Multiple Disease Prediction

P. Meena, Department of CSE, Narayana Engineering College, Gudur.
M.Subhashini, Assistant professor, Department of CSE, Narayana Engineering College, Gudur.

Abstract: *Predicting diseases through machine learning has gained significant attention due to its potential in enhancing healthcare outcomes. This project focuses on developing a disease prediction system capable of forecasting multiple conditions based on patient demographics, medical history, and symptom. Machine learning assist with foreseeing all kinds of diseases more effectively and help treat patients. Predictive analysis with the assistance of efficient multiple machine learning algorithms helps to predict the disease more correctly to treat patients in a better way. Using machine learning algorithms can lead to rapid disease prediction with high accuracy. machine learning algorithms are applied for classifying disease based on these clinical features. Finally, we compare the results obtained by different machine learning classification algorithms and visualize the result in a graph*

INDEX TERMS*: Precision Medicine, Clinical Decision Support(CDS),supervisedlearning, predictive modeling, machine learning,natural language processing.*

I.INTRODUCTION

In healthcare, machine learning is transforming how we predict diseases by using computers to analyze patient data. This project focuses on using these advanced techniques to predict multiple diseases accurately. By looking at patient information such as age, medical history, genetic data, and symptoms, we aim to create models that can forecast diseases before they become serious. The main goal is to develop these models so they can not only detect diseases early but also help doctors personalize treatments for each patient. This could lead to better outcomes for patients and more efficient use of healthcare resources. To achieve this, we'll start by gathering a diverse range of healthcare data and preparing it carefully for analysis. We'll then explore different machine learning methods to find the ones that work best for predicting diseases. Throughout this project, we'll prioritize patient privacy and follow strict ethical guidelines to ensure that data is handled responsibly and securely. Ultimately, our hope is that this research will empower healthcare providers with powerful tools to make better decisions and improve patient care. By harnessing the potential of machine learning, we aim to make healthcare more personalized, effective, and accessible for everyone. Throughout the project, we will evaluate various machine learning algorithms, including supervised learning classifiers and ensemble methods, to determine which models perform best in predicting diseases across different patient populations. Rigorous validation and evaluation metrics such as accuracy, precision, recall, and area under the curve will gauge the robustness and reliability of these models.

II .METHODOLOGY

In the proposed system, we present a GUI model, which is used by Naïve Bayes method, Decision Tree method, Support Vector Machine method and Random Forest method and MLP, the five diverse machine learning algorithms for disease prediction. The evaluating and comparative study between five different machine learning classification algorithms is made using performance metrics such as accuracy. The approach includes four steps. Firstly, important clinical features are selected. Secondly, an input of symptoms is taken from the user. Third, machine learning algorithms are applied for classifying disease based on these clinical features. Finally, we compare the results obtained by different machine learning classification algorithms and visualize the result in a graph. In this way with the expanding accessibility of electronic health data, increasingly hearty and progressed computational methodologies, for example, machine learning has gotten progressively functional to apply and investigate in disease prediction zone.

2.IMPLEMENTATION

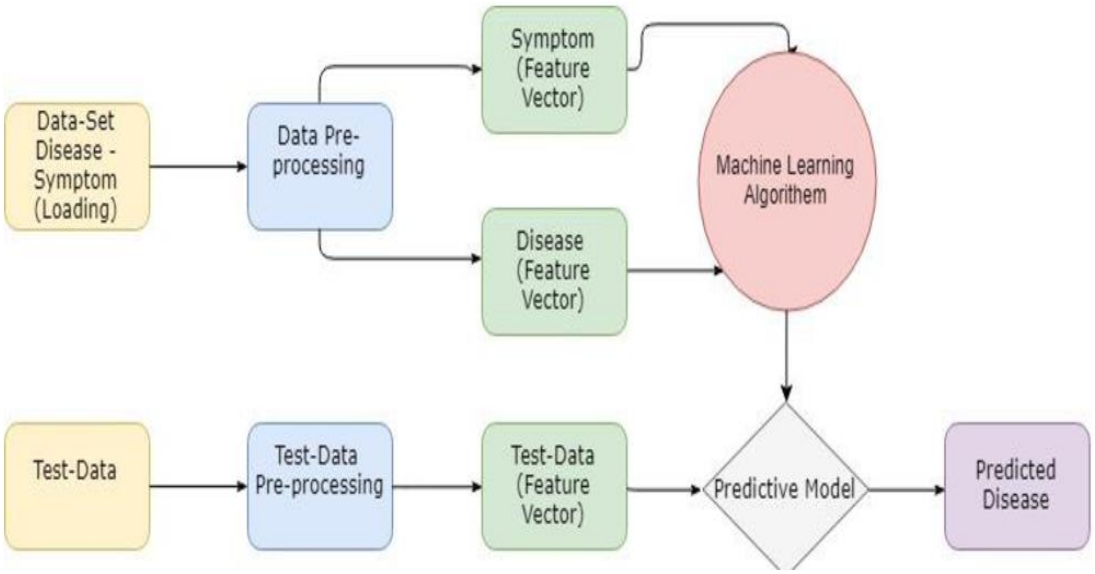
Gathering the datasets: We gather all the r data from the kaggle website and upload to the proposed model

Generate Train & Test Model: We have to preprocess the gathered data and then we have to split the data into two parts training data with 80% and test data with 20%

Run Algorithms: For prediction apply the machine learning models on the dataset by splitting the datasets in to 70 to 80 % of training with these models and 30 to 20 % of testing for predicting

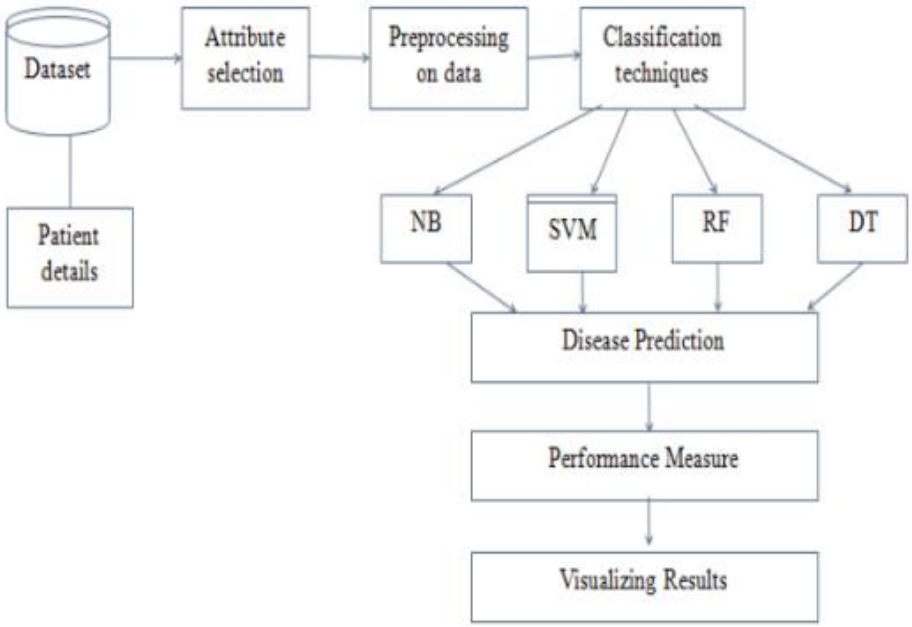
Input data: In this module we will give symptoms as our input based on we will get output

Predict output: in this module we will get output i.e disease name



The approach includes four steps.

- Firstly, important clinical features are selected.
- Secondly, an input of symptoms is taken from the user.
- Third, machines learning algorithms are applied for classifying disease based on these clinical features.
- Finally, we compare the results obtained by different machine learning classification algorithms and visualize the result in a graph.
- In this way with the expanding accessibility of electronic health data, increasingly hearty and progressed computational methodologies,
- for example, machine learning has gotten progressively functional to apply and investigate in disease prediction zone.



III.FUNCTIONAL REVIEW

The term functional review defines the different type of data collection and processing also its easy to train and test the data there are:

1.Data Collection

2.Data Preprocessing

3.Training And Testing

4.Modiling

5.Predicting

DATA COLLECTION:

Data collection for multiple disease prediction using machine learning involves gathering diverse and comprehensive datasets that encompass various aspects of patient health and medical history.

Patient Demographics: Information such as age, gender, ethnicity, and geographical location.

Medical History: Previous diagnoses, treatments, surgeries, and hospitalizations. Symptoms and Clinical Signs: Presenting symptoms recorded during visits or examinations.

Laboratory Results: Blood tests, genetic tests, biomarkers, and other diagnostic results.

DATA PREPROCESSING:

In most cases, an imbalanced dataset signifies that there are fewer examples of a minority class in the dataset for a machine-learning algorithm to learn the decision boundary. Data preprocessing serves as the foundation for developing and refining machine learning models for disease prediction. It ensures that the data used for training and testing is clean, standardized, and appropriately structured to maximize the accuracy and reliability of predictions. By carefully implementing each step of data preprocessing, the project can effectively leverage machine learning to enhance diagnostic accuracy, support personalized treatment plans, and ultimately improve patient outcomes in healthcare settings.

TRAINING AND TESTING:

The training phase involves using labeled data (where the outcome or diagnosis is known) to train the machine learning model to recognize patterns and relationships between input features (predictors) and the target variable (disease presence or outcome).

Steps Involved: Data Splitting: The dataset is typically divided into two subsets:

Training Set: This subset (often around 70-80% of the data) is used to train the model. The model learns from this data by adjusting its parameters through iterative optimization processes (e.g., gradient descent in neural networks, tree splitting in decision trees).

Validation Set: In some cases, a validation set (usually around 10-20% of the data) is used to tune hyperparameters (like learning rate, regularization strength) and assess model performance during training.

Once the model is trained on the training data, it needs to be evaluated on unseen data to assess its generalization ability—how well it performs on new, previously unseen data. Steps Involved: Testing Set: The remaining portion of the dataset (not used in training) serves as the testing set. It is crucial that the model does not see this data during training to avoid bias in performance evaluation. Evaluation Metrics:

Accuracy: Proportion of correctly predicted instances among all instances.

Precision: Proportion of true positive predictions among all positive predictions.

Recall (Sensitivity): Proportion of true positives correctly identified among all actual positives.

F1-score: Harmonic mean of precision and recall, balancing between the two metrics.

Modeling:

In the context of disease prediction using machine learning refers to the process of selecting and training algorithms that can effectively learn patterns from data to make accurate predictions about the presence, progression, or risk of diseases. Here's an overview of the key aspects involved in modeling.

PREDICITION:

Prediction in the context of machine learning refers to using trained model to make predictions or decision on new unseen data.

Input data: providing the preprocessed input data to the trained model data.The input data should be in the same as the model was trained on.

Output prediction: The format and interpretation depend on the specific problem and type of model used.

IV.RESULT AND DISCUSSION

Training & Splitting

```
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.33, random_state=42)
```

Model selection

```
from sklearn.model_selection import cross_val_score
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt
```

Fig 3:Training and splitting the data

```
from sklearn.svm import SVC
SVM = SVC(kernel='linear')
SVM.fit(X_train, y_train)
predictions = SVM.predict(X_test)
val1 = (accuracy_score(y_test, predictions)*100)
print("*Accuracy score for SVM: ", val1, "\n")
print("*Confusion Matrix for SVM: ")
print(confusion_matrix(y_test, predictions))
print("*Classification Report for SVM: ")
print(classification_report(y_test, predictions))

from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, predictions)
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, predictions)
print(cm)
```

Random Forest

```
from sklearn.ensemble import RandomForestClassifier
RF = RandomForestClassifier()
RF.fit(X_train, y_train)
predictions = RF.predict(X_test)
val3 = (accuracy_score(y_test, predictions)*100)
print("*Accuracy score for RF: ", val3, "\n")
print("*Confusion Matrix for RF: ")
print(confusion_matrix(y_test, predictions))
print("*Classification Report for RF: ")
print(classification_report(y_test, predictions))
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, predictions)
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, predictions)
print(cm)

plt.matshow(cm)
plt.title('Confusion matrix of the classifier\n')
plt.xlabel('Predicted')
plt.ylabel('True')
plt.colorbar()
plt.show()
```

```
*Accuracy score for RF:  100.0
```

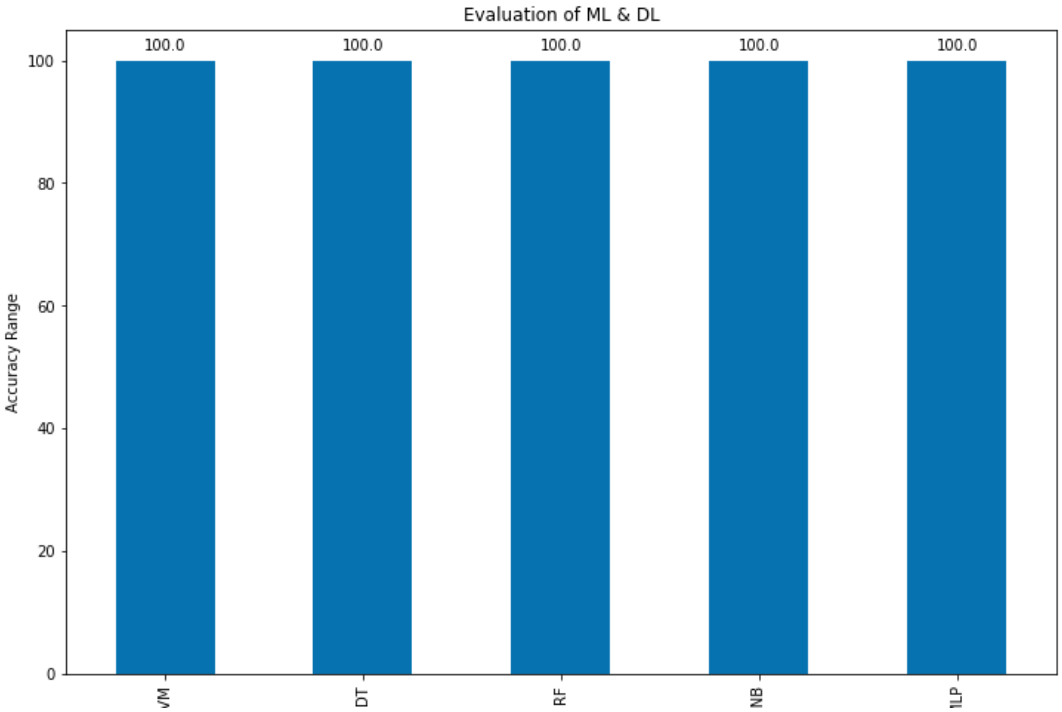
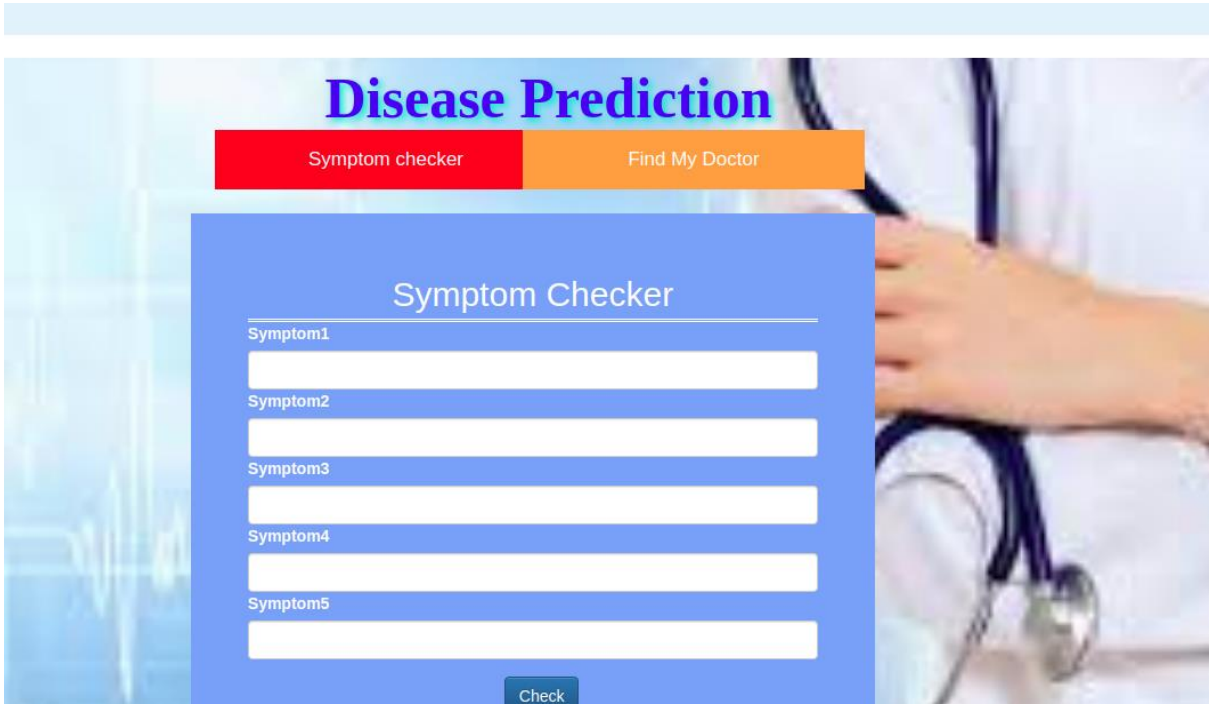


Fig 3:From the above the all are giving better accuracy for prediction



V.CONCLUSION

The fundamental aim of the paper is to anticipate all the more precisely the event of the ailment utilizing Machine Learning procedures. We used four algorithms like DecisionTree, Random Forest, SVM and Naïve Bayes. These algorithms are used to evaluate the parameters like Accuracy, Precision, Recall and F-score on a disease dataset by considering the symptoms as input variables. Among all the algorithms on the comparison, the highest accuracy for predicting the disease is given by the Naïve Bayes algorithm. Further enhancements can be added to this system because the new disease may arise in future. These new diseases can be predicted by adding new symptoms to the data set. By using more powerful machine learning supervised algorithms, the GUI is modified user friendly to provide more detail information about the disease to the patients and can be further extended by an additional component suggesting medicine to the patient if necessary on emergency.

VI. REFERENCES

[1] Thailand Motor Vehicle Registered. CEIC Data Global Database.

[2] Linda, S., Can public transport compete with the private car? Iatss Research, 2003.27(2): p. 27-35.

[3] CO2 emissions (metric tons per capita) - Thailand. THE WORLD BANK.

[4] Cohen, A.J., et al., Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet, 2017. 389(10082): p. 1907-1918.

[5] Litman, T. and D. Burwell, Issues in sustainable transportation. International Journal of Global Environmental Issues, 2006. 6(4): p. 331-347.

[6] Liu, M. and N. Choosri.. A technical solution to improve the red cab for touring in Chiang Mai: Chinese tourists' perspective. in 2016 Chinese Control and Decision Conference (CCDC). 2016. IEEE

[7] Farooq, M.U., A. Shakoor, and A.B. Siddique. GPS based Public Transport Arrival Time Prediction. in 2017 International Conference on Frontiers of Information Technology (FIT). 2017. IEEE.

[8] Bin, Y., Y. Zhongzhen, and Y.Baozhen, Bus arrival time prediction using support vector machines. Journal of Intelligent Transportation Systems, 2006. 10(4): p. 151-158.

[9] Maiti, S., et al. Historical data based real time prediction of vehicle arrival time. in 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). 2014. IEEE.

[10] Fan, W. and Z. Gurmu, Dynamic travel time prediction models for buses using only GPS data. International Journal of Transportation Science and Technology, 2015. 4(4): p. 353-366.