International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Comprehensive Approach For Symptoms-Driven Multiple Disease Detection using Machine Learning Algorithms

[1]Dr. Sreedhar Bhukya, [2]D. Saikeerthana

Submitted: 25/01/2024 **Revised**: 03/03/2024 **Accepted**: 11/03/2024

Abstract: Computer-Assisted; In medical analysis, diagnosis is a rapidly developing, multifaceted topic of research. The development of computer-aided diagnostic applications has garnered significant attention in recent years due to the potential for seriously misleading medical therapies resulting from errors in medical diagnosis systems. It is essential to use machine learning (ML) to computer-aided diagnostic testing. An item, like bodily organs, cannot be correctly identified by a simple equation. For this reason, pattern recognition essentially requires learning from examples. Pattern recognition and machine learning (ML) have the potential to increase the accuracy of disease approach and diagnosis in the field of biomedicine. They also honour the impartiality of the decision-making process. Creating an excellent, automated system for the analysis of high-dimensional, multi-modal biomedical data may be accomplished with the help of machine learning (ML). This survey research examines the similarities and differences of many machine learning algorithms for the diagnosis of different illnesses, including diabetes and heart disease. It focuses on a collection of machine learning methods and algorithms used in disease detection and decision-making, such as Random-Forest, Naive Bayes Classifier, Decision Tree, and Voting Classifier.

***Keywords**: Voting classifier, random forest, decision tree, patient, doctor*

1. Introduction

Every day, a vast amount of patient-related data is gathered by our medical care department, including clinical assessments, necessary boundaries, examination reports, therapy follow-ups, medication selections, and more. Sadly, though, it isn't properly explored and mined. It is either handled in the record studio as hard circle area or as bundles of paper sheets. Experts and investigators alike are hurried and anxious due to this massive amount of data. In the government and some organisations, highly skilled analysts handle the majority of the information. In the future, we shall be supervised by the precision and advancement of automated systems. The feasibility of therapies, clinical trials, drugs, and the identification of links between clinical and decision information to use machine learning systems will be helpful to executives handling a variety of illnesses.Data mining is becoming more and more necessary in the clinical and medical fields. When specific information mining techniques are applied correctly, important data can be extracted from massive data sets, enabling healthcare professionals to make quick decisions and improve patient services. Utilising the grouping to aid the doctor is the soul point. Illnesses and conditions related to physical fitness,

like diarrhoea, chicken pox, migraines, diabetes, impetigo, jaundice, dengue, and so on, can have a serious influence on a person's health and occasionally even cause death if left untreated. By "mining" their enormous information base, the medical services sector can identify a potent dynamic. For instance, by eliminating linkages and hidden examples from the data set. Machine learning models such as the Random-Forest, Naive Bayes, Decision Tree, and Voting Classifiers can provide an answer for the current situation.the typical collection of distinct models and methods, we have developed a robotized framework that is able to identify and extract confidential information associated with infections from a historical dataset of diseases and their side effects. The primary goal of the work is

1. To describe the diseases using various algorithms, such as Voting Classifier, Random Forest ,Naive Bayes, and Decision Tree.
2. To determine which major risk factors are responsible for these diseases.
3. Comparing various arrangement techniques and determining which characterization strategy works best with the information provided.
4. To look into how changing one risk factor for another during characterization affects the other (e.g., diabetes by hypertension, cardiovascular sickness, or smoking).

[1]*Professor, Department of Computer Science and Engineering Sreenidhi institute of Science and Technology Hyderabad, India sreedharb@sreenidhi.edu.in*
[2]*Student of M.Tech, Department of Computer Science and Engineering Sreenidhi institute of Science and Technology Hyderabad, India Keerthidanam22@gmail.com*

In order to assess symptom data and forecast the probability of different diseases, this research presents a novel method of disease detection called "Symptom-Driven Disease Detection," which makes use of machine learning techniques. Our technique takes into account a broad variety of symptoms and their possible connections with various diseases at the same time, in contrast to standard diagnostic methods that concentrate on certain diseases or symptoms.

The conventional method of diagnosing illness is systematically assessing the symptoms of the patient, then focusing on particular tests and examinations to confirm or rule out particular illnesses. Although this approach has shown some promise, it frequently takes a long time and significantly depends on the knowledge of medical specialists. Furthermore, it could miss any connections between symptoms and illnesses that are not immediately noticeable. So, In this paper, we get the disease names accurately.

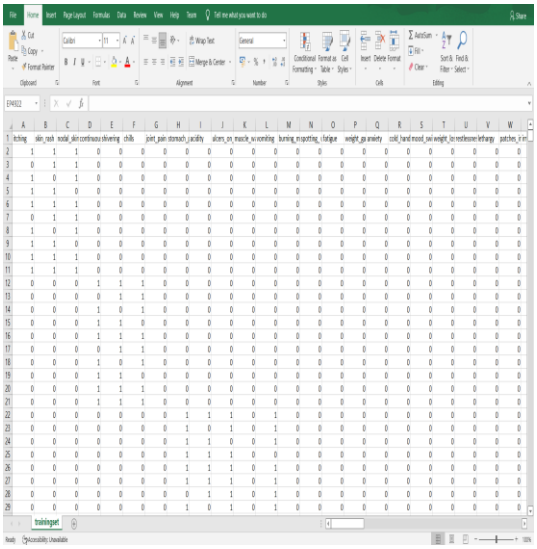
2. Literature

ImplementAccording to a summary provided by McKinsey [1], 80% of clinical consideration charges in the United States are used for treating persistent illnesses, and half of all Americans suffer from at least one continuous condition. As our standards for daily conveniences rise, the prevalence of chronic illness is rising. The annual average expenditure of the United States on treating chronic illnesses is 2.7 trillion USD. This amount represents eighteen percent of the US GDP annually. Persistent infections are also a major problem for medical services in many other countries. In China, chronic diseases are the main cause of mortality; 86.6% of deaths in the country are attributed to chronic illnesses, citing a 2015 Chinese report on nutrition and chronic infections. That is why doing danger assessments for chronic conditions is essential. As clinical information has evolved [2], creating electronic health records has grown more advantageous [3].

Moreover, [4-6] was the first to offer a superior heterogeneous vehicular telematics viewpoint that was bioinspired in order to accomplish the setup of state-of-the-art heterogeneous vehicular organisations to facilitate the assortment of dissimilar clients' health-related continual big information[7-10].

Dataset Collection

I'm incorporating the Disease Prediction dataset, which comes from the Kaggle data collection, in this study. 4921 cases total in the dataset, and 133 qualities or characteristics—which are depicted in figure—are classified as illness symptoms.



1. METHODOLOGY

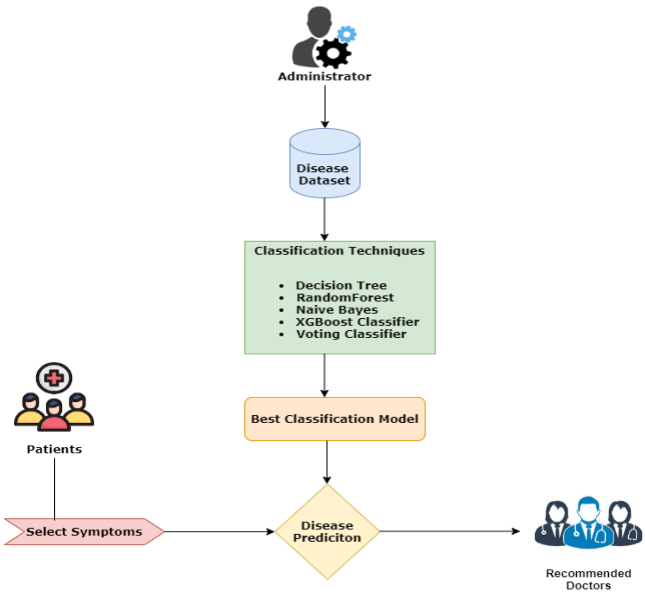


Fig 1: shown the patient details

Figure 1: shows Both physicians and patients will register with this system, and the administrator will use a variety of disease datasets to assess classification algorithms and identify the optimal model with the highest accuracy. After the patient authenticates, the system will use the symptoms they have chosen as input to predict the ailment using the best classification model. It will then present a suggested doctor's list based on the expected diagnosis.

DATA COLLECTION

Collect a thorough dataset from reputable medical sources or healthcare databases that includes symptom recordings and matching disease diagnoses. The process of measuring and obtaining information on desired variables in a way that makes it possible to find and apply the data-related questions in different kinds of study is known as data collection[11-14].

DATA PREPROCESSING:

Scaling numerical features, encoding categorical variables, and handling missing values will all help to clean up the dataset. To train models effectively, make sure the data is clean and well-maintained.Using Panda's library, I loaded the dataset and divided it into independent variables like dataset features and dependent variables like goal feature during the data pre-processing stage[15-17]. In this case, the anticipated disease name is the target feature, and the traits are treated as symptoms.

Feature extraction Finding relevant elements from the dataset that are suggestive of various conditions is known as feature extraction. These attributes could include demographic information, medical history, and specific symptoms that thepatient has described.

Feature Selection: To enhance model performance and computational efficiency, determine which characteristics are most significant using methods like feature importance ranking or dimensionality reduction. **Base Classifiers**: Select from a range of simple classifiers, such as Random Forest(RF), Decision Trees(DT), Support Vector Machines (SVM), Naive Bayes classifiers(NV), that have been trained on various feature space subsets.

Ensemble learning: To increase overall prediction accuracy and resilience, train the basic classifiers separately on the dataset and then aggregate their predictions using the Voting Classifier.

Voting Classifier Algorithm:Make use of the Voting Classifier algorithm, which aggregates the predictions of several base classifiers by means of voting procedures (e.g., majority voting or weighted voting).

Combination of multiple classifiers: The Voting Classifier algorithm combines the forecasts from several base classifiers, each of which was trained on the same dataset but with distinct feature sets or techniques.

This diversity lowers the chance of overfitting and aids in capturing various facets of the data.

Voting mechanisms: The Voting Classifier uses two primary categories of voting mechanisms:

i)Hard Voting: A simple majority vote among the base classifiers determines the projected class label in a hard vote.
ii)Soft Voting: The predicted class label in soft voting is determined by selecting the class with the highest average likelihood.

Weighted voting: Base classifiers in the ensemble may optionally be given weights, which would increase the weight of classifiers that perform better or are more reliable.

Flexibility: The Voting Classifier may be used with a variety of fundamental classifier combinations, including k-nearest neighbours, logistic regression, decision trees, support vector machines, and more. This flexibility allows for the creation of many ensembles that are tailored to the distinct characteristics of the dataset.

VOTING CLASSIFIER WORKS AS:

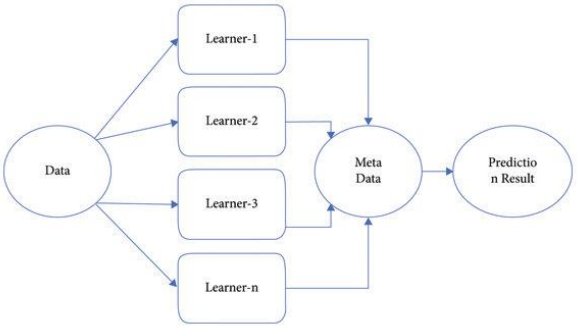
TRAINING PHASE:Using different algorithms or feature subsets, each base classifier in the ensemble is trained independently on the trainingdataset.
Every basic classifier gains the ability to forecast by training on the basis of its comprehension of the data and the underlying patterns.

VOTING PHASE: The Voting Classifier aggregates the predictions made by the base classifiers using the chosen voting method (hard or soft voting) during the voting phase.
In a hard vote, the class label that shows up most frequently in the base classifiers' predictions is selected as the winning prediction.

By summing the projected probabilities that the base classifiers assigned to each class predict, the class with the greatest average probability is selected through soft voting.

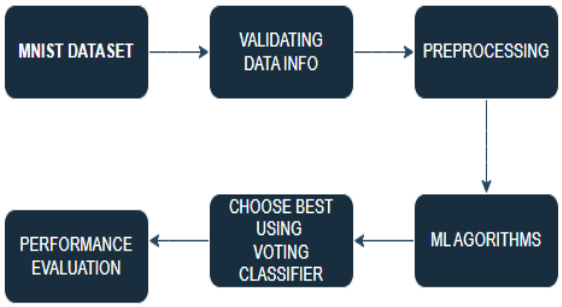
PREDICTION PHASE Once trained, the Voting Classifier can make predictions on new unseen data by passing it through each base classifier and combining their predictions using the voting mechanism.

To sum up, the Voting Classifier algorithm is an effective ensemble learning technique that enhances prediction accuracy and robustness by combining the predictions of several basic classifiers. It is a well-liked option for many different machine learning classification jobs due to its adaptability and efficiency.



Flowchart of voting classifier

3. System Design



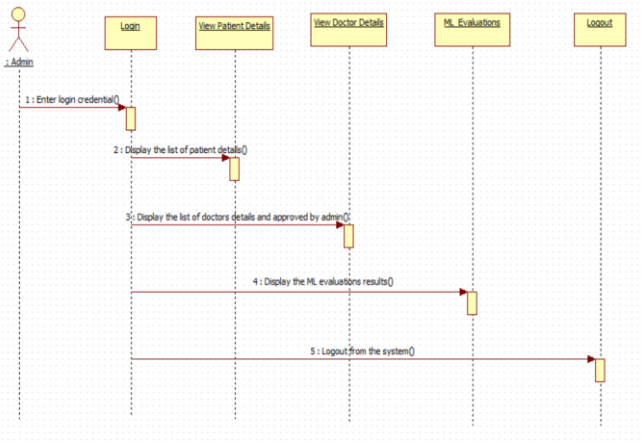
The above figure shows flow of the data where mnist dataset is used after taking the symptoms and validating all the values we perform preprocessing techniques .After that the data will go through ML algorithms. Using Voting classifier we get the best value.

Database Schema

	Field Name	Datatype	Len
*	sno	int	10
	patient_uid	varchar	500
	patient_name	varchar	500
	patient_age	varchar	500
	patient_gender	varchar	500
	symptoms	varchar	1000
	disease	varchar	500
	doctor_uid	varchar	500
	apdate	varchar	500
	aptime	varchar	100
	status	varchar	100

The bkappointments table will store the doctor booking appointment details which is updated by patients.In the database schema we have data for patient, doctor, book appointments. In this schema we have serial number, patient _uid, patient name, patient age, patient gender, symptoms, disease, doctor_uid and status.

FUNCTIONAL REQUIREMENTS



Figureshows the administrator will execute all objects sequentially when they are login with this system successfully.

Classifier Methods

RandomForest:

Regression and classification issues are the main applications for supervised machine learning technique called as the RF classifier. The Random Forest (RF)will forecast the outcomes by considering several decision tree outcomes and determining which outcome is more likely to be a voter than a projected class. To predict illness in this system, I utilised RandomForestClassifier(RFC), which is imported from the sklearn ensemble package. According on the test findings, the RF classifier has an 88% accuracy rate.

Naïve Bayes:

A supervised machine learning technique that uses the Bayes theorem as its foundation is the Naive Bayes classifier. The Bayes theorem will determine the likelihood of each anticipated sickness, giving the maximum likelihood of a predicted disease. For illness prediction, I had imported the BernoulliNB classifier from the sklearn naive_bayes package. The experimental findings show an accuracy of 86.17 percent for the NB classifier.

Decision Tree:

For the purpose of addressing classification issues, the supervised learning classifier known as the DT classifier is a tree-based predicted classifier. The IF-THEN approach will be used by the decision tree classifier to forecast the illness. With the use of a dataset, the decision tree classifier will build a tree structure during training. Then, using testing inputs, it will compare each child node to the predicted class leaf node.

The DecisionTreeClassifier, a preset classifier imported from the sklearn. tree package, is utilised in this system to forecast diseases. Based on testing results, the decision tree provided an accuracy of 82.18 percent..

VotingClassifier

To improve prediction precision, I've included a voting classifier to this investigation. This estimator is regarded as machine learning as it will learn from the estimators or model classifiers and provide predictions based on the voting of each estimator output. Soft voting and hard voting are the two accessible voting procedures.. In this case, the output class's estimated output probability will serve as the basis for soft voting, while the anticipated output class will determine hard voting.

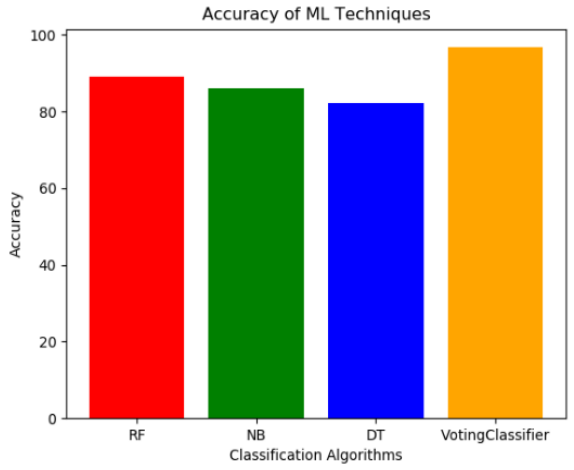
Based on test findings, the voting classifier achieves a 95% accuracy rate in predicting diseases. The preconfigured VotingClassifier, which is also imported from the sklearn ensemble package, is used to forecast diseases.

TECHNOLOGY

- **Data Gathering :** Putting together your data is the first step towards tackling any machine learning problem. We'll be using a Kaggle dataset for this problem. This dataset consists of two CSV files, one for testing and one for training. The dataset has 133 total columns, of which 132 display the symptoms and the final column the prognosis.
- **Data Cleaning:** Cleaning is the most important stage in a machine learning project. The quality of our data determines how well our machine-learning model performs. Therefore, before giving the data to the model for training, it must always be cleaned. With the exception of the objective column, prognosis, which is a string type that is transformed to numerical form using a label encoder, every column in our dataset is numerical.
- **Building Model:** Once gathered and cleaned, the data may be utilised to train a machine learning model. This cleaned data will be used to train the Random Forest, Naive Bayes, and Support Vector classifiers. We'll use a confusion matrix to evaluate the model's quality.
- **Inference** After the models have been trained, we will be able to predict the illness based on the input symptoms by combining the predictions of all three models.
- Consequently, generally, our prognosis is more accurate and dependable.

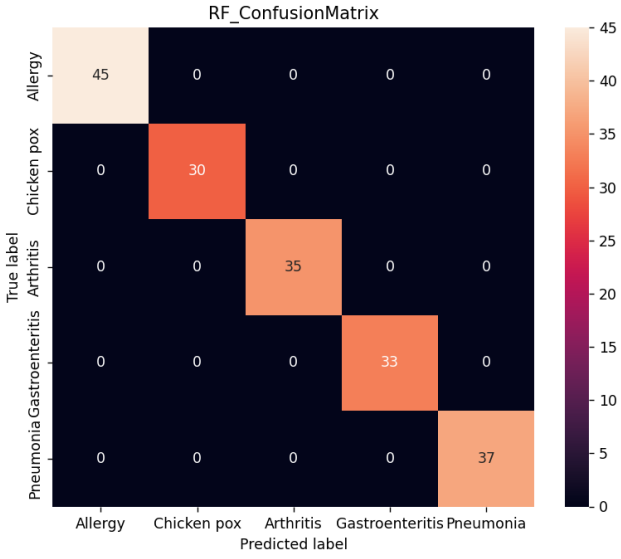
BAR GRAPH:

In the bar visualisation, this system will display the accuracy of four classifiers: the voting classifier (94.64 percent), NB (86.17 percent), DT (82.18 percent), and RF (88.00 percent). Therefore, when compared to other machine learning methods; the Voting Classifier has produced the best accuracy results.



BAR GRAPH

CONFUSION MATRIX



The above confusion matrix shows the following diseases are predicted with number of times correctly based on

the RF classifier, such as, Allergy disease is predicted 45 times, chicken pox is predicted 30 times, arthritis is predicted 35 times, gastroenteritis is predicted 33 times, and Pneumonia is predicted 37 times.

System Testing

Validation testing

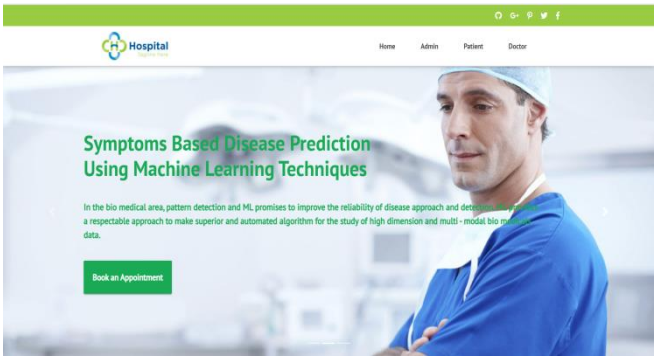
The front-end application, including the registration and login forms, will have undergone validation testing. Therefore, it will verify during validation testing whether or

not the essential fields have content entered. The user will receive a warning message if a field is not filled out correctly, prompting them to fill up the relevant data. As a result, the form should not be submitted with any blank fields.

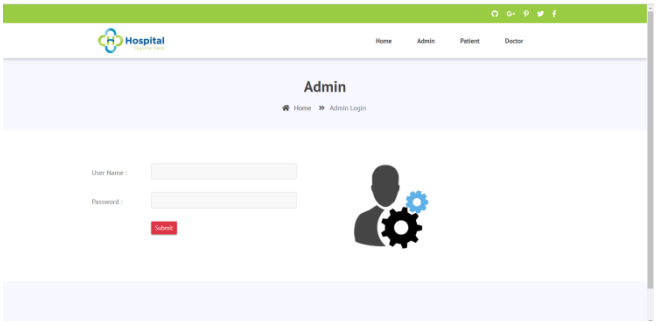
Verification Testing

Once the validation testing is over, the verification testing is carried out. Validation testing verifies that the field contents match those on the back-end server, such as a database. The end-user will get a warning box noting them of the invalid content if any of the fields' contents are missing from the database server. Thus, the final user will understand and provide the appropriate material.

OUTPUTS:

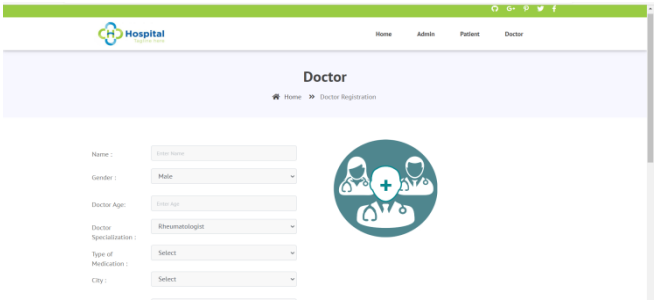


The image above illustrates how running the index.py Python file will cause the application's primary index page to open.



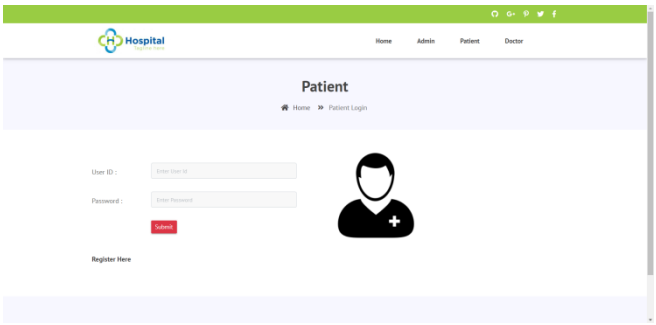
Admin Login

An admin login page, as seen in figure, is the page that appears when I click on an admin link. The administrator will enter their legitimate login and password here to log in.



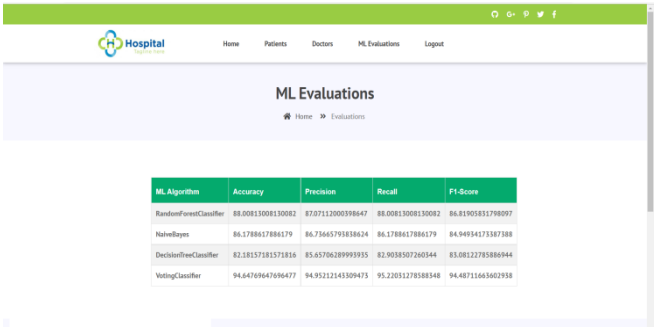
Doctor registration

When I click the Register Here link from the doctor login page, the doctor registration page will open. The doctor will complete the registration form on this page by entering the required fields, which include name, gender, age of the doctor, specification, kind of drug, etc.



PATIENT PAGE TO LOGIN

A patient link will take me to the patient login page, as seen in figure, when I click on it. The patient will use their legitimate user ID and password to log in here.



ML Evaluation

This picture shows the all-machine learning classifier's performances and matrices.

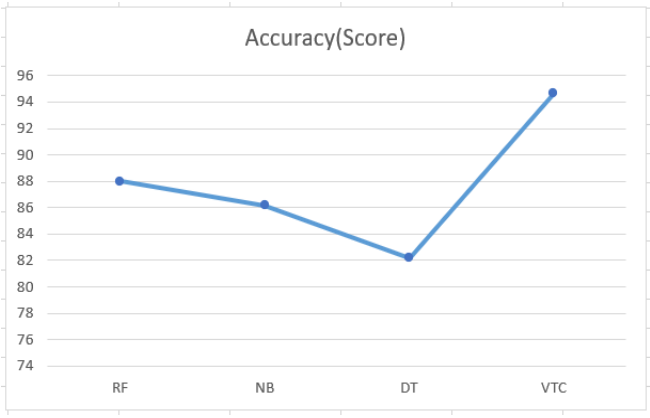
In this instance, the dataset for sickness prediction will be split into two categories: 70% will be used as the training dataset, and the remaining 30% will be used as the testing dataset.

These training and testing datasets are trained using the RF, NB, DT, and voting classifiers to calculate different matrices and all-model outputs, such as accuracy, precision, recall, and F1-score.

TEST CASES

Sl.no	Pre-requisite	Test Input	Expected outcome
1.	Positive test: A Known disease detected.	Flu-related symptoms, such as fever, coughing, and bodily pains	"Influenza" or the related illness code designating influenza
2.	Negative Test: No Illness Found	symptoms (such as sneezing and watery eyes) that don't match any recognised disease	No disease detected" or similar message indicating no match
3.	Performance Test: Situation with High Load	varying the number of concurrent users or requests to replicate times of peak use	Even with a high load, the system maintains acceptable response times.
4.	Security Check: Patient Data Protection	Injection attacks, efforts to gain unauthorised access, and breaches of patient privacy	Patient information is protected, and the system guards against breaches and unwanted access.

Numerous situations are covered by these test cases, such as performance under load, security issues, and positive and negative illness detection. While the anticipated outcome values describe the expected outcomes or behaviours for each test case, the potential values reflect other inputs or situations that will be examined.



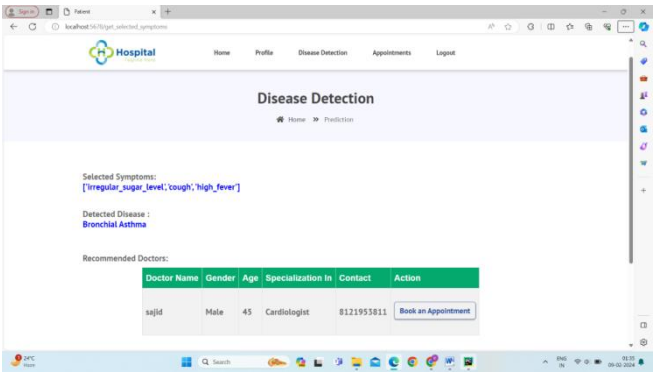
Line Chart with Accuracy Score of ML Classifier

This system generates the line chart graph with the accuracy score of performance metrics of the ML classifier

Table.3. ML and DL models performance comparisons**:**

Algorithm name	Accuracy
Random forest	88%
Naive Bayes	8.1%
Decision Tree	82.18%
Voting Classifier	95%

RESULT SCREEN SHORTS



This is the result page which displays the predicted disease with given multiple disease.

4. Conclusion

Statistical prediction models that can't produce high-quality results have overtaken the evaluation industry. Statistical models are not effective in preserving generalised knowledge because they cannot handle broad data points and missing values. These are all the reasons

why MLT is valuable. Machine learning is used in many areas, including data mining, image identification, natural language processing, and illness diagnosis. Potential answers can be found in machine learning for each of these areas.This work examines many machine learning methods for the diagnosis of different diseases such as diabetes and heart conditions. Due to their ability to precisely characterise the characteristic, the majority of models have produced great outcomes. Ninety-five percent of the maximum categorization accuracy is offered.

These methods offer chances for a better decision-making process and are highly helpful for the examination of specific issues.

References

[1] S. Mitra, S.K.Pal & Mitra , P., Data mining in soft computing framework: A survey, IEEE transactions on neural networks, 13(1), 314,2018.

[2] Krzysztof J. Cios, G.William Moore, Uniqueness of medical data mining, Artificial Intelligence in Medicine 26, 1–24, 2017.

[3] Parvez Ahmad, Saqib Qamar, Syed QasimAfser Rizvi, Techniques of Data Mining in Healthcare: A Review, International Journal of Computer Applications (0975 – 8887) Volume 120 – No.15, June 2017.

[4] Hsinchun Chen, Sherrilynne, S. Fuller, Carol Friedman and William Hersh, Knowledge Management, Data Mining and text mining inmedical informatics.

[5] V. krishnaiah, G. Narsimha, & N. Subhash Chandra, A study on clinical prediction using Data Mining techniques, International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN 2249-6831 Vol. 3, Issue 1, 239 248, March 2017.

[6] Divya Tomar and Sonali Agarwal , A survey on data mining approaches for healthcare, International Journal of Bio-Science and Bio-Technology Vol.No.5, pp. 241-266, 2017.7.

[7] Mohammed Abdul Khalid, Sateesh kumar Pradhan, G.N.Dash, F.A.Mazarbhuiya, A survey of data mining techniques on medical data for finding temporally frequent diseases", International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 12, December 2018.

[8] S.D.Gheware, A.S.Kejkar, S.M.Tondare, Data Mining: Task, Tools, Techniques and Applications, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 10, October 2017.

[9] Yongjian Fu , Data Mining : Tasks, Techniques and Applications

[10] http://academic.csuohio.edu/fuy/Pub/pot97.pdf

[11] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2017.

[12] G. Beller, J. Nucl. Cardiol. "The rising cost of health care in the United States: is itmaking the United States globally noncompetitive?" vol. 15, no. 4, pp. 481-482, 2018.

[13] Pang-Ning Tan, Michael Steinbach ,Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2016.

[14] Gosain, A.; Kumar, A., "Analysis of health care data using different data miningtechniques," Intelligent Agent & Multi-Agent Systems, 2017. IAMA 2009, International Conference on, vol. no., pp.1,6, 22-24 July 2018.

[15] Dr. M.H.Dunham, "Data Mining, Introductory and Advanced Topics", Prentice Hall, 2017. 14. A. S. Elmaghraby, et al. Data Mining from multimedia patient records. 6, 2017.

[16] Nada Lavrac, BlažZupan, "Data Mining in Medicine" in Data Mining and Knowledge

[17] Discovery Handbook, 2018.

[18] Sreedhar Bhukya, Quality-aware energy efficient scheduling model for fog computing comprised IoT network, Volume 97, January 2022, 107603.

[19] Sreedhar Bhukya, QOS Based Service Composition for Various Cloud Users using Chebyshev Distance & Evolutionary Fitness Function, ISSN NO : 1006-6748, 2021.