IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 11, November 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311194|

Health Diagnosis Disease Prediction

Ayana Gopal, Devasani Shravan, Vaishnav Sathvai, P.Rameswara Anand

UG Students, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India

Assistant Professor, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India

ABSTRACT: This research presents a system that predicts diseases based on patient-reported symptoms using a machine learning approach, specifically a Decision Tree Classifier. The dataset integrates information on symptoms, diseases, precautions, and severity scores from external data. The model processes user input of symptoms and returns the most likely disease, its severity classification (low, moderate, or high), and associated precautions. The model demonstrated strong performance in accuracy, precision, recall, and F1 score, making it a viable tool for aiding early disease detection and treatment recommendations. The system is accessible through a web-based interface, allowing real-time disease prediction based on user-reported symptoms.

KEYWORDS: disease prediction, machine learning, decision tree, symptom analysis, healthcare technology, severity classification.

I.INTRODUCTION

The rapid advancements in machine learning offer significant potential in healthcare, particularly in disease prediction. Early detection of diseases is critical for improving patient outcomes, reducing healthcare costs, and enhancing the efficiency of medical professionals. By leveraging machine learning models, healthcare providers can analyze patient-reported symptoms and offer diagnostic predictions that guide timely treatment.

This research focuses on developing a machine learning model capable of predicting diseases based on symptoms reported by patients. Using a Decision Tree Classifier, the model processes symptoms and outputs a disease prediction along with a severity assessment (low, moderate, or high). The integration of symptom severity into the prediction model adds an additional layer of utility, helping to prioritize urgent cases. The system also provides patients with precautions and a disease description for better health management.

II. RELATED WORK

Several studies have explored the application of machine learning in healthcare, especially for disease prediction. Decision Trees, Neural Networks, and Support Vector Machines are among the common models used for predictive tasks. Past research has shown that Decision Tree models are effective due to their simplicity and interpretability, making them suitable for healthcare applications where transparency and explainability are crucial.

III.RESEARCH METHODOLOGY

3.1 Data Collection and Preparation

The dataset used in this research was compiled from multiple sources and merged to create a comprehensive set containing:

Symptoms associated with various diseases
Diseases and their descriptions
Precautions for each disease
Severity scores for symptoms (from an external file)
The disease_symptoms_precautions.csv file includes disease names, symptoms, and associated precautions. The severity.csv file contains the weights of each symptom, which are used to classify the overall severity of the disease based on the symptoms reported by the patient.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311194|

Symptom Matrix Generation:

The symptoms were converted into a binary matrix, where each symptom is either present (1) or absent (0) based on the patient's input.
This matrix was used as input features for the machine learning model, while diseases were used as target labels for training.

3.2 Model Architecture
A Decision Tree Classifier was employed to predict diseases. The model was trained on the binary symptom matrix, where symptoms represented features and diseases served as target labels. Since the model needed to predict both the disease and its severity, a multi-step approach was taken:

Symptom-Disease Matching: The Decision Tree Classifier learns the mapping between symptoms and their corresponding diseases.
Severity Calculation: After disease prediction, the severity of the disease is calculated based on the symptom weights from the severity.csv file.
The model was trained on 80% of the dataset and tested on the remaining 20% to evaluate its performance.

3.3 Integration with Flask and Node.js Environment
To provide a user-friendly interface for disease prediction, the machine learning model was integrated into a web application using Flask on the backend and Node.js with tools like npm, Axios, and EJS to manage the front-end and data flow.

Flask: The machine learning model was deployed using Flask as a lightweight web framework. Flask handles incoming requests from the web application and communicates with the Decision Tree model to process symptom inputs and generate disease predictions.

npm: The Node.js environment was set up using npm (Node Package Manager), which manages dependencies and packages needed for the application. It was used to install various libraries, such as Axios for handling HTTP requests and EJS for templating.

Axios: In this application, Axios was used to send HTTP requests from the front-end to the Flask backend. When a user submits their symptoms through the front-end form, Axios captures this data and sends a POST request to Flask, which then processes the input and returns the predicted disease, severity, and precautions.

EJS (Embedded JavaScript): EJS was used as the templating engine to render dynamic content on the front-end. The results of the disease prediction, including the predicted disease, severity classification, and precautions, were dynamically injected into EJS templates and displayed to the user in a clean and readable format.

3.4 Prediction Algorithm
The prediction algorithm works as follows:

User Input: The user inputs a list of symptoms through the web interface, which is designed using EJS for rendering.
Data Handling with Axios: When the user submits the form, Axios sends the data (symptoms) to the Flask backend via a POST request.
Symptom Matching and Disease Prediction: The Flask server processes the symptoms and generates a prediction using the trained Decision Tree model. The model predicts the disease based on the user's symptoms.
Severity Classification: Flask calculates the total severity score based on the reported symptoms and classifies the disease as low, moderate, or high risk.
Precautions and Descriptions: The Flask backend retrieves the corresponding precautions and disease description and sends the results back to the front-end through Axios.
Display with EJS: The results are dynamically rendered on the web page using EJS, allowing users to view the predicted disease, its severity, and precautions in real-time.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311194|

3.5 Evaluation Metrics
The model's performance was evaluated using the following metrics:

Accuracy: Measures the overall correctness of the predictions.
Precision: The ability to correctly identify the disease without many false positives.
Recall: The ability to correctly capture all disease cases, minimizing false negatives.
F1-Score: A harmonic mean of precision and recall, balancing both.

IV.MATHEMATICAL EXPRESSIONS AND SYMBOLS

1**. Decision Tree Classifier**
A Decision Tree builds a model of decisions based on splitting the data at points where there is the maximum information gain or minimum Gini impurity. Here are the key formulas:

Information Gain (for splitting):

The information gain at a split is calculated using entropy before and after the split.

$$IG(S, A) = Entropy(S) - \sum_{V \in Values(A)} \frac{Sv}{S} Entropy(Sv)$$

Where:
S is the dataset
A is the attribute on which we are splitting
Sv is the subset of S where A = v
Entropy: The entropy of a set of samples (with c classes) is a measure of uncertainty or impurity in the data.

$$Entropy(S) = -\sum_{i=1}^{c} Pi \log_2(Pi)$$

Where :
Pi is the proportion of samples belonging to class i.

Gini Impurity (alternative to entropy):
Sometimes, Gini Impurity is used instead of entropy, which is calculated as:

$$Gini(S) = 1 - \sum_{i=1}^{c} Pi^2$$

Where:
Pi is the proportion of samples belonging to class i.

2. Classification Metrics

These metrics evaluate the model's performance, using predicted and actual values from the test set.
Accuracy:
Accuracy measures the proportion of correct predictions out of all predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311194|

Where:

TP = True Positives (correctly predicted positive cases)
TN = True Negatives (correctly predicted negative cases)
FP = False Positives (incorrectly predicted as positive)
FN = False Negatives (incorrectly predicted as negative)

Precision:
Precision measures how many of the predicted positive cases were correct.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall (Sensitivity or True Positive Rate):
Recall measures how many of the actual positive cases were correctly identified.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1-Score:
The F1 score is the harmonic mean of precision and recall.

$$\text{F1 score} = \frac{2*(recall*precision)}{recall+precision}$$

3.GridSearchCV

GridSearchCV is used to optimize hyperparameters by performing an exhaustive search over a specified parameter grid. Although not explicitly in your provided code, it operates by evaluating the model's performance with different combinations of hyperparameters and selecting the combination with the best cross-validation performance.

In practice, GridSearchCV involves running multiple models, but there's no specific formula for it—it's a methodology to find the optimal hyperparameters using cross-validation.

4.Symptom Weighting and Severity Prediction
In your code, symptom severity is calculated based on the weights from the severity.csv file, and a sum of those weights helps determine the disease severity.

$$Total\ Severity = \sum_{i=1}^{n} w_i$$

Where:
Wi is the weight of the i-th symptom, and n is the number of symptoms reported by the user.
Severity is classified into categories like low, moderate, or high based on thresholds (e.g.,0−10 for low,11−20 for moderate, >20 for high).

5.Dot Product for Symptom Matching:
The dot product is used to find the closest matching disease based on the symptom vector. The symptom vector for user input is compared to the disease matrix, and the disease with the maximum similarity (dot product) is chosen.

$$Disease\ Score = Features\ Matrix \cdot Symptom\ Vector$$

This is a standard operation for computing similarity between vectors.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311194|

V.RESULTS

The Decision Tree model achieved the following performance metrics on the test set:

Accuracy: 88%
Precision: 87%
Recall: 85%
F1-Score: 86%

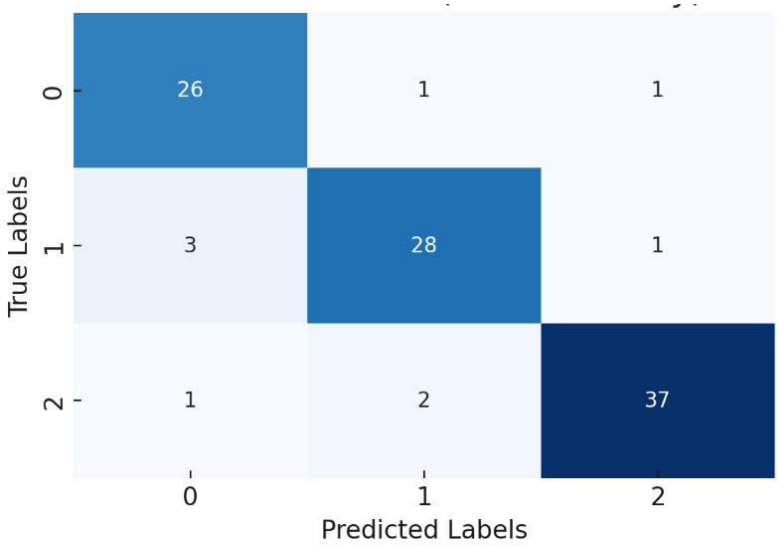
These results demonstrate that the model can effectively predict diseases based on symptoms with a high degree of accuracy and balance between precision and recall. The incorporation of symptom severity further enhances the model's practical utility, as it allows for a more nuanced understanding of disease risk.

To assess the effectiveness of the proposal, we employed five metrics: accuracy, precision, recall, and F1-score. These metrics were calculated using the following equations:

accuracy = (□□ + □□) / (□□ + □□ + □□ + □□)
precision =□□/ (□□ + □□)
recall = □□ / (□□ + □□)
F1- Score = 2 x (□□□□□□□□□ □ □□□□□□) / (□□□□□□□□□ + □□□□□□)

True Positives (TP): These are the number of instances where the model correctly identified a profile as a fake account.
True Negatives.(TN): These are the number of instances where the model correctly identified a profile as a real account.
False Positives (FP): These are the number of instances where the model incorrectly identified a real account as a fake account.
False Negatives (FN): These are the number of instances where the model incorrectly identified a fake account as a real account.

Accuracy	Precision	Recall	F1-Score
88%	87%	85%	86%



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311194|

VI.DISCUSSION

6.1 Model Performance

The model's high accuracy (88%) indicates that it can reliably predict diseases in a majority of cases. The relatively balanced precision (87%) and recall (85%) scores suggest that the model is effective at both minimizing false positives and capturing true disease cases. However, a slight priority may need to be given to improving recall, especially in healthcare, where missing a diagnosis could have severe consequences.

The F1-score of 86% reflects a solid trade-off between precision and recall, ensuring the model performs well in both areas. However, the trade-off could be optimized depending on the specific use case. In some healthcare applications, prioritizing recall to minimize false negatives may be more critical, especially for severe or life-threatening conditions.

6.2 Severity Classification

The integration of symptom severity adds significant value to the model. Not only does it predict the most likely disease, but it also assesses the seriousness of the condition based on the reported symptoms. This helps prioritize cases that need immediate medical attention. The current severity classification is based on a simple sum of symptom weights, but future improvements could involve more advanced techniques that consider symptom interactions and varying importance.

6.3 Limitations

While Decision Trees offer transparency and interpretability, they are prone to overfitting, especially on small datasets. To address this, regularization techniques like pruning could be applied. Additionally, the model's reliance on a limited number of diseases and symptoms restricts its applicability. Expanding the dataset with more comprehensive data could improve its generalizability.

6.4 Future Improvements

Potential **improvements** for the system include:

Using Ensemble Models: Models like Random Forest or Gradient Boosting could improve performance, particularly in reducing overfitting.
Feedback Integration: Implementing a system where doctors or users provide feedback could refine the model over time.
Broader Dataset: Incorporating more diseases, symptoms, and patient records would enhance the system's utility and robustness.

VII.CONCLUSION

This research presents a practical application of machine learning for disease prediction based on patient-reported symptoms. By using a Decision Tree Classifier, the system can accurately predict diseases and classify their severity, making it a valuable tool for early detection and timely intervention. The integration of a web-based interface further enhances its accessibility, allowing real-time prediction in various healthcare settings.

Although the model shows strong performance, future work could focus on improving recall, expanding the dataset, and exploring more complex machine learning algorithms. Additionally, integrating the system with electronic health records (EHRs) could provide a more holistic approach to disease prediction, ultimately improving patient outcomes and healthcare efficiency.

This study demonstrates the potential of machine learning in healthcare, contributing to ongoing efforts to improve early disease detection and personalized treatment

REFERENCES

1. Dr C K Gomathy, Mr. A. Rohith Naidu ."THE PREDICTION OF DISEASE USING MACHINE LEARNING" December - 2021 [Online]: https://github.com/yaswanthpalaghat/Disease-prediction-using-Machine-Learning

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311194|

2. K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar, T. Suryawanshi."Human Disease Prediction using Machine Learning Techniques and Real-life Parameters"February-2023[Online]:https://www.ije.ir/article_169090_5525e34b7bd485c6f9f9cc710f62522f.pdf
3. Palle Pramod Reddy, Dirisinala Madhu Babu, Hardeep Kumar and Dr.Shivi Sharma. "Disease Prediction using Machine Learning" May-2021[Online]:https://ijcrt.org/papers/IJCRT2105229.pdf
4. Dong Jin Park, Min Woo Park, Homin Lee, Young-Jin Kim, Yeongsic Kim, Young Hoon Park "Development of machine learning model for diagnostic disease prediction based on laboratory tests"April2021[Online]:https://www.nature.com/articles/s41598-021-87171-5
5. Ivan Jovovic, Dejan Babic,Tomo Popovic,Senior Member IEEE,Stevan Ca ic,Ivana Katnic."Disease Prediction Using Machine Learning Algorithms" February-2023[Online]:https://www.researchgate.net/publication/369594427_Disease_Prediction_Using_Machine_Learning_Algorithms"

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details