IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SYMPTOMS BASED DISEASE DIAGNOSIS USING MACHINELEARNING

[1] Vasavi Sravanthi Balusa, [2] P Jaya Praksh Rao, [3] Adithya vardhan Reddy [4] G Shiva Sanjana

[1] Assistant Professor

[1] Department of Computer Science and Engineering,

[1] Methodist College of Engineering and Technology, Hyderabad, India.

Abstract: Now a days, people face various diseases due to environmental condition and their living habits. So the prediction of disease at an earlier stage becomes an important task. The correct prediction of disease is the most challenging task. To overcome this problem Machine Learning plays an important role to predict the disease. Medical science has a large amount of data growth per year. Due to the increasing amount of data growth in the medical and healthcare field the accurate analysis of medical data has been benefits from early patient care. With the help of disease data, machine learning finds hidden pattern information in a huge amount of medical data. With the help of disease symptoms data set disease prediction is done by using Machine learning algorithms like Random forest and Support Vector Machine.

Index Terms - **Machine Learning, Disease, Symptoms, Efficiency, Supervised, Accuracy.**

I. INTRODUCTION

Health is one of the most important assets of our life, in today's hustle and bustle of life, most people neglect their health.

Disease diagnosis is a critical and decisive aspect of healthcare. Accurate and timely diagnosis can help in effective treatment and management of a patient's health problems. The conventional diagnostic methods have been based on the physician's clinical experience and expertise. Healthcare systems worldwide are facing increasing pressure due to rising patient numbers, limited resources, and the need for accurate and timely diagnoses.

Machine Learning can help us overcome these challenges and improve our health such as detecting diseases, reducing costs, diagnosing diseases, recommending treatments, and designing health policies. In today's era of advancing technology and healthcare, predictive analytics has emerged as a powerful tool for early detection and proactive management of diseases. Our project seeks to the capabilities of machine learning algorithms to accurately predict potential diseases based on reported symptoms, paving the way for timely interventions and improved healthcare outcomes.

I.I OBJECTIVE

Personalized Model: To develop a ML model that analyze the patient symptoms to identify potential health and risks to provide timely diagnosis. Comprehensive Information: In addition to predictions, the app offers comprehensive information about the predicted disease, including descriptions, precautions, medications, diet recommendations, and workout tips. Accurate Predictions: Leveraging machine learning, our model provides accurate disease predictions based on the symptoms provided by the user. Increase Accessibility: Provide a user-friendly platform for individuals to explore health concerns, particularly for those lacking immediate access to healthcare professionals.

I.II PROBLEM STATEMENT

The project aims to tackle the prevalent issue of diagnosing health conditions accurately from symptoms. By allowing users to input their symptoms and promptly receiving potential diagnoses, the system offers a convenient solution. This helps individuals dealing with overlapping symptoms to distinguish between different diseases efficiently. The platform streamlines the process of identifying health concerns, enhancing accessibility to accurate medical information.

II. SYSTEM METHODOLOGY

Data Collection and Preprocessing: Module Description: This module focuses on collecting comprehensive datasets that encompass a wide range of symptoms, diseases. The collected data needs to be pre-processed to ensure it is clean, normalized, and transformed into a suitable format for model training.

Implementation:

Utilizes public healthcare databases, and curated datasets to gather relevant healthcare information. Employ data preprocessing libraries in Python, such as Pandas and NumPy, to clean and preprocess the collected data. This involves handling missing values, normalizing the data, and transforming categorical data into numerical formats suitable for machine learning algorithms.

Model Development:

Module Description: Develop predictive models that can analyze the pre-processed data and provide disease predictions and health recommendations. The models must be able to accurately classify symptoms and predict potential diseases.

Implementation: Use machine learning libraries such as Scikit-learn to build and train the predictive models. This involves selecting appropriate algorithms, training the models on the pre-processed datasets, and validating their performance using techniques like cross- validation. Continuous training with new data ensures the models improve over time.

Query and Response Handling:

This component handles the communication between the user and the system. It takes the user's input query and sends it to the preprocessing unit and then receives the processed results to be displayed back to the user.

Preprocessing:

This stage involves preparing the data for analysis by the machine learning (ML) algorithms. It ensures that the data is clean, normalized, and in a format suitable for the ML models.

Components:

Label Encoding: Converts categorical data (e.g., symptom names) into a numerical format that can be used by ML algorithms.

Data Splitting: Divides the data into training and testing sets to enable model training and validation.

Machine Learning Model Training:

Random Forest : Random Forest is a powerful machine learning algorithm. By leveraging an ensemble of decision trees, Random Forest can effectively analyze diverse features extracted from symptoms, allowing for robust detection capabilities. In our research, we employ Random Forest as a key component of our proposed system, aiming to enhance the accuracy and efficiency of disease detection.

Support Vector Machine (SVM): SVMs are effective in high-dimensional spaces and are particularly useful for classification tasks involving complex decision boundaries. The SVM model is trained and evaluated to provide the prediction accuracy.

Database:

The database stores comprehensive healthcare information, including symptoms, diseases, medication, diet, workout and precautionary measures. It acts as the knowledge base for the system.

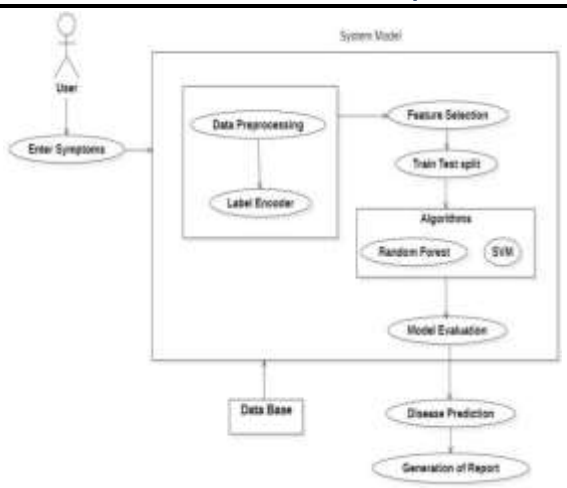

Fig 1: System Architecture

III. IMPLEMENTATION

III.I DATA BASE

The data base stores the training data for the machine learning model. This data is likely a large collection of examples consisting of patient symptoms and their corresponding diagnoses. The quality of the data heavily influences the accuracy of the model's predictions.

Fig 2: Data sets

III.II PRE-PROCESSING

This step involves cleaning the data, handling missing values, and formatting the data for use by machine learning algorithms.

```
dataNormalized1 = preprocessing.normalize(data, norm = 'l1')
dataNormalized2 = preprocessing.normalize(data, norm = 'l2')
print("L1 Normalized Data", dataNormalized1)
print("L2 Normalized Data", dataNormalized2)
```

```
L1 Normalized Data [[ 0.45132743 -0.25663717  0.2920354 ]
 [-0.0794702   0.51655629 -0.40397351]
 [ 0.609375    0.0625     -0.328125  ]
 [ 0.33640553 -0.4562212  -0.20737327]]
L2 Normalized Data [[ 0.75765788 -0.43082507  0.49024922]
 [-0.12030718  0.78199664 -0.61156148]
 [ 0.87690181  0.08993875  0.47217844]
 [ 0.55734935 -0.75585734 -0.34357152]]
```

Fig 3: Pre Proccessing

III.III DATA SPLITTING

The train-test split is a technique for evaluating the performance of a machine learning model. The data is split into two sets: a training set and a test set. The training set is used to train the machine learning model, and the test set is used to evaluate the performance of the model on unseen data. The size of the training and test sets is typically 80/20.

III.IV MODEL TRAINING

SUPPORT VECTOR MACHINE

```
In [22]: svc = SVC(kernel='linear')

         # Train the model
print("\n" + "="*40 + "\n")
svc.fit(X_train, y_train)
```

Fig 4: Support Vector Machine

A support vector machine (SVM) is a machine learning algorithm that can be used for classification tasks. SVMs work by finding a hyper plane that separates the data points of one class from the data points of another class.

RANDOM FOREST MODEL

```
RandomForest = RandomForestClassifier(n_estimators=100, random_state=42)
   # Train the model

RandomForest.fit(Xi_train, yi_train)
```

Fig 5: Training Data Set

Random forest trains multiple decision tree models on random subsets of the training data.

III.V EVALUATION METRICS

Once the models are trained, they are evaluated on the test data. The model evaluation involves comparing the predictions of the model to the actual values in the test set. There are a number of different metrics that can be used to evaluate the performance of a machine learning model.

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.99	0.96	67
1	0.99	0.98	0.98	82
2	0.89	0.92	0.90	85
3	1.00	0.99	0.99	69
4	0.94	0.99	0.96	75
5	0.96	1.00	0.98	68
6	1.00	0.96	0.98	69
7	0.99	0.95	0.97	77
8	0.98	1.00	0.99	58
9	0.97	0.97	0.97	69
10	0.98	1.00	0.99	65
11	0.99	0.98	0.98	82
12	1.00	0.97	0.99	74
13	1.00	1.00	1.00	79
14	1.00	0.96	0.98	73
15	0.92	1.00	0.96	82
16	0.94	0.97	0.95	75
17	1.00	0.93	0.96	68
18	0.99	1.00	0.99	69
19	1.00	1.00	1.00	83
20	1.00	1.00	1.00	66
21	1.00	1.00	1.00	78
22	1.00	0.99	0.99	83
23	0.99	0.96	0.98	84
24	1.00	1.00	1.00	68
25	1.00	0.98	0.99	82
26	0.98	0.98	0.98	66
27	0.93	0.99	0.96	69

Fig 6: Classification Report

```
========================================

RandomForest Accuracy: 0.9795379537953796
 RandomForest Confusion Matrix:
[[66,  0,  1, ...,  0,  0,  0],
 [ 0, 80,  0, ...,  2,  0,  0],
 [ 0,  0, 78, ...,  0,  0,  0],
 ...,
 [ 0,  0,  0, ..., 70,  0,  0],
 [ 1,  0,  0, ...,  0, 81,  0],
 [ 0,  0,  0, ...,  0,  0, 61]]

----------------------------------------
```

Fig 7: Random Forest Accuracy

```
====================================

SVC Accuracy: 0.969964087495919
 SVC Confusion Matrix:
[[73,  0,  0, ...,  1,  0,  0],
 [ 0, 73,  1, ...,  4,  0,  0],
 [ 0,  0, 82, ...,  0,  0,  0],
 ...,
 [ 0,  0,  0, ..., 66,  0,  0],
 [ 0,  0,  0, ...,  0, 83,  0],
 [ 0,  0,  0, ...,  0,  0, 63]]

------------------------------------
```

Fig 8: SVM Accuracy

IV RESULTS

This project leverages machine learning algorithms to predict diseases based on user- provided symptoms. Two primary models wereused. Below are the results and accuracy of themodels:

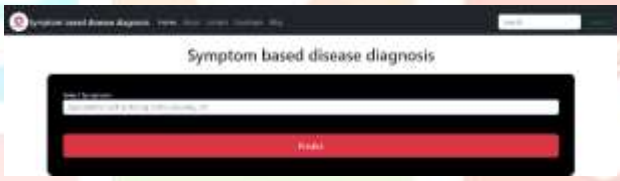

Fig 9: Home page of user interface

Fig 10: Result Disease Name

V CONCLUSION

In conclusion, symptom-based disease prediction using machine learning holds significant potential to improve early detection, enhance healthcare efficiency, and empower users with initial information for informed decisions. However, it is crucial to use this technology as a complementary tool alongside professional medical evaluation and treatment plans.

REFERENCES

1. Goyal, V. A., Parmar, D. J., Joshi, N. I., & Champanerkar, K. (2020). Medicine recommendation system. Medicine (Baltimore), 7(3).
2. Naveenkumar, S., Kirubhakaran, R., Jeeva, G., Shobana, M., & Sangeetha, K. (2021). Smart health prediction using machine learning. International Research Journal on Advanced Science Hub (IRJASH), 3(3), 124-128.
3. Yelne, R. A., & Raut, A. (2022, June). Health Care Digitalization and Machine Learning. In 13th International Conference on Advances in Computing, Control, and Telecommunication Technologies, ACT 2022.
4. Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. Current genomics, 22(4), 291.
5. Bao, Y., & Jiang, X. (2016, June). An intelligent medicine recommender system framework. In 2016 IEEE 11Th conference on industrial electronics and applications (ICIEA) (pp. 1383-1388). IEEE.
6. Chen, R. C., Chiu, J. Y., & Batj, C. T. (2011, July). The recommendation of medicines based on multiple criteria decision making and domain ontology—An example of anti-diabetic medicines. In 2011 International Conference on Machine Learning and Cybernetics (Vol. 1, pp. 27-32). IEEE.
7. Doulaverakis, C., Nikolaidis, G., Kleontas, A., & Kompatsiaris, I. (2012). GalenOWL: Ontology-based drug recommendations discovery. Journal of biomedical semantics, 3, 1-9.
8. Medvedeva, O., Knox, T., & Paul, J. (2007). Diatrack: web-based application for assisted decision-making in treatment of diabetes. Journal of Computing sciences in Colleges, 23(1), 154-161.
9. Dahiwade, D., Patle, G., & Meshram, E. (2019, March). Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1211-1215). IEEE.