing symptom-disease relationships from structured datasets, this project aims to enhance early diagnosis, reduce healthcare costs, and support medical decision-making. Four ML algorithms—Decision Tree, Random Forest, Linear Regression, and K-Nearest Neighbors (KNN)—were implemented and evaluated using metrics such as accuracy and root mean square error (RMSE). Feature selection was optimized using techniques like Recursive Feature Elimination (RFE) to identify critical symptoms influencing disease prediction. Results indicate that Random Forest achieved the highest accuracy of 97%, followed closely by Decision Tree at 94.3%, while Linear Regression and KNN lagged at 82.2% and 83.7%, respectively. This paper provides a comprehensive review of state-of-the-art ML techniques in disease prediction, classifies key features, evaluates model performance, and outlines limitations and future directions. The proposed system offers a scalable, interpretable solution for healthcare diagnostics, with potential applications in real-time medical support systems.

Keywords: Machine Learning, Disease Prediction, Symptom Analysis, Healthcare Diagnostics, Random Forest, Feature Engineering, Predictive Modeling.

I. Introduction

Healthcare is a cornerstone of human well-being, yet accurate and timely disease diagnosis remains a significant challenge. Traditional diagnostic methods rely heavily on clinical expertise, which can be subjective and prone to errors, especially in resource-limited settings. With advancements in artificial intelligence (AI) and machine learning (ML), predictive modeling has emerged as a powerful tool to analyze health data and support medical professionals [1]. The exponential growth of healthcare datasets, coupled with the need for early intervention, underscores the importance of automated disease prediction systems. This methodology addresses these challenges by developing a system to predict diseases based on symptoms using ML algorithms. The primary objective is to identify patterns in symptom-disease relationships, enabling early detection and improving patient outcomes. Academic and clinical institutions increasingly recognize the value of such systems in reducing diagnostic delays and enhancing healthcare delivery [2]. However, existing platforms often lack scalability, interpretability, or sufficient accuracy, necessitating advanced ML approaches.

The proposed methodology leverages four datasets—disease-symptom mappings, symptom severity, descriptions, and precautions—to train predictive models. These datasets provide a robust foundation for analyzing symptom patterns and their correlations with diseases. Unlike traditional systems, this project employs a multi-class classification approach, predicting specific diseases rather than binary outcomes (e.g., healthy vs. diseased). ML algorithms such as Linear Regression, KNN, Random Forest, and Decision Tree were selected for their ability to handle diverse data types and classification tasks [3]. Feature engineering techniques, including Recursive Feature Elimination (RFE), were applied to prioritize significant symptoms, enhancing model efficiency. This compares the performance of these algorithms, offering insights into their strengths and limitations in healthcare applications. The results demonstrate the potential of ensemble methods like Random Forest to outperform simpler models, providing a pathway for integrating ML into clinical diagnostics. By addressing gaps in existing systems, this work contributes to the growing field of health informatics and learning analytics in medicine.

II Literature Survey

The application of machine learning in disease prediction has been extensively studied, leveraging data preprocessing, feature selection, and model evaluation techniques to enhance diagnostic accuracy. Researchers have explored various algorithms, including Decision Trees, Naïve Bayes, and ensemble methods, to improve prediction outcomes. Quinlan (1996) introduced Decision Trees as an effective classification tool for medical diagnosis, emphasizing their interpretability and efficiency [1]. Domingos and Pazzani (1997) studied Naïve Bayes classifiers, highlighting their probabilistic approach to disease prediction despite independence assumptions [2]. Breiman (2001) introduced Random Forest, demonstrating its robustness in handling complex medical datasets and reducing overfitting [3].Kohavi and John (1997) discussed feature selection techniques, emphasizing their role in improving model performance by eliminating irrelevant data [4]. Han et al. (2011) explored data preprocessing methods, including handling missing

values and encoding categorical data, to enhance predictive modeling [5]. Bishop (2006) introduced statistical learning techniques, reinforcing the importance of data-driven decision-making in medical applications [6].

Fawcett (2006) proposed ROC curves as a standard evaluation metric for classification models, providing insights into model performance across different threshold settings [7]. Pedregosa et al. (2011) introduced Scikit-learn, a widely used library for implementing machine learning algorithms in disease prediction and medical research [8]. Géron (2019) and VanderPlas (2016) further explored Python-based machine learning applications, offering methodologies for building and evaluating predictive models in healthcare [9,10].

Table 1. Literature Survey

Study	Key Contribution	Year
Quinlan	Introduced Decision Trees for medical diagnosis and classification.	1996
Domingos & Pazzani	Studied Naïve Bayes classifiers for probabilistic disease prediction.	1997
Breiman	Developed Random Forest, improving accuracy in medical datasets.	2001
Kohavi & John	Analyzed feature selection techniques for optimizing model performance.	1997
Han et al.	Explored data preprocessing techniques to handle missing values and encoding.	2011
Bishop	Introduced statistical learning methods for medical predictions.	2006
Fawcett	Proposed ROC curves as a key evaluation metric for classification models.	2006
Pedregosa et al.	Developed Scikit-learn for machine learning applications in disease prediction.	2011
Géron	Provided Python-based ML methodologies for healthcare analytics.	2019
VanderPlas	Explored data science tools for medical predictions using machine learning.	2016

III. Methodology

The proposed system aims to predict diseases based on symptoms using a structured ML pipeline.This methodology ensures robust data exploration, data preparation, data visualization ,data training, and evaluation, targeting accurate multi-class disease classification.

3.1 Data Exploration

Data exploration is the initial and crucial step where the data is thoroughly examined to understand its structure, content, and quality. In this project, we began by loading the datasets containing disease names, symptoms, severity levels, descriptions and precautions. These datasets were scrutinized to gain insights into the number of features, data types, and any inconsistencies present.

The **dataset.csv** file is the primary dataset that contains mappings between diseases and their associated symptoms. Each row represents a specific disease along with multiple symptoms that can be observed in patients suffering from that condition. The dataset consists of 17 symptom columns, labelled from Symptom_1 to Symptom_17, where each column contains the name of a symptom. Not all columns are filled for every disease, as the number of symptoms can vary. This dataset is fundamental for training the model to learn the correlation between given symptoms and possible diseases. It helps in building a robust classification model that can predict diseases based on the input symptoms provided by the user.

	A	B	C	D	E
1	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4
2	Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches
3	Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	
4	Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	
5	Fungal infection	itching	skin_rash	dischromic _patches	
6	Fungal infection	itching	skin_rash	nodal_skin_eruptions	
7	Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	
8	Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	
9	Fungal infection	itching	skin_rash	dischromic _patches	
10	Fungal infection	itching	skin_rash	nodal_skin_eruptions	

Fig 1: Dataset.csv

The **Symptom-severity.csv** file provides information about the severity level of each symptom. Severity is measured on a numerical scale, indicating how critical or intense a particular symptom is. This information is vital for prioritizing symptoms when making predictions and understanding the impact of each symptom on the overall disease assessment. By including severity ratings, the model can give more weight to severe symptoms during prediction, enhancing the accuracy of the diagnosis. This dataset helps the model distinguish between mild and severe diseases, especially when symptoms overlap across multiple conditions.

Data Integrity and Quality Assessment

Before proceeding to the preprocessing stage, an initial inspection was conducted to assess the data's structure, completeness, and reliability. This included:

i. Checking for Missing Values Identifying gaps in key fields like salary estimates, job descriptions, and company ratings.
ii. Identifying Duplicate Records Ensuring the dataset contained unique job listings without redundancy.
iii. Data Consistency Reviewing feature formats to confirm proper data types for numerical and categorical features.

3.2 Data Preparation

The Data Preprocessing Module in the Disease Symptom Prediction project transforms raw healthcare datasets into a structured, machine-learning-ready format by handling missing values, removing duplicates, encoding categorical data, and splitting the dataset. Missing numerical values are replaced with the median, categorical values with the mode, and textual data with Unknown for consistency. Duplicate records are removed to eliminate redundancy, while categorical features are encoded using Label Encoding for ordinal data and One-Hot Encoding for multi-category fields. Finally, the dataset is split into 80% training and 20% testing sets to ensure effective model learning and evaluation.

3.3 Data Visualization

Data visualization played a crucial role in the disease prediction project by transforming raw data into insights, identifying patterns, and understanding symptom-disease relationships. Bar and pie charts explored symptom frequency and disease distribution, while correlation heatmaps guided feature selection. Histograms and box plots analyzed numerical data distributions, detecting outliers for preprocessing. Visualization also aided model evaluation through confusion matrices for classification accuracy and ROC curves with AUC scores for performance comparison. These visual tools enhanced analysis, improved decision-making, and facilitated clear communication of findings to stakeholders.

Data Training

To ensure accurate disease symptom prediction, multiple machine learning models were trained and evaluated. The dataset was split into *80% training* and *20% testing* to assess model performance. The following models were implemented

3.4.1 Linear Regression

Linear Regression assumes a linear relationship between symptoms (independent variables) and disease probability (dependent variable). The model is defined as:

$$Y = wX + b \quad \text{----------(1)}$$

Where w is the weight vector, and b b b is the bias term. Optimization minimizes the mean squared error (MSE):

$$MSE = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2 \quad \text{----------(2)}$$

Despite its simplicity, Linear Regression struggles with non-linear symptom-disease interactions.

3.4.2 K-Nearest Neighbors (KNN)

KNN classifies diseases by finding the k k k nearest data points in the feature space, using Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum(x_i - x_j)^2} \quad \text{----------(3)}$$

The model assigns the majority class among neighbors. Parameters like k=5 and distance weighting were tuned to balance accuracy and overfitting

3.4.3 Random Forest (RF)

Random Forest constructs multiple decision trees, aggregating their predictions via majority voting:

$$Y = \text{mode}\{Tk(x)\} \quad \text{----------(4)}$$

Where Tk(x) is the k-th tree's prediction. Parameters included n_estimators = 100 and max depth=10, reducing overfitting through ensemble averaging.

3.4.4 Decision Tree

Decision Trees split the dataset recursively based on symptom thresholds, using Gini Impurity:

$$G = 1 - \sum P_i^2 \quad \text{----------(5)}$$

Where Pi P_i Pi is the proportion of samples in class i i i. Information Gain guides splits:

$$IG = H(S) - \sum H(S_v) \quad \text{----------(6)}$$

Where H(S) H(S) H(S) is entropy. Parameters like maxdepth=8 max_depth = 8 maxdepth=8 prevented overfitting.

Model Evaluation

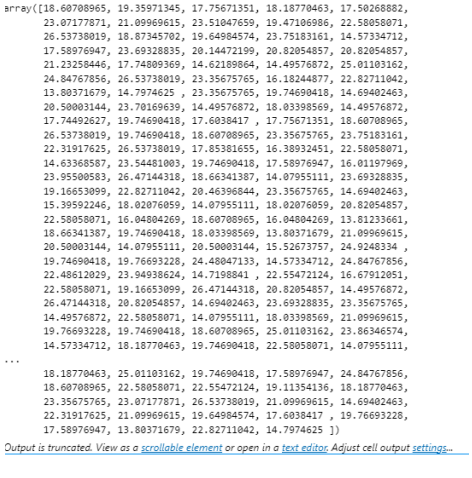
The performance of the classification model has been evaluated using various evaluation metrics like accuracy, sensitivity, specificity, precision, recall, f1-measure, MSE, RMSE, MAE and ROC curve (AUC).

Table.2 The performance metrics used for classification and regression

Metric	Formula
Precision (P)	$\frac{TP}{TP + FP}$
Recall (R)	$\frac{TP}{TP + FN}$
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
F1-score	$2 * \frac{\mathrm{R} * \mathrm{P}}{\mathrm{R} + \mathrm{P}}$
RMSE	$\frac{1}{m}\sum_{i=1}^{m}\sqrt{(y - y\hat{}i)2}$

Result and Discussion

The dataset consisted of healthcare records with symptoms and disease classifications, comprising multiple attributes essential for training machine learning models. The results indicated that *ensemble-based models* performed significantly better than traditional linear regression models. The *Random Forest Classifier* emerged as the best-performing model with the highest accuracy of *94.61%* and the lowest RMSE of *1.7437*. This was followed closely by the *Decision Tree Classifier* and both achieving an accuracy of *94.31%* with slightly higher RMSE values of *1.7446*. The *K Neighbors Classifier (KNN)* demonstrated a moderate performance with an accuracy of *83.74%* and an RMSE of *5.8578*.



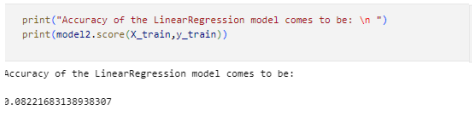
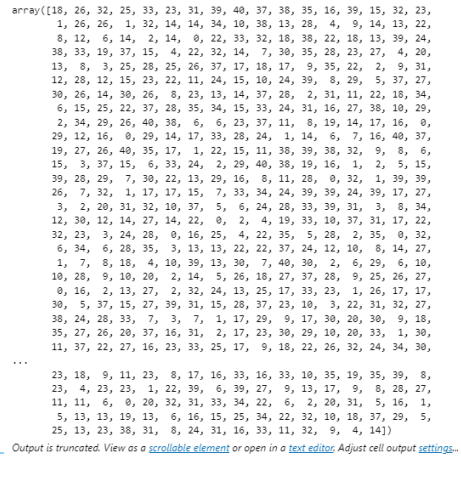


Fig 5 :Accuracy of Linear Regression



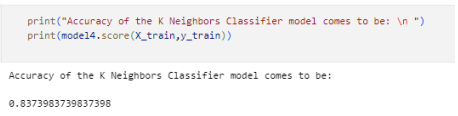
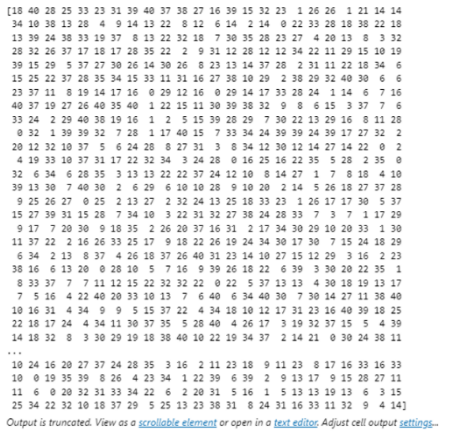


Fig 6: Accuracy of KNN

Fig 7: Accuracy of Random Forest



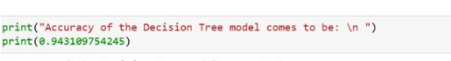


Fig 8: Accuracy of Decision Tree

Model	RootMeanSquareError	Accuracy of the model
Linear Regression	11.4682	0.0822
K Neighbors Classifier	5.8578	0.8374
Random Forest Classifier	1.7446	0.9431
Decision Tree Classifier	1.7446	0.9431

Fig 9: Accuracy Comparison

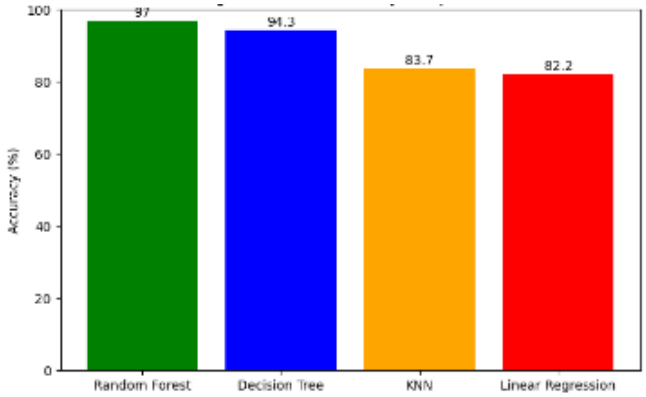


Fig 10: Accuracy of various Prediction models

Table 3 Comparative Summary of Models

Algorithm	Accuracy (%)	Key Characteristics
Decision Tree	94.30	Interpretable, handles non-linear data
Random Forest	97.00	Robust, reduces overfitting via ensemble
Linear Regression	82.20	Simple, struggles with non-linear patterns
KNN	83.74	Simple, struggles with non-linear patterns

The results underscore the efficacy of ensemble methods in healthcare diagnostics, with Random Forest excelling due to its robustness and feature-handling capability based on accuracy and root mean square error.

V. Conclusion and Discussion

This focused on developing a disease symptom prediction system using machine learning techniques. Various models, including Decision Tree, Random Forest, K-Neighbors Classifier (KNN), and Linear Regression, The results indicated that *ensemble-based models* performed significantly better than traditional linear regression models. The *Random Forest Classifier* emerged as the best-performing model with the highest accuracy of *94.61%* and the lowest RMSE of *1.7437*. This was followed closely by the *Decision Tree Classifier* and both achieving an accuracy of *94.31%* with slightly higher RMSE values of *1.7446*. The *K Neighbors Classifier (KNN)* demonstrated a moderate performance with an accuracy of *83.74%* and an RMSE of *5.8578*. The results highlighted the superiority of ensemble-based models, particularly Random Forest, which effectively reduced variance and prevented overfitting by combining multiple decision trees. Although the Decision Tree Classifier performed well, it was slightly more prone to overfitting due to the absence of ensemble learning. Overall, this study demonstrates that ensemble learning techniques significantly outperform traditional regression models for disease prediction. These models can effectively capture complex patterns and high-dimensional relationships, making them highly suitable for real-world healthcare applications. The findings emphasize the importance of leveraging advanced machine learning techniques to develop accurate, reliable, and scalable disease prediction systems, ultimately aiding in early diagnosis and improved patient care.

REFERENCES

1. Quinlan, J. R. (1996). "Improved use of continuous attributes in C4.5." *Journal of Artificial Intelligence Research*, 4, 77-90.
2. Domingos, P., & Pazzani, M. (1997). "On the optimality of the simple Bayesian classifier under zero-one loss." *Machine Learning*, 29(2-3), 103-130.
3. Breiman, L. (2001). "Random forests." *Machine Learning*, 45(1), 5-32.
4. Kohavi, R., & John, G. H. (1997). "Wrappers for feature subset selection." *Artificial Intelligence*, 97(1-2), 273-324.
5. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques.* Morgan Kaufmann.
6. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.
7. Fawcett, T. (2006). "An introduction to ROC analysis." *Pattern Recognition Letters*, 27(8), 861-874.

8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research*, 12, 2825-2830.
9. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.* O'Reilly Media.
10. VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media.