Disease Prediction

PAVAN SHIVASHANKAR
20023029
Pavanshivashankar11@gmail.com

JUNCHENG ZHOU
20022811
izhou59@stevens.edu

***Abstract*—Predire is a web-based application that will help to predict diseases and provide an educational environment for students and patients. This project will provide an effective platform for the medical students and patients. There's been an increase in disease prediction models but there have never been models which predicts more than one variety of disease in a wider spectrum. This project focuses and tackles this research problem by using classification algorithm of data science on meticulously curated medical dataset.**

Different algorithms like Decision Tree, Random Forest Classifier is used for the classification of the dataset, where the users predict their disease based on 5 input symptoms. With the power of the powerful machine learning techniques our website can predict disease with great accuracy. The website also provides news and research paper related to medical field for students and doctors.

For this project we are going to use various classification algorithm to the transformed dataset with visualization

I. Introduction

Predire is a website that uses Machine Learning algorithms to predict diseases with the help of symptoms provided by our users/patients. It also provides Healthcare related news and research papers for doctors and medical students. This website caters not only to the public but also to hospitals and private corporations. With the integration of powerful machine learning techniques and accurate medical data, we have developed a website that can predict 41 diseases. It predicts diseases that range from common diseases such as common cold, fever, etc to severe diseases such as hepatitis, tuberculosis, diabetes, etc The main objective of making this project (Predire) is to create an environment where the user can browse through various Health-related news, site curated research papers and get an accurate synopsis of the disease they might have while using our Prediction page. As this is a Web Application, it is easily accessible from anywhere in this world as long as there is a stable internet and hardware, and software requirements are met. The User Interface of the website is meticulously designed for a user-friendly experience wherein the user can easily navigate through the website using our well-designed Home page and Navigation bar. The main highlight of the website is our Prediction model.

The healthcare industry has experienced a revolution in diagnostic procedures due to the incorporation of modern technologies. This has allowed for new insights into illness patterns and prognosis. This research explores the complex field of disease prediction using machine learning (ML) with the goal of improving patient outcomes and early detection. Traditional diagnostic methods are being supplemented by machine learning (ML) algorithms that can interpret complicated patterns, correlations, and hidden linkages within datasets as the volume and complexity of medical data continue to rise. Accurate illness prediction not only changes the preventive healthcare picture, but it also opens the door for individualized and focused interventions. The goal of this research is to use machine learning (ML) methods, like Gaussian Naive Bayes, Random Forest, Decision Trees, and Support Vector Machines (SVM), to predict and assess the occurrence of diseases based on a variety of symptoms. The potential for a paradigm shift in healthcare delivery is presented by the convergence of medical expertise and state-of-the-art technology in disease prediction. This could usher in an era where proactive measures can reduce health risks, enhance patient outcomes, and promote overall well-being. The met conclusion of this project is to predict user's diseases accurate, however the research is labelled to be in progress as with more medical data, more accurate, descriptive and intricate prediction models can me made.

II. Related Work

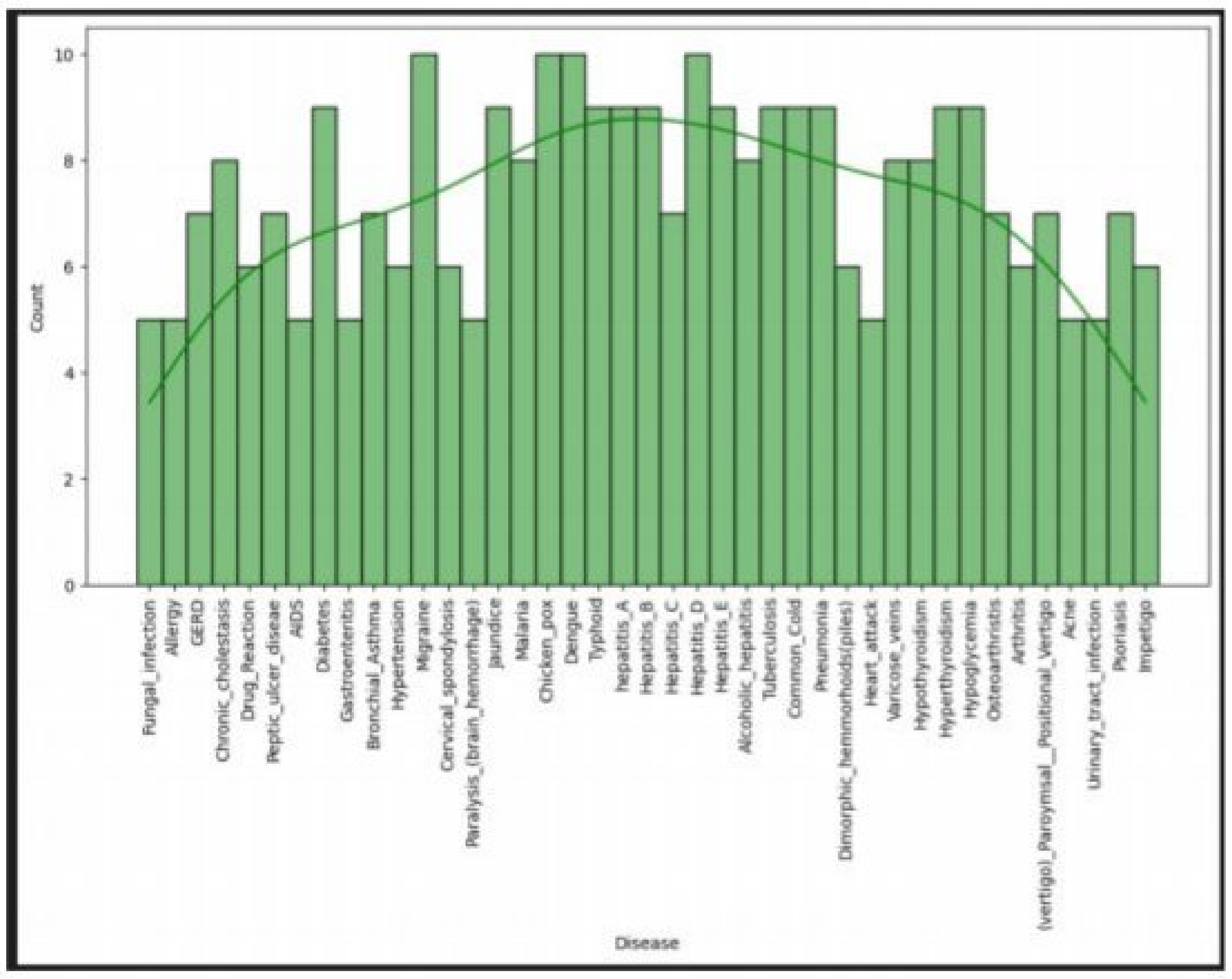
Healthcare Professionals: This dataset can be used for clinical analysis, research studies, and epidemiological investigations pertaining to various diseases by doctors, medical practitioners, and researchers. It might help to comprehend the frequency and symptom patterns among people with particular medical disorders. Medical Researchers: The dataset can be utilized by researchers who are interested in certain diseases or conditions to investigate correlations between symptoms, age, gender, and other factors. The development of fresh perspectives, therapeutic approaches, and preventative measures can benefit from this data. Healthcare Technology Companies:

This dataset can be used by companies creating AI algorithms, diagnostic tools, or healthcare applications to train and validate their models. Based on patient features and symptoms, the data can help construct prediction models for illness diagnosis or monitoring.

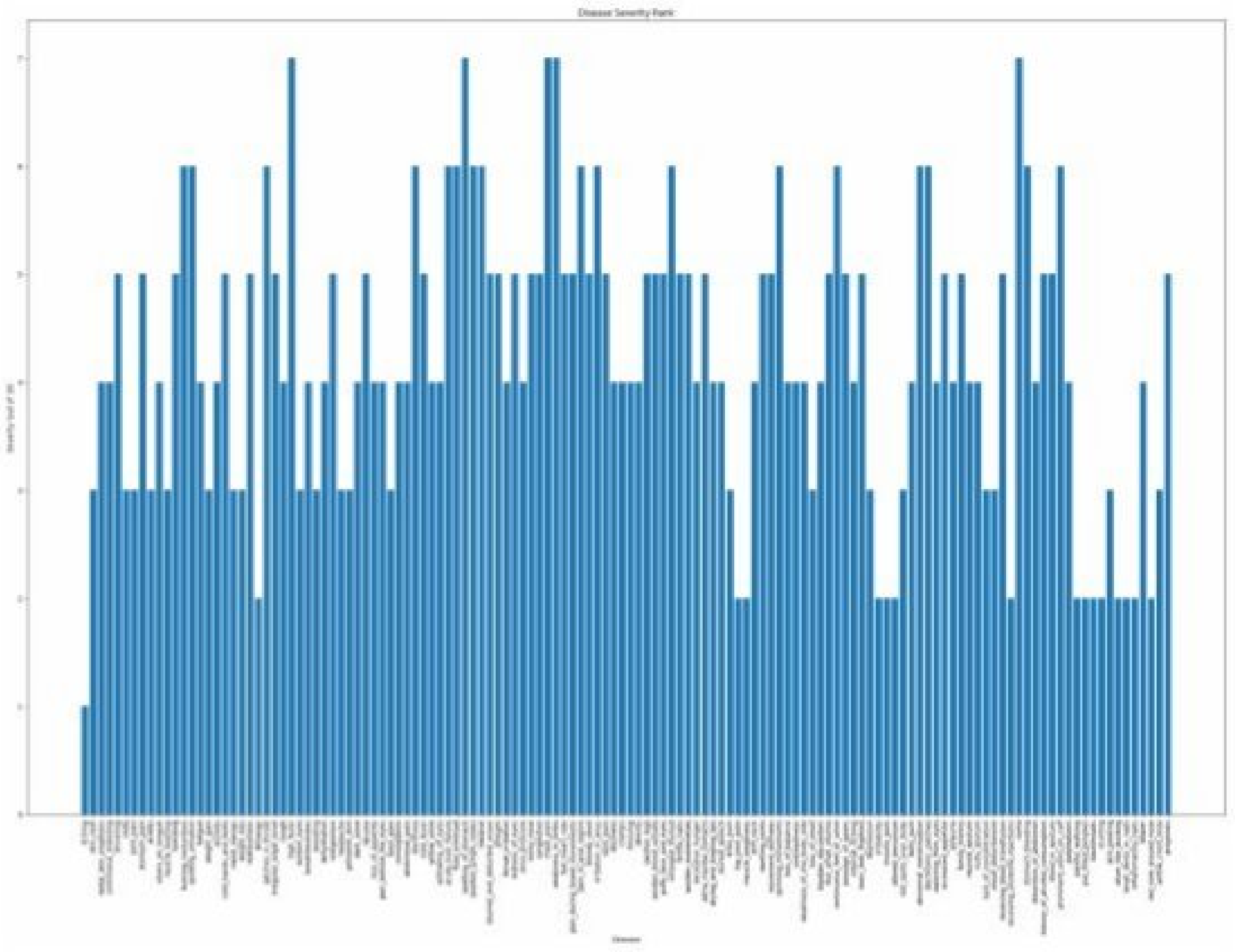
III. Our Solution

A. Description of Dataset

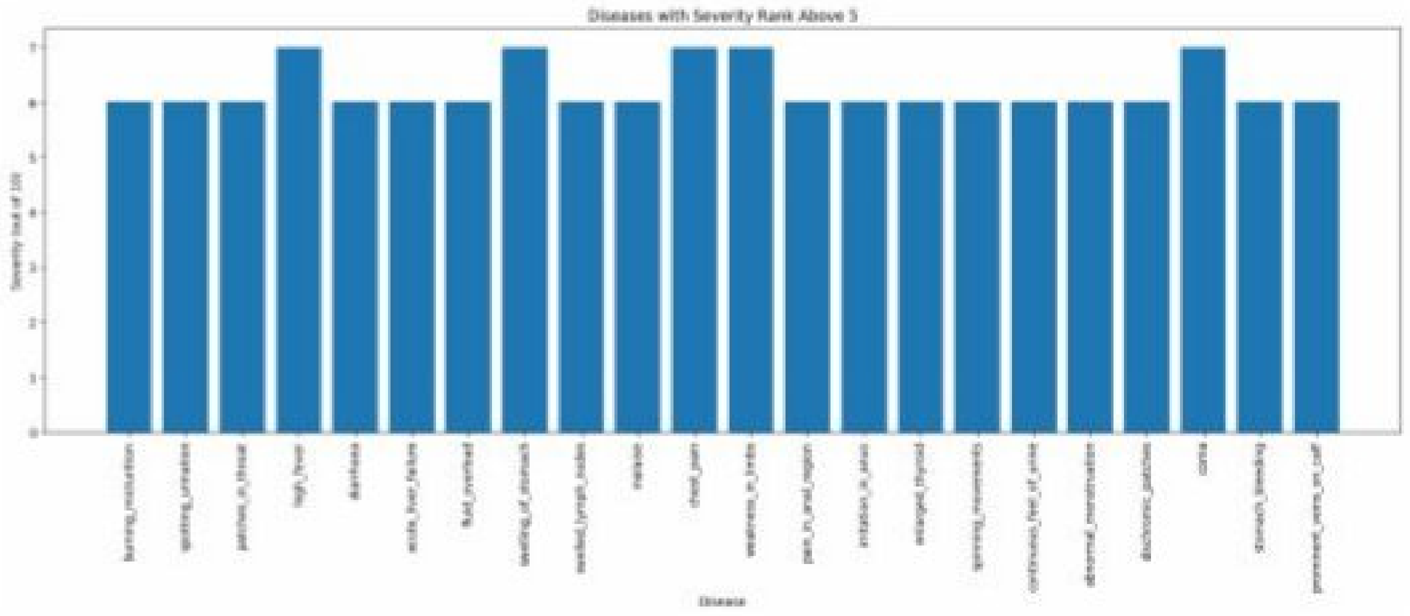
This dataset comprises details about different diseases and their corresponding symptoms. Each entry in the dataset represents a particular disease, and the columns encompass the disease name alongside binary markers for symptoms. Symptoms are denoted as binary values, where 1 signifies the existence of a symptom, and 0 denotes its absence. The dataset serves as a fundamental resource for discerning patterns and connections between diseases and symptoms. The existing systems do not provide prediction of multiple diseases. They are focussed on a single disease such as Diabetes, Tuberculosis, Heart disease etc. Since they are solely focused on prediction, they do not have a lot of versatality. Even though they specialise in a single disease, no descriptive analysis is provided for their users. Also, they do not provide an environment where the user can learn or gain knowledge about topics related to their application. All the existing systems are web/mobile applications where no importance is given to the user interface of their application. To resolve the problems occurring in the existing system, we came up with a solution to create a web application that handles all these problems and provides a better way to work. Our application will be easier for the doctors, patients and students to use as it will be user friendly. We have put in a lot of effort in designing a modern, elegant yet efficient website. We have adopted a different prediction approach from the existing systems to provide fast and accurate results. We provide a subtle description for the disease with a list of precautions to take.



The plot above shows us that which of the diseases have the highest number of symptoms. After this step we can dwell ourselves into the other dataset in our project which tells us about the severity of the diseases at hand. We ranked the symptoms from 0-10 based on their severity level. The graph below shows us the Symptoms with the highest of severity (for our case we took the base rank as 5) to gain knowledge of the symptoms which require the most rapid action.

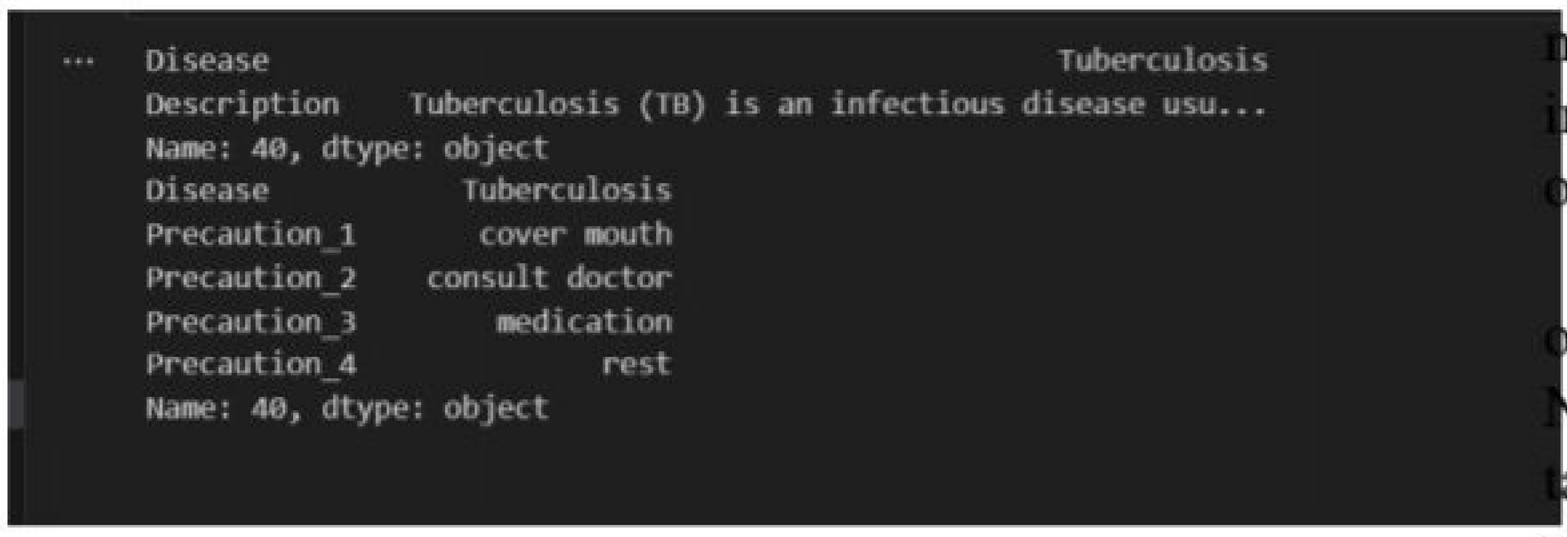


The Symptoms and Severity Levels dataset provides comprehensive information about various symptoms associated with their corresponding severity levels. This provides us with a quantitative measure of the intensity or impact of the symptom in a given case. The dataset enables the identification of patterns, trends, and correlations that can contribute to a more nuanced understanding of symptomatology in various health conditions. The graph below shows us the Symptoms with the highest of severity (for our case we took the base rank as 5) to gain knowledge of the symptoms which require the most rapid action.



The Disease Descriptions dataset serves as a complement to the Diseases and Symptoms dataset, providing in-depth textual descriptions for each disease. Each entry corresponds to a specific disease, featuring a column dedicated to detailed descriptions that illuminate the characteristics, causes, and effects of the respective diseases. These descriptions play a crucial role in enhancing the understanding of disease nature and contribute to improving the interpretability of machine learning models.

On the other hand, the Precautions dataset contains precautionary measures and recommendations linked to various diseases. Each row represents a specific disease, with columns outlining precautionary advice and recommended actions for each condition. These precautions encompass a range of strategies, including lifestyle adjustments, medical interventions, and other preventive measures. The dataset is invaluable for delivering actionable information to users or patients, guiding them on how to minimize risks associated with specific diseases.



```
...   Disease                                          Tuberculosis
      Description    Tuberculosis (TB) is an infectious disease usu...
      Name: 40, dtype: object
      Disease             Tuberculosis
      Precaution_1         cover mouth
      Precaution_2      consult doctor
      Precaution_3          medication
      Precaution_4                rest
      Name: 40, dtype: object
```

B. Machine Learning Algorithms

For the dataset, we have decided to use random forest. Random Forest is often considered a powerful and effective algorithm for predicting diseases due to several characteristics that make it well-suited for this task. Diseases often exhibit complex and non-linear relationships with multiple variables. Random Forest is capable of capturing complex patterns and non-linear relationships in the data, making it suitable for modeling the intricate nature of diseases.

Random Forest provides a feature importance score for each variable, indicating the contribution of each feature to the predictive accuracy. This is crucial in the medical field, as it helps identify the most relevant factors contributing to a disease, aiding medical professionals in understanding the underlying mechanisms. Medical datasets can often contain outliers or noisy data. Random Forest is robust to outliers, as it builds multiple trees and aggregates their predictions, reducing the impact of outliers on the overall model Medical datasets may have missing values for certain features. Random Forest can handle missing data well by imputing missing values based on available information and still providing accurate predictions. Other algorithms which we used to train our model; Support Vector Machines (SVM): Because SVM can handle high-dimensional data and nonlinear correlations, it is useful in the prediction of diseases. SVM is well-suited for situations where distinguishing subtle differences between various health conditions is necessary because it is an excellent tool for recognizing complicated patterns inside datasets. SVM is a useful tool in illness prediction models because of its resilience in processing a variety of data types, such as genetic, imaging, and clinical data.

2. Decision Trees: Decision trees are useful in disease prediction scenarios where model transparency is crucial since they are easily interpretable and comprehensible. Decision trees are flexible for medical datasets with a variety of information kinds since they can handle both continuous and categorical variables. Furthermore, Decision Trees are a natural way to capture the significance of features, which helps physicians comprehend the variables that influence disease prognoses.

3. AdaBoost: AdaBoost is an ensemble learning technique that builds a strong prediction model by combining several weak learners. By concentrating on cases that earlier models misclassified, AdaBoost performs exceptionally well in enhancing classification accuracy in disease prediction. By reducing biases, this adaptive boosting method improves the model's overall performance. AdaBoost works very well with imbalanced datasets, which are a common problem in the field of medical data analysis.

4. Gaussian Naive Bayes: When working with continuous information, naive Bayes models—in particular, Gaussian Naive Bayes—are excellent choices for illness prediction tasks. Even with a small amount of training data, they function effectively and are computationally effective. Since Gaussian Naive Bayes assumes feature independence, it makes modeling easier and works well for medical datasets with a lot of variables. Because of its probabilistic structure, predictions may be easily interpreted.

C. Implementation Details

STEP 1:-Define the Problem: The initial step is to clearly articulate the problem at hand. This involves specifying the disease to be predicted and defining the target variable. Understanding the objective of the prediction task helps in determining the features that will contribute to the model's learning process. We first import all the libraries necessary for the Implementation and visualization. Then we Load data into our model.

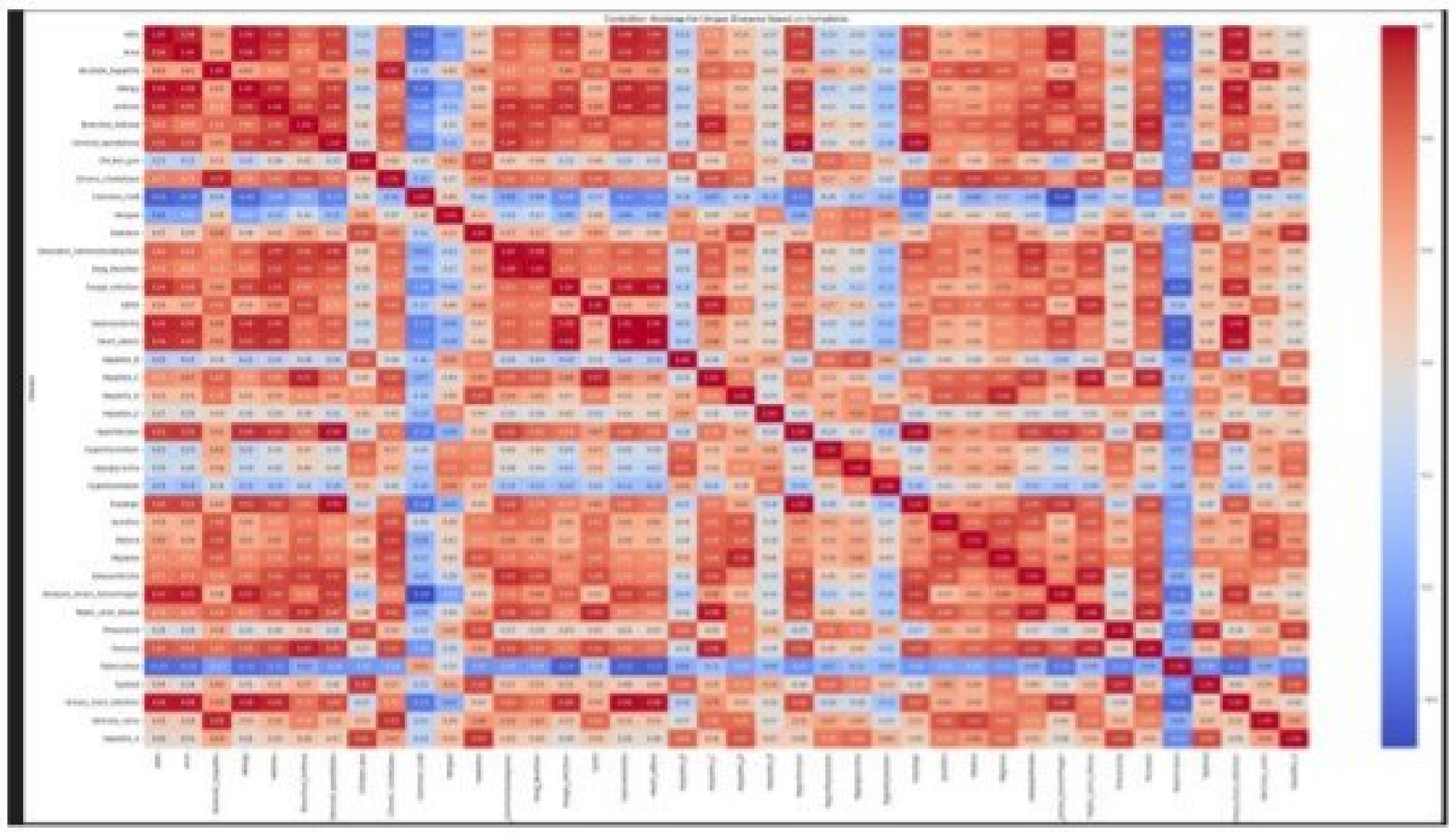
Collect Data: Data collection is fundamental to any machine learning project. It is imperative to ensure the quality of the data, addressing issues such as cleanliness and completeness, as the model's performance heavily relies on the quality of the input data. Through the dataset, we understand the severity as well as the symptoms which has the highest influence in certain diseases.

Exploratory Data Analysis EDA: Once the data is collected, exploratory data analysis EDA is conducted. EDA involves statistical and visual methods to explore the characteristics of the dataset. Visualization techniques, such as histograms and scatter plots, aid in understanding the distribution of features and identifying potential patterns or trends within the data.

A useful technique for displaying disease relationships is a heatmap. Through analysis of the heatmap's hue and intensity, you can determine:

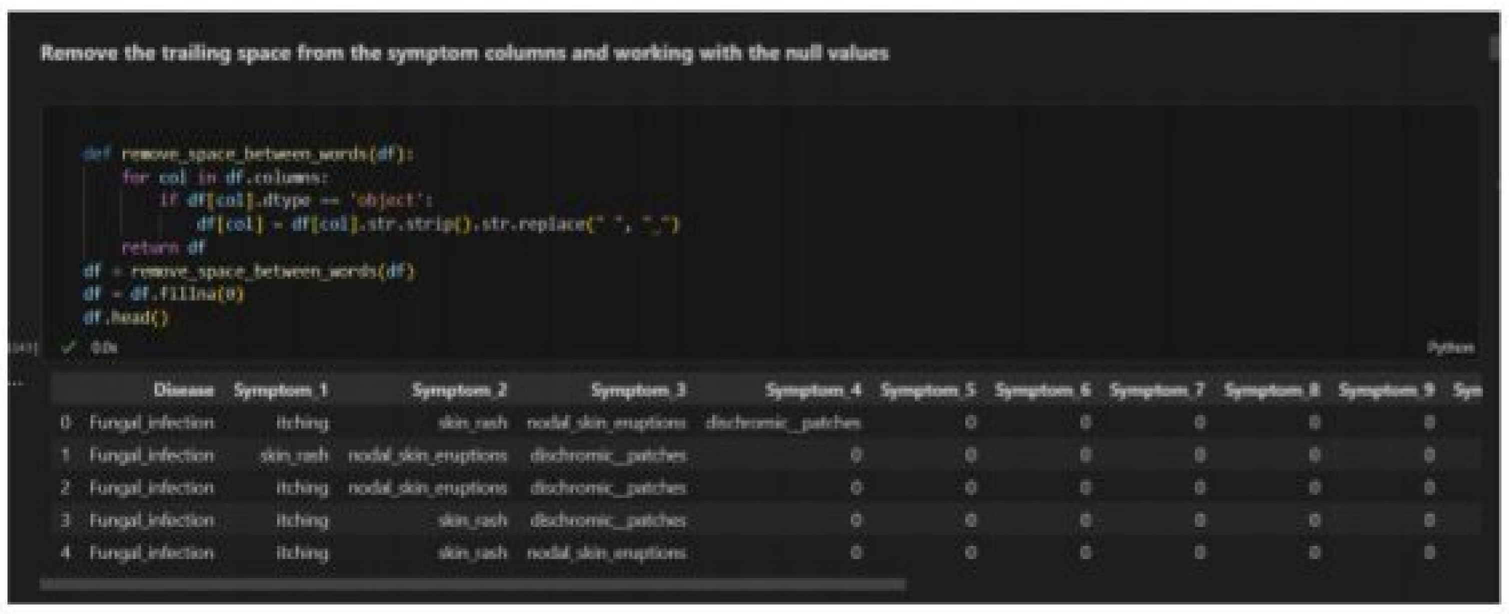
Positive Correlation: The likelihood of diseases developing simultaneously is increased when there is a positive correlation between them. A strong positive link between such diseases is indicated by bright or vivid colors in the heatmap. This information is useful for comprehending co-occurrence patterns and could have consequences for future research, treatment planning, or diagnosis.

Negative association: On the other hand, illnesses that have a negative association are not as likely to coexist. In the heatmap, darker or colder hues indicate a negative correlation. Recognizing illnesses that may exclude one another or present with contradictory symptoms requires an understanding of negative correlations. From the graph below we can confirm the correalations between the diseases provided in the dataset.



The heatmap's color intensity corresponds to the correlation's strength. Lighter hues imply a weaker association, while brighter hues indicate a stronger correlation. You can learn more about the connections between diseases and spot possible patterns or clusters by examining the heatmap. This data can direct further investigation, study, or improvement of disease diagnostic prediction algorithms. Please offer more information if you would want to talk about any patterns you've noticed or if you have any specific queries concerning the heatmap!

Data Preprocessing: Data preprocessing is a critical step to ensure that the data is in a suitable format for machine learning algorithms. This involves handling missing data through imputation or removal, encoding categorical variables into numerical representations, and scaling or normalizing numerical features to bring them to a consistent scale.



Remove the trailing space from the symptom columns and working with the null values

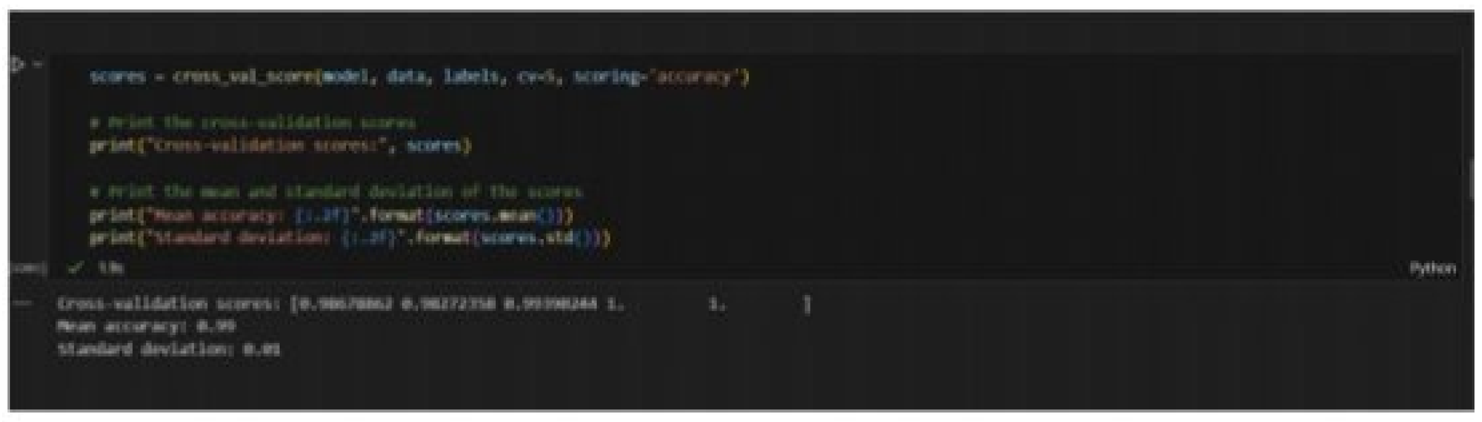
```python
def remove_space_between_words(df):
    for col in df.columns:
        if df[col].dtype == 'object':
            df[col] = df[col].str.strip().str.replace(" ", "_")
    return df
df = remove_space_between_words(df)
df = df.fillna(0)
df.head()
```

	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7	Symptom_8	Symptom_9	Sym
0	Fungal_infection	itching	skin_rash	nodal_skin_eruptions	dischromic__patches	0	0	0	0	0	
1	Fungal_infection	skin_rash	nodal_skin_eruptions	dischromic__patches	0	0	0	0	0	0	
2	Fungal_infection	itching	nodal_skin_eruptions	dischromic__patches	0	0	0	0	0	0	
3	Fungal_infection	itching	skin_rash	dischromic__patches	0	0	0	0	0	0	
4	Fungal_infection	itching	skin_rash	nodal_skin_eruptions	0	0	0	0	0	0	

Here we remove the trailing space from the symptom columns and worked with the null values. We also found that the data had duplicate values which were processed in this step.

Feature Selection: Identifying and selecting the most relevant features is essential for model performance. In the context of Random Forest, which inherently handles feature importance, techniques such as examining feature importance scores generated during training can be employed. This step helps streamline the model by focusing on the most impactful variables.

Model Training: The dataset is then split into training and testing sets which in our model we split it into 80:20 split to facilitate model training and evaluation The model is trained using the training dataset. With the help of this training dataset we are able to apply various powerful machine learning algorithms such as Random forest, Naive Bayes, SVM, AdaBoost and Decision Tree. WE then Adjust hyperparameters to optimize model performance. Use techniques like cross-validation to ensure the model generalizes well to unseen data. During this phase, the Random Forest algorithm learns the underlying patterns and relationships within the data, enabling it to make predictions based on new, unseen data.after the analysis we encode the data parameter to let it fit in the machine learning algorithm.



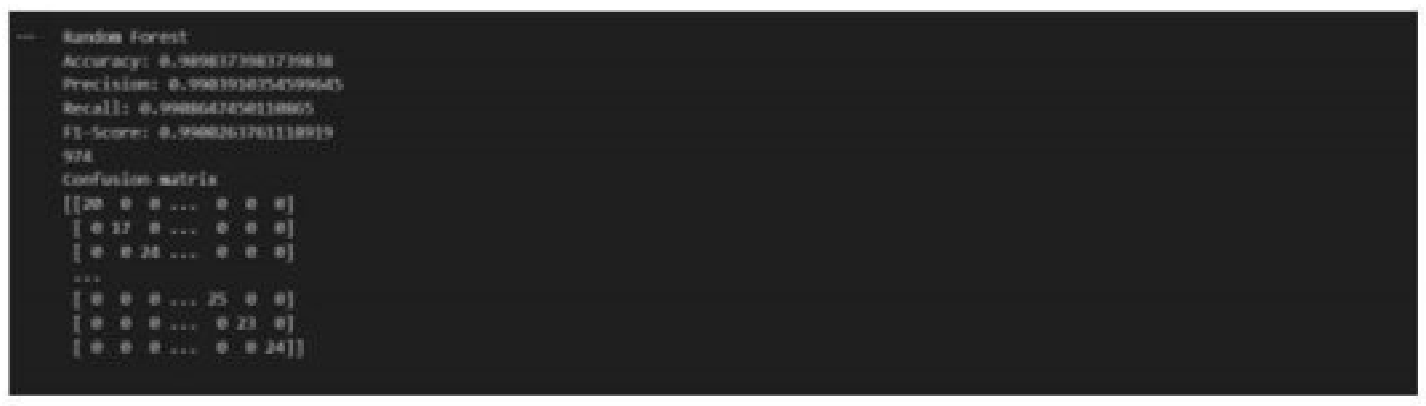
```python
scores = cross_val_score(model, data, labels, cv=5, scoring='accuracy')

# Print the cross-validation scores
print("Cross-validation scores:", scores)

# Print the mean and standard deviation of the scores
print("Mean accuracy: {:.2f}".format(scores.mean()))
print("Standard deviation: {:.2f}".format(scores.std()))
```

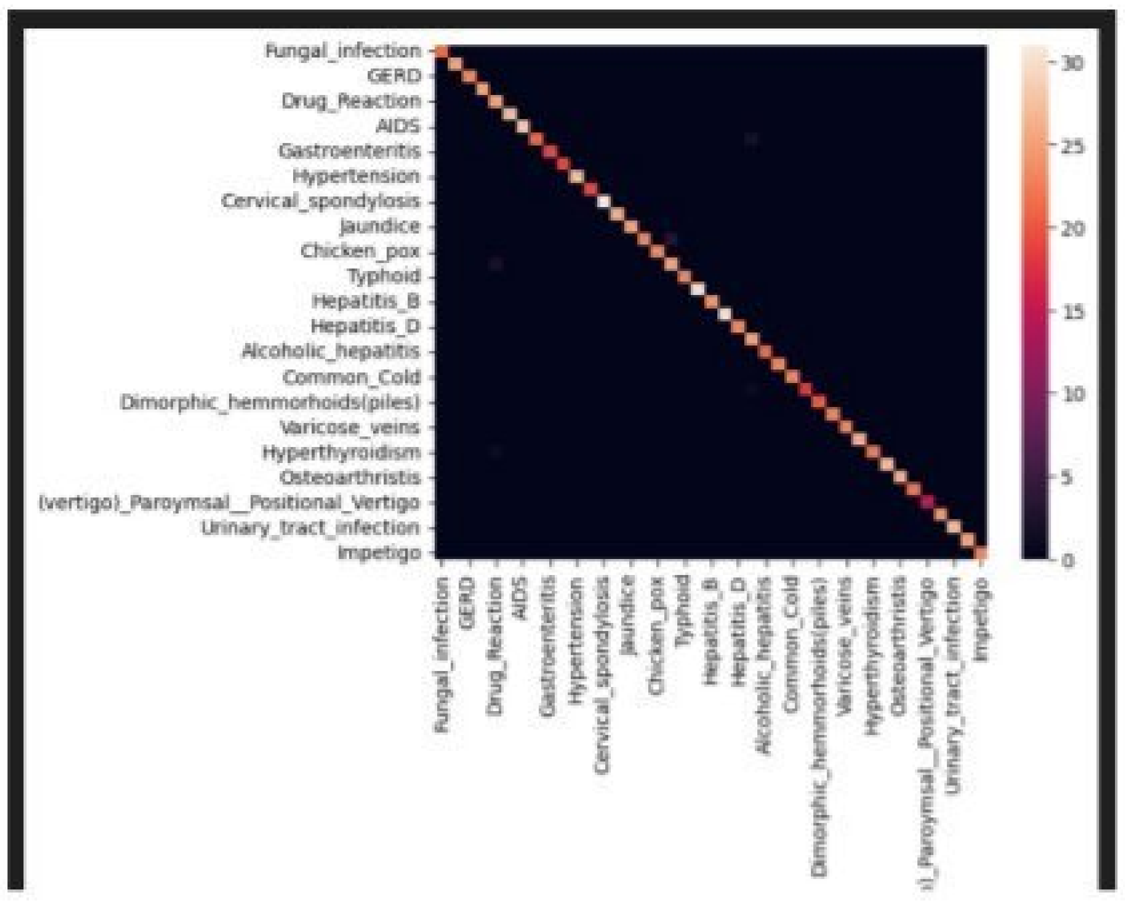
```
Cross-validation scores: [0.98678862 0.98272358 0.99390244 1.         1.        ]
Mean accuracy: 0.99
Standard deviation: 0.01
```

Model Evaluation: To assess the model's performance, it is evaluated using the testing dataset. Various metrics, such as accuracy, precision, recall, and F1-score, are calculated to gauge how well the model generalizes to new, unseen data. This step is crucial for understanding the model's strengths and potential areas for improvement.



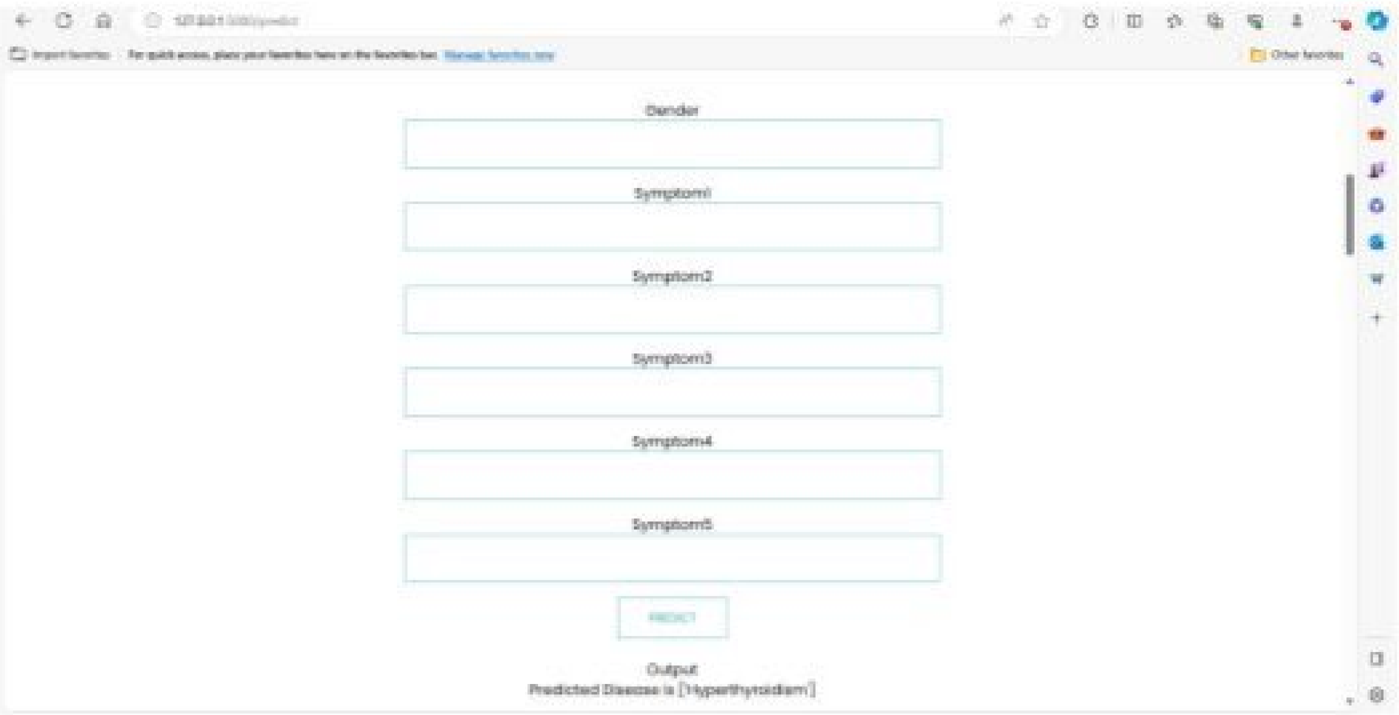
```
Random Forest
Accuracy: 0.9898373983739838
Precision: 0.9903910354599645
Recall: 0.9908647450110865
F1-Score: 0.9900263761110919
974
Confusion matrix
[[20  0  0 ...  0  0  0]
 [ 0 17  0 ...  0  0  0]
 [ 0  0 24 ...  0  0  0]
 ...
 [ 0  0  0 ... 25  0  0]
 [ 0  0  0 ...  0 23  0]
 [ 0  0  0 ...  0  0 24]]
```

As we can see from the image above that the random forest algorithm provides us with a 99 percent of accuracy for the prediction which makes it the best model available to give prediction for our website. Here we also plotted a heatmap for the same. The concentration of the colors of the diagonals in random forest heatmap tells us that the model is making correct predictions and which is the same case for the other models which we have used which shows us the area where the areas where the model is failing to make a correct prediction



The main goal of our project (Predire) is to develop a space where users can peruse a variety of health- related news, research papers selected by the site, and receive an accurate overview of any diseases they may be at risk for utilising our Prediction page. Given that it is a web application, it can be easily accessed from anywhere in the world as long as the necessary hardware, software, and internet conditions are satisfied. The website's User Interface was carefully created for a user-friendly experience, allowing visitors to simply navigate through it utilising the Home page and Navigation bar. Our Prediction model is the website's primary selling point. The model, which has a predict button and 5 symptom inputs, is simple to grasp. For the convenience of the user, we have also included thorough instructions on how to operate the model. There is also a Blog/News section where we offer news feeds about medical topics from around the globe. By delivering the most recent research papers in the field of medicine, this project is also attempting to establish a setting that caters to medical students.

To make the predictive power of our machine learning models accessible to users, we developed a web application. Using the Flask framework, we seamlessly integrated the trained models into the backend, allowing users to input their medical information through an intuitive user interface. The web application facilitates real-time predictions and serves as a user-friendly platform for disease risk assessment. The website which we created for the users to input their symptoms and get their predictions of the disease with utmost precision and at the leisure of staying at home just at the tip of their hands. The users get a fast and accurate model to predict disease just by adding the symptoms which is easy to use and user friendly.



Gender

Symptom1

Symptom2

Symptom3

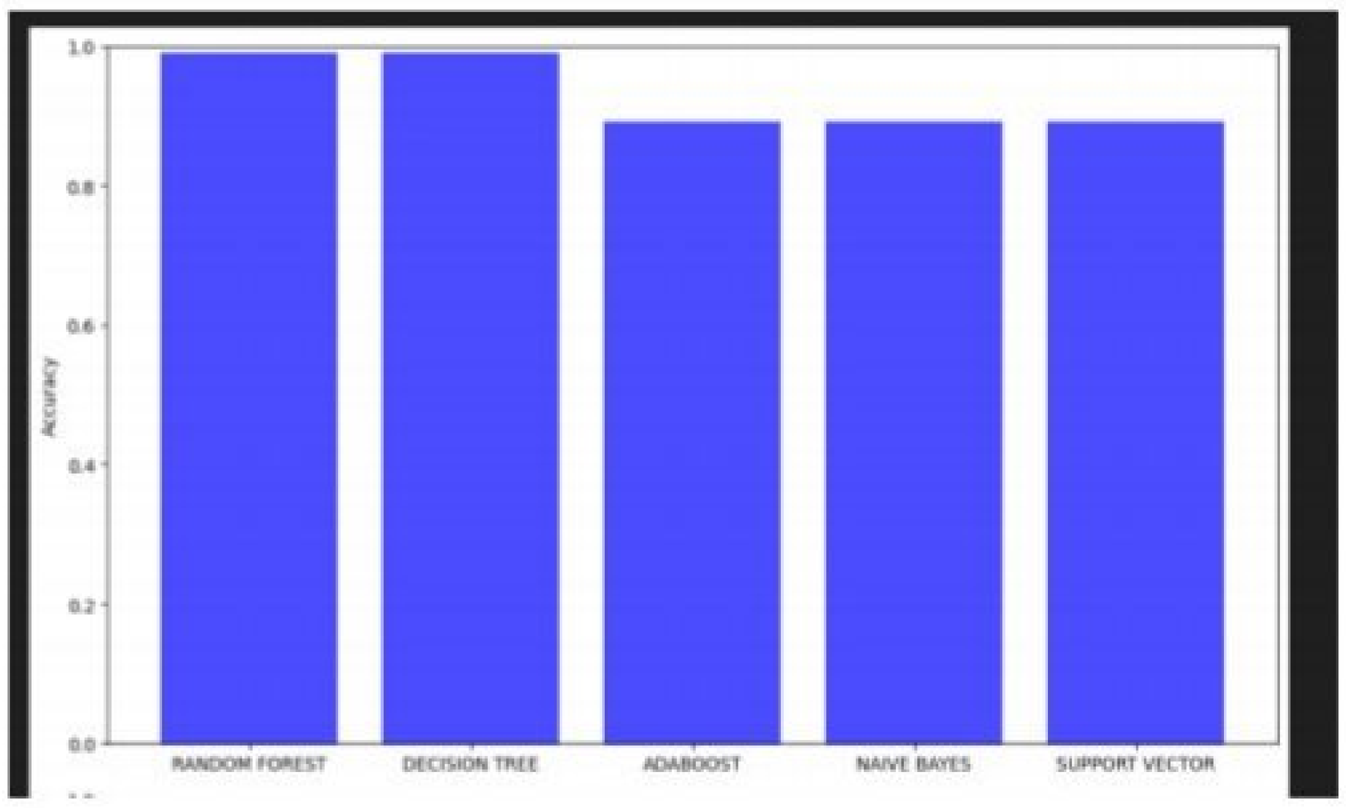
Symptom4

Symptom5

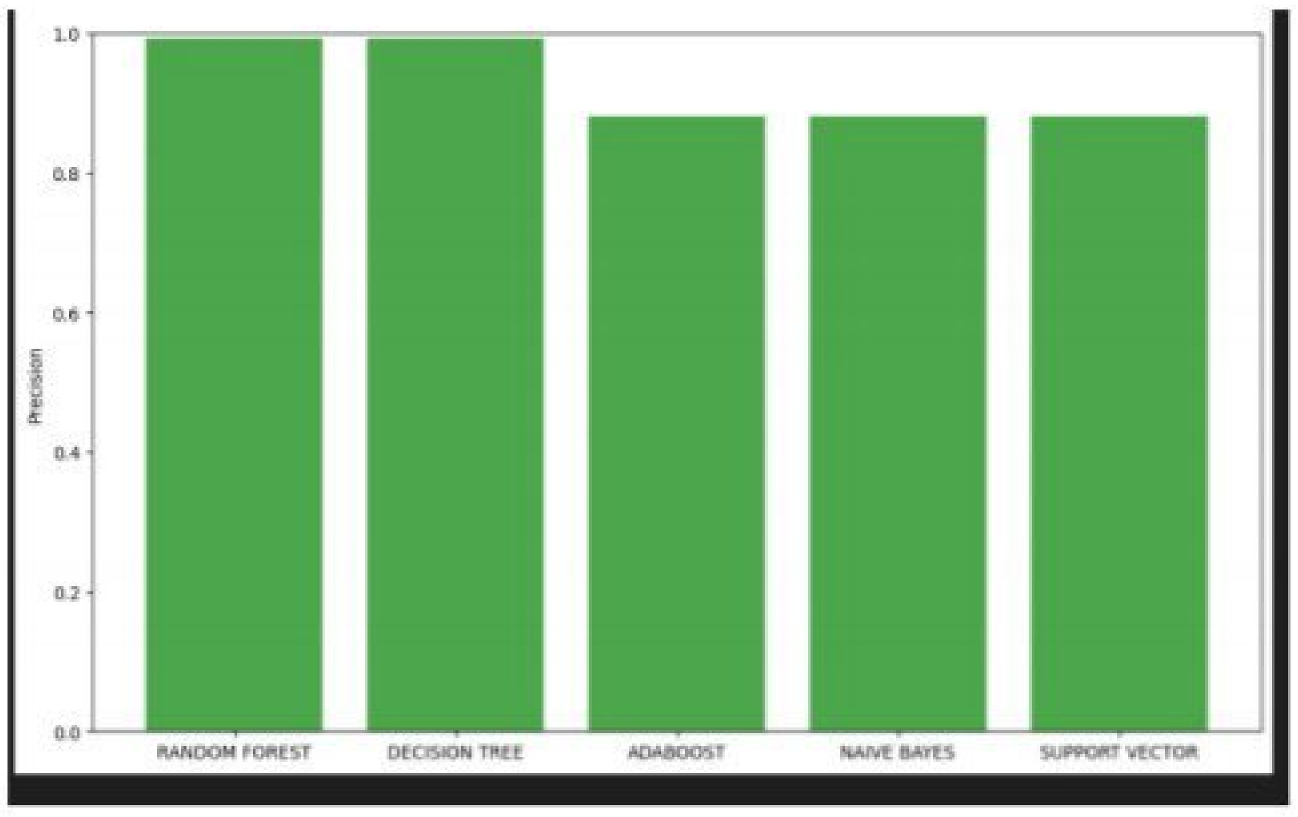
Output
Predicted Disease is ['Hyperthyroidism']

IV. Comparison

Among the algorithms under consideration, decision tree and random forest models consistently showed the highest accuracy in a comparative comparison. Decision trees are well-known for being easily understood and straightforward, which makes them useful for identifying certain patterns in data. An ensemble technique based on decision trees called random forests uses several trees to reduce overfitting and improve prediction accuracy. However, models with comparatively lesser accuracy in disease prediction were Adaboost, Support Vector Machines (SVM), and Gaussian Naive Bayes. Adaboost, while effective in boosting the performance of weak classifiers, may face challenges in capturing intricate relationships within complex medical datasets. Even though SVM is effective in high-dimensional environments, it may not perform well in noisy or big datasets. Because Gaussian Naive Bayes relies on the concept of feature independence, it may oversimplify complex correlations found in data connected to diseases.



Here in the Graphs we can see the comparison between the different algorithms which we used to train model and the accuracy as well as the precision obtained from each and every algorithm. After applying the algorithms to our model and comparing the accuracies of the same we can conclude that Random Forest algorithm has the highest accuracy following decision tree algorithm. We can see that adaboost and other algorithms fall short because the decision tree has already grown very complex. The results for Precision is same as that for the accuracy with random forest and decision tree as prevailing algorithms with the highest precision.



It's important to remember that these models' performance can change based on the features of the dataset and the type of sickness that's being predicted. Furthermore, the results can be greatly impacted by the features used, the hyperparameters adjusted, and the data preprocessing techniques used. The choice of an appropriate model for illness prediction should ultimately be determined by carefully analyzing the data and taking into account the unique difficulties presented by the healthcare industry.

V. Future Directions

In the future, new features can be added to improve the accuracy of the prediction model. For example, new disease data have an impact on the number of systems seeking medical treatment and the hospital. In addition, in the face of a larger amount of data, we can use the cloud architecture in this paper to carry out distributed computing.

We can also work on our UI by providing the user with information curated specifically for them through their past diagnosis and store it in the database which can further help the doctors as well as patient to keep track of their past diseases.

We can also add a hospital or clinic recommendation for the disease predicted to further assist the user about the next steps they can take.

Time series analysis of patient data, which looks at a patient's potential for a disease as well as when it might occur in their lifetime, is a promising tool for the future. Integrating genotype and phenotype data into the model is another approach that might be taken, which would provide researchers with an even more thorough understanding of diseases than they now have.

VI. Conclusion

Our machine learning-based disease prediction model can be an accurate tool for diagnosing various medical conditions. By leveraging diverse datasets containing symptoms, disease descriptions, and precautions, our model can demonstrate promising results in early diagnosis.

The integration of machine learning techniques enhances the efficiency of disease prediction, providing valuable insights for healthcare professionals and most importantly individuals.

This project underscores the transformative impact of machine learning in healthcare and emphasizes the need for ongoing research in this dynamic field. With the use of blockchain-system, this can be transformed into a promising system which will have centralized data of patients and help them with better and improved diagnosis."

References

1. Mrs. Jayashree L K, Sushmita Mayapur, T D Vyshnavi, T D Prathyusha, V K Kavya, 'HEART DISEASE PREDICTION SYSTEM' (Issue 06, Vol. 6 J 2019)

2. S. A. Riffat, F. Harun and T. Hassan, "An Interactive Tele-Medicine System via Android Application", 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), pp. 148-152, 2020.

3. R. Lee, K. Chen, C. Hsiao and C. Tseng, "A Mobile Care System with Alert Mechanism", IEEE Transactions on Information Technology in Biomedicine, vol. 11, no. 5, pp. 507-517, Sept 2007

4. A. Jayed Islam, M. Mehedi Farhad, S. Shahriar Alam, S. Chakraborty, M. Mahmudul Hasan and M. Siddat Bin Nesar, "Design Development and Performance Analysis of a Low-Cost Health-Care Monitoring System ", 2018 International Conference on Innovations in Science Engineering and Technology (ICISET), pp. 401-406, 2018.

5. A. Tikotikar and M. Kodabagi, "A survey on technique for prediction of disease in medical data," 2017 International Conference on Smart Technologies for Smart Nation (SmartTechCon), 2017, pp. 550- 555. https:// doi: 10.1109/SmartTechCon.2017.8358 432.

6. S. Grampurohit and C. Sagarnal, "Disease Prediction using Machine Learning Algorithms," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-7, https://doi: 10.1109/INCET49848.2020.91541 30.