# DISEASE PREDICTION USING ENSEMBLE LEARNING


Nandhitha Kothi
Computer science
Illinois State University
Normal, Illinois, USA
nkothi@ilstu.edu

Divya Konanki
Computer science
Illinois State University
Normal, Illinois, USA
dkonank@ilstu.edu

Sai Sree Harsha Idarapalli
Computer science
Illinois State University
Normal, Illinois, USA
sidarap@ilstu.edu

Venkata Jaswanth Malineni
Computer science
Illinois State University
Normal, Illinois, USA
vmaline@ilstu.edu



*Abstract* - **The healthcare industry is among the most vital for any nation. The health sector needs to be advanced with new technologies like artificial intelligence, machine learning, big data, etc. due to the growing population. Physicians utilize symptoms to diagnose illnesses. A specific disease is brought on by a particular set of symptoms. With the aid of machine learning algorithms, diseases can be predicted by analyzing these patterns of symptoms throughout a medical database. When the model is given symptoms, a machine learning model is developed that has the ability to predict the illness. Many deaths are avoided when diseases are diagnosed early. This machine learning model aids in the early diagnosis of illnesses by doctors.**

**A dataset with various symptoms and the corresponding diseases is taken for this research. This dataset is used to train and test algorithms such as Decision Tree, Random Forest, SVM, and Naïve Bayes. The disease is then predicted using the maximum votes from the ensemble learning process.**



**Keywords:** Ensemble Learning, Decision Tree, SVM, Random Forest Classifier, Naïve Bayes, Python.


## I. INTRODUCTION

Large data sets, often measured in petabytes, are found in the medical domain. Modern data mining techniques are used to deal with massive amounts of data and uncover hidden patterns. Nowadays, there are considerably more patients than doctors in the world. Thus, it is challenging for physicians to review all patient data and determine each patient's ailment. There are illnesses that can be fatal if care is not received in a timely manner. A specific disease is brought on by a particular set of symptoms. Using machine learning approaches to identify these patterns of illnesses and symptoms throughout a medical database. When the model is given symptoms, a machine learning algorithm is developed that has the ability to predict the illness. We used separate training and testing datasets to construct this model. This includes signs of the corresponding illnesses. There are 41 distinct disease types and 131 symptoms in these databases. There are 4921 samples in the training dataset and 41 samples in the testing dataset.

Algorithms for machine learning Using training datasets, Random Forest, Decision Tree, SVM, and Naïve Bayes algorithms are taught, while testing datasets are used to assess how well these algorithms predict diseases. Then, with the use of ensemble techniques, these four algorithms serve as the foundational models and provide a single optimal result.

## II. LITERATURE SURVEY

Numerous investigations have been carried out to predict diseases from the symptoms that are present. One person created a model based on statistics that could determine whether an individual had influenza by using machine learning techniques. There were a lot of healthy adults and teenagers among them who had fever and at least two additional symptoms that were consistent with influenza. Out of 3744 individuals, 2470 have been determined to have had an influenza diagnosis in a lab. These data serve as the foundation for their conclusions. Using a random-forest machine-learning algorithm, disease was anticipated based on symptoms. The method produced low time and low-cost results for disease prediction. Using the symptoms alone, the random forest machine learning algorithm can diagnose a disease.

"Machine Learning for Disease Prediction" is the name of the system that produced results for disease prediction with little time and expense. Data analysis is an essential component of all disciplines, and data mining is a step that integrates that strategy into the healthcare foundation. The process of making predictions about data for healthcare is known as data mining. The scope of medical care is quickly growing. The current one describes the general state of health care system and forecasts healthcare data in addition to (i) assessing, (ii) managing, and (iii) communicating. The idea of machine learning is applied to data pertaining to illnesses. The retrieval and treatment procedures in these kinds of processes are To do this, data analysis was employed. Decision trees are therefore used to predict disease outbreaks. It works pretty well.

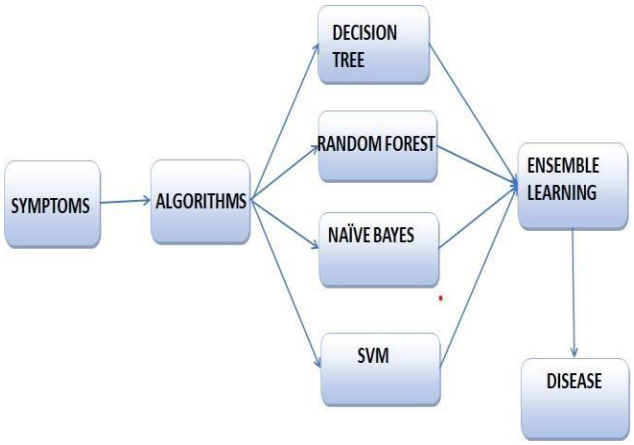
The concept of " Enhancing disease prediction by machine learning," which refers to using machine learning to improve disease prediction, is presented by the author of this paper [9], and this concept-based experiment shows that this is the outcome. Big data is improving this kind of data by increasing the amount of medical data. By using the genetic algorithm, this idea recovers missing data and incorporates medical data into the dataset. This system makes use of both.

## III. METHODOLOGY

### 1. PROCEDURE

Steps to build the disease prediction model:
Step 1: We read and process the dataset in this first step.
Step ii: The dataset that can fit and work with our model has now been processed.
Step iii: Our machine learning models, Decision Tree, Random Forest, Support Vector Machine, and Naïve Bayes, are trained using the training dataset.
Step IV: A testing dataset is used to assess how well these algorithms perform.
Step V: To produce the best results, all four models are combined to create basic models.
Step vi: To display the output, a GUI window is created. This graphical user interface window has three fields for symptoms that the user can enter. A final prediction button allows the user to click on the button to see the output.

### 2. FLOWCHART

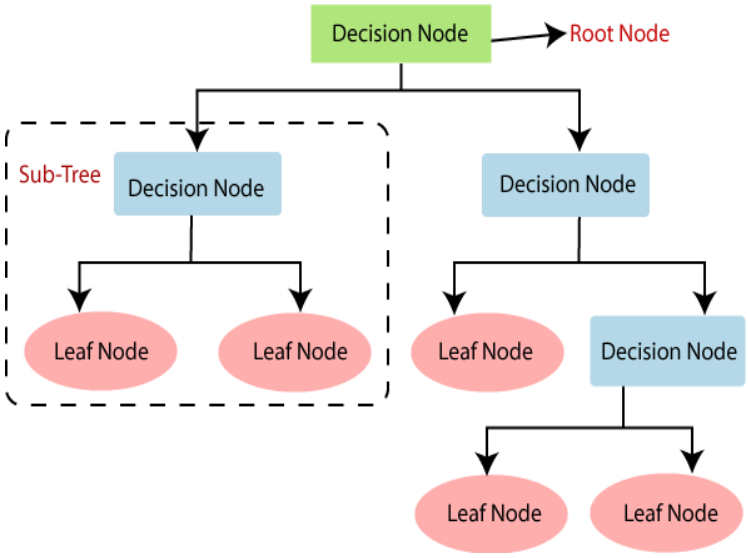



### 3. ALGORITHMS

Decision Tree

With internal nodes representing dataset features, branches representing decision rules, and each leaf node representing the result, a decision tree is a tree-structured classifier. Decision Node and Leaf Node are the two nodes in a decision tree. In contrast to leaf nodes, which are the outcomes of decisions, decision nodes are used to make any kind of decision and have many branches. Simply put, a decision tree divides itself into subtrees based on whether the answer to a question Yes or No.

The decision tree algorithm internally constructs the tree when it receives a dataset. The decision tree algorithm must choose how to build the tree and, in the process, must choose which node should be the root. All of the independent variables are potential candidates for the root node; however, the internal algorithm must determine how and under what conditions to split the independent variables before deciding which one to use for the decision tree.



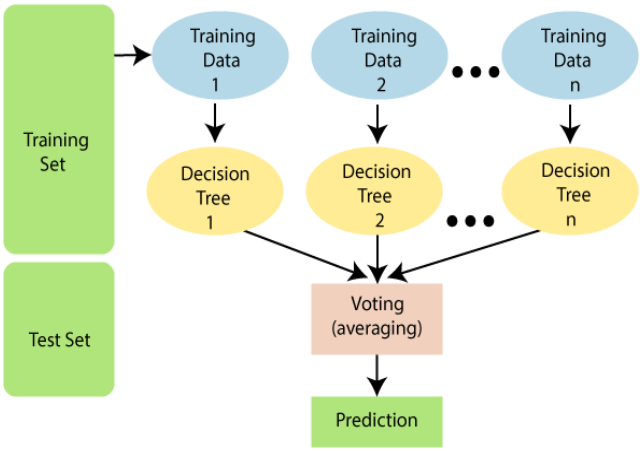

Random Forest Classifier

To increase dataset accuracy, a classifier called random forest averages several decision trees on different subsets of the given dataset. Taking predictions from each decision tree, the random forest predicts the result according to the majority vote of predictions, as opposed to depending just on one decision tree. The accuracy improves with the number of trees in the forest. The concept of group learning serves as its foundation. To put it simply, Random Forest is an assemblage of Decision Trees.

Working:

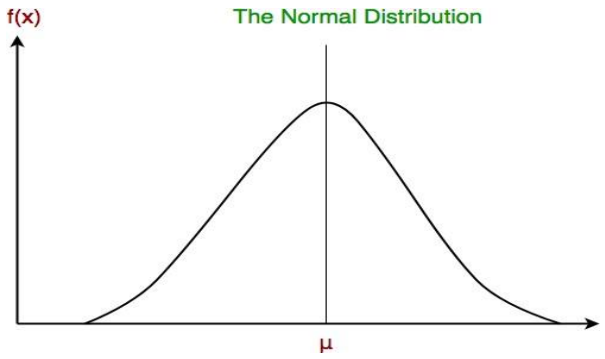
- A dataset with m total symptoms is selected at random to yield K symptoms. For these k symptoms, k decision trees are then constructed.
- Continue doing this n times so that n trees are created for each k symptoms.
- A random variable is assigned to each of the n-built decision trees to predict the Disease. We have a total of n Decision trees, which result in n disease predictions, once the predicted disease is saved.

- Counts the output from n decision trees that appeared the most frequently.



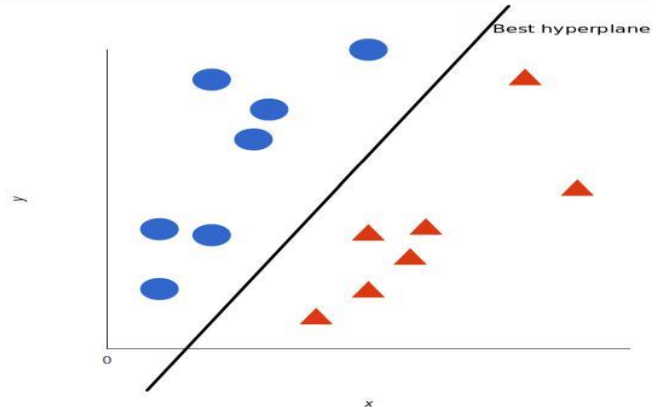

Naïve Byes

The foundation of Naïve Bayes Classifier is Bayes' Theorem. Because it makes predictions based on an object's probability, it is a probabilistic classifier. Each pair of features being classified is independent of the other, and this is the common principle shared by the family of algorithms. A variation of Naive Bayes that uses the Gaussian normal distribution and continuous data is called Gaussian Naive Bayes. A symmetric probability distribution around the mean of the Gaussian distribution shows that data close to the mean happen more frequently than data far from it. Gaussian Nave Bayes is used when we assume that each feature's continuous variables all have a Gaussian distribution.




Support Vector Machine

One well-liked supervised learning algorithm for both regression and classification issues is support vector machine, or SVM. However, classification problems are the main application for it in machine learning. The objective of the SVM algorithm is to determine the optimal boundary or line for classifying n-dimensional space into groups so that new data points can be effortlessly inserted into the appropriate group at a later time. The best decision boundary is known as a hyper plane. SVM selects the extreme points/vectors that contribute to the creation of the hyper plane. The algorithm is referred to as a Support Vector Machine, and support vectors are the extreme cases.




## 4. ENSEMBLE LEARNING

The process of combining multiple models to produce a single output is known as ensemble learning. The primary goal of this procedure is to improve performance. While there are many approaches to group learning, some basic group strategies include 1) Max Voting: This method, which is primarily used for classification issues, prints the majority prediction that is obtained from each model.

2) Averaging: Every prediction is taken into account as an output in this average. Both regression and classification issues can benefit from this.

3) Weighted Average: This is an extra measure of average. Each model in this has a weight assigned to it according to its significance.

Additionally, there are a few sophisticated ensemble techniques like boosting, bagging, and stacking.

## 5. SYSTEM REQUIREMENTS

Hardware Requirements

- RAM: 4GB
- HDD: 40GB min
- PROCESSOR: i5 or above
- OS: WINDOWS 10

Software Requirements

- Programming language: Python 3.8.5
- OS: Windows

## 6. SOFTWARE DESCRIPTION

**PYTHON:**

High-level programming languages include Python. Guido Van Rossum introduced and developed Python in 1991. This language is very efficient because of its many unique features. Among them are: 1. This language is user-friendly, which makes it simple to comprehend and write code in.

2. It is possible to write complex problems in a few lines or less.

3. Python operates one line at a time.

4. Python also has support for object-oriented languages, which makes it easier for programmers to write reusable code.

5. A plethora of libraries are available for a variety of fields, including machine learning and GUI development.

## IV. IMPLEMENTATION:

### 1. DATA COLLECTION:

Our dataset was gathered from the Kaggle website and includes disease symptoms. We used distinct training and testing datasets that included the symptoms of the corresponding diseases. There are 41 distinct disease types and 131 symptoms in these datasets. There are 4921 samples in the training dataset and 41 samples in the testing dataset. The illnesses that we have consumed are as follows:

| | |
|---|---|
| Fungal infection | Hepatitis C |
| Allergy | Hepatitis D |
| GERD | Hepatitis E |
| Chronic cholestasis | Alcoholic hepatitis |
| Drug Reaction | Tuberculosis |
| Peptic ulcer disease | Common Cold |
| AIDS | Pneumonia |
| Diabetes | Dimorphic hemorrhoids(piles) |
| Gastroenteritis | Heart attack |
| Bronchial Asthma | Varicose veins |
| Hypertension | Hypothyroidism |
| Migraine | Hyperthyroidism |
| Cervical spondylosis | Hypoglycemia |
| Paralysis (brain hemorrhage) | Osteoarthritis |
| Jaundice | Arthritis |
| Malaria | (vertigo) Paroxysmal Positional Vertigo |
| Chicken pox | Acne |
| Dengue | Urinary tract infection |
| Typhoid | Psoriasis |
| hepatitis A | Impetigo |
| Hepatitis B | |

| | |
|---|---|
| skin_rash | bruising |
| nodal_skin_eruptions | obesity |
| continuous_sneezing | swollen_legs |
| shivering | swollen_blood_vessels |
| chills | puffy_face_and_eyes |
| joint_pain | enlarged_thyroid |
| stomach_pain | brittle_nails |
| acidity | swollen_extremeties |
| ulcers_on_tongue | excessive_hunger |
| muscle_wasting | extra_marital_contacts |
| vomiting | drying_and_tingling_lips |
| burning_micturition | slurred_speech |
| spotting_ urination | knee_pain |
| fatigue | hip_joint_pain |
| weight_gain | muscle_weakness |
| anxiety | stiff_neck |
| cold_hands_and_feets | swelling_joints |
| mood_swings | movement_stiffness |
| weight_loss | spinning_movements |
| restlessness | loss_of_balance |
| lethargy | unsteadiness |
| patches_in_throat | weakness_of_one_body_side |
| irregular_sugar_level | loss_of_smell |
| cough | bladder_discomfort |
| high_fever | foul_smell_of urine |
| sunken_eyes | continuous_feel_of_urine |
| breathlessness | passage_of_gases |
| sweating | internal_itching |
| dehydration | toxic_look_(typhos) |

These are the symptoms what we have taken:

| | |
|---|---|
| indigestion | depression |
| headache | irritability |
| yellowish_skin | muscle_pain |
| dark_urine | altered_sensorium |
| nausea | red_spots_over_body |
| loss_of_appetite | belly_pain |
| pain_behind_the_eyes | abnormal_menstruation |
| back_pain | dischromic _patches |
| constipation | watering_from_eyes |
| abdominal_pain | increased_appetite |
| diarrhoea | polyuria |
| mild_fever | family_history |
| yellow_urine | mucoid_sputum |
| yellowing_of_eyes | rusty_sputum |
| acute_liver_failure | lack_of_concentration |
| fluid_overload | visual_disturbances |
| swelling_of_stomach | receiving_blood_transfusion |
| swelled_lymph_nodes | receiving_unsterile_injections |
| malaise | coma |
| blurred_and_distorted_vision | stomach_bleeding |
| phlegm | distention_of_abdomen |
| throat_irritation | history_of_alcohol_consumption |
| redness_of_eyes | fluid_overload |
| sinus_pressure | blood_in_sputum |
| runny_nose | prominent_veins_on_calf |
| congestion | palpitations |
| chest_pain | painful_walking |
| weakness_in_limbs | pus_filled_pimples |
| fast_heart_rate | blackheads |
| pain_during_bowel_movements | scurring |
| pain_in_anal_region | skin_peeling |
| bloody_stool | silver_like_dusting |
| irritation_in_anus | small_dents_in_nails |
| neck_pain | inflammatory_nails |
| dizziness | blister |
| cramps | red_sore_around_nose |
| | yellow_crust_ooze |

**2. MERITS AND DEMERITS**

1. MERITS:

✓ We can cure them by knowing the disease and being able to predict it at an early stage.

✓ Benefits both physicians and patients.

✓ It can determine which patients are susceptible to certain illness or ailments.

✓ After that, clinicians can take the necessary actions to reduce or eliminate the risk, which will enhance patient care and prevent future hospital admissions.

2. DEMERITS:

✓ It is impossible to accurately predict two diseases with identical symptoms. Not entirely accurate.

## V. CONCLUSION AND FUTURE SCOPE

Illness Forecasting The application of machine learning algorithms offers a user-friendly setting for illness prediction based on the provided symptoms. As a result, by providing their symptoms, the user can predict their illness. Future research will concentrate on giving patients the appropriate medication and medical attention as soon as possible in order to build the best infrastructure and the most practical and expedient medical fields. In the future, the optimal subset of attributes for disease prediction will be obtained by reducing the real size of the dataset using generic techniques. The result will dictate how accurate the illness prediction was. The goal of improvements is to increase efficiency and uniformity.

## VI. REFERENCES


1. Sneha Grampurohit 2020, “Disease Prediction using Machine Learning Algorithms”(Jun. 2020)IEEE

2. Marouane Fethi Ferjani 2020, “Disease Prediction Using Machine Learning”(Dec. 2020)

3. Faliang Huang 2009, “Research on Ensemble Learning”(Sep. 2009)IEEE

4. Shahadat Uddin,Arif Khan “Comparing Different Supervised Machine Learning Algorithms for Disease Prediction” Accessed on: 21 December, 2019[Online].Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8

5. Qulan,J.R. 1986. “Induction of Decision Trees”. Mach.Learn. 1.1(Mar.1986),81-10

6. Sayantan Saha, Argha Roy Chowdhuri et,al “Web Based Disease Setection System”,IJERT, 4,April2013

7. Palli Suryachandra, Prof. Venkata Subba Reddy, “Comparison of Machine Learning algorithms For Breast Cancer”, IEEE