DISEASE PREDICTION BY SYMPTOMS

Developed By:

H.T.NO	STUDENT NAME
2203A51077	G.Anjali
2203A51441	R.Vani Priya
2203A51395	A.Deepshitha
2203A51303	K.Upasana

Under the Guidance of

Dr.N.Venkatesh

Assistant professor

Submitted to

School of Computer Science and Artificial Intelligence

SR University

Ananthasagar (V), Hasanparthy (M), Hanamkonda (Dist.) – 506371

www.sru.edu.in

DEPARTMENT OF COMPUTER SCIENCE
&
ENGINEERING

CERTIFICATE

This is to certify that the **AIML**-Course Project report entitled "DISEASE PREDICTION BY SYMPTOMS" is a record of bonafide work carried out by the students G.Anjali,R.Vani Priya,A.Deepshitha,K.Upasana bearing RollNo(s) 2203A51077, 2203A51441,2203A51395,2203A51303 during the academic year 2023-24 in partial Fulfillment of the award of the degree of Bachelor of Technology in computer science & Engineering by the **SR UNIVERSITY**,Anantasagar

Lab In-charge **Head of the Department**

TABLE OF CONTENTS

Disease prediction by symptoms and patient profile Using Machine Learning

ABSTRACT

This research paper examines the importance of disease symptoms in the diagnosis and management of various illnesses. The paper highlights the challenges of accurately identifying and interpreting disease symptoms, particularly when they are nonspecific. The research paper also discusses the implications of delayed diagnosis and the negative impact it can have on patient outcomes.

KEYWORDS

disease symptoms, management, outcomes, quality of life, symptoms, diagnosis.

INTRODUCTION

When it comes to healthcare, disease symptoms are important markers that aid in diagnosing the underlying issue and creating treatment strategies that are appropriate. Healthcare professionals can improve patient outcomes by accurately diagnosing patients and providing timely interventions by closely monitoring their symptoms.

Effective disease management depends on accurate diagnosis and interpretation of disease symptoms. The purpose of this research paper is to examine the importance of disease symptoms and the difficulties involved in identifying and interpreting them.

One branch of AI called machine learning (ML) uses data as an input resource . When predefined mathematical functions are used, classification or regression is produced, which is frequently challenging for humans to achieve. For example, by utilising ML, it is frequently easier to locate cancerous cells in a microscopic image, which is usually difficult to do just by looking at the images.

Furthermore, the most recent study demonstrates MLBDD accuracy of above 90% because to advancements in deep learning, a type of machine learning . Alzheimer's disease, heart failure, breast cancer, and paroxysmal ventricular dystrophy are among the conditions that machine learning can identify. The application of machine learning (ML) algorithms in the detection of diseases shows how useful technology is in the medical field.

Only a handful of the numerous tough fields to handle in a nutshell are recent advances in machine learning difficulties, such as imbalanced data, ML interpretation, and ML ethics in medical domains . In order to shed light on the current trend, approaches, and issues related to

ML in disease diagnosis, we provide a review in this paper that focuses on the novel applications of ML and DL in this field. We also provide an overview of development in this field.

We begin by outlining several methods to machine learning and deep learning techniques and particular architecture for detecting and categorizing various forms of disease diagnosis.

Machine learning algorithms have demonstrated their value as useful instruments in disease diagnosis through the examination of patterns and correlations in large datasets. Machine learning algorithms can identify anomalies and enable clinicians to identify potential disease signs even before specific symptoms manifest by training models on patient electronic health records. A delayed diagnosis may have a number of effects on patients' overall health. Accurately and promptly identifying disease symptoms is essential for effective disease management and better patient outcomes.

In addition to improving disease diagnosis, accurate identification and interpretation of disease symptoms also play a vital role in the development and implementation of precision medicine.

Disease symptoms are essential for both diagnosing and treating a wide range of illnesses. Health care professionals can provide patients with appropriate treatment and care by accurately identifying and interpreting disease symptoms. This can lead to improved patient outcomes and quality of life. It is crucial to remember that the specific illness and individual circumstances can have a significant impact on the symptoms of a disease. The onset of a disease is frequently accompanied by non-specific symptoms, which makes diagnosis difficult and frequently delayed.

The ML algorithms are generally classified into three categories such as supervised, unsupervised, and semisupervised . However, ML algorithms can be divided into several subgroups based on different learning approaches.

Algorithm and Dataset Analysis

A disease prediction dataset is a useful set of information that is used in epidemiology and healthcare to predict the likelihood, course, or occurrence of different diseases within a population. Numerous variables, including demographic data, medical history, lifestyle factors, and environmental circumstances, are often included in these datasets. These variables are essential for comprehending and forecasting illness trends.
These datasets are used by academics and medical practitioners to create data-driven insights and predictive models that help with early disease identification, prevention, and improved treatment. These databases are essential for improving public health outcomes because they enable us to better anticipate and manage health risks through the analysis of historical and current data.

	Disease	Fever	Cough	Fatigue	Difficulty Breathing	Age	Gender	\
0	Influenza	1	0	1	1	19	0	
1	Common Cold	0	1	1	0	25	0	
2	Eczema	0	1	1	0	25	0	
3	Asthma	1	1	0	1	25	1	
4	Asthma	1	1	0	1	25	1	

	Blood Pressure	Cholesterol Level	Outcome Variable
0	Low	0	1
1	0	0	0
2	0	0	0
3	0	0	1
4	0	0	1

The relevant literature states that CNN, SVM, and LR are the most often used simple algorithms in the development of MLBDD models. For example, Kalaiseldvi et al. (2020) proposed a CNN-based approach for the diagnosis of brain tumours; Dai et al. (2019) used CNN to develop a device intelligence app for the detection of skin cancer; Fathietal. (2020) used SVM to classify liver diseases; Sing et al. (2019) used SVM to classify patients with symptoms of heart disease; and Basheer et al. (2019) used Logistic Regression to diagnose heart disease.

The most popular machine learning methods for disease diagnosis are shown in Figure . The larger and bolder text highlights the significance and frequency of the algorithms used in MLBDD. We can confirm based on the figure that

Word cloud for most frequently used ML algorithms in MLBDD publications.

Because publicly accessible data sets do not require permission and provide adequate information to conduct their research, the majority of MLBDD researchers use them. However, many studies used privately collected or owned data, either because of their uniqueness based on patient set-up or to get a result with real data, as manual patient data gathering is time-consuming [46, 55, 56, 68, 70]. The most often used datasets in disease diagnosis areas are the Parkinson, PIMA, and Cleveland heart disease datasets. Table 11 enumerates publicly accessible datasets and resources that could prove beneficial for upcoming scholars and practitioners.

CSV FILE:

	A	B	C	D	E	F	G	H	I	J	K
1	Disease	Fever	Cough	Fatigue	Difficulty E	Age	Gender	Blood Pres	Cholesterc	Outcome	Variable
2	Influenza	Yes	No	Yes	Yes	19	Female	Low	Normal	Positive	
3	Common (	No	Yes	Yes	No	25	Female	Normal	Normal	Negative	
4	Eczema	No	Yes	Yes	No	25	Female	Normal	Normal	Negative	
5	Asthma	Yes	Yes	No	Yes	25	Male	Normal	Normal	Positive	
6	Asthma	Yes	Yes	No	Yes	25	Male	Normal	Normal	Positive	
7	Eczema	Yes	No	No	No	25	Female	Normal	Normal	Positive	
8	Influenza	Yes	Yes	Yes	Yes	25	Female	Normal	Normal	Positive	
9	Influenza	Yes	Yes	Yes	Yes	25	Female	Normal	Normal	Positive	
10	Hyperthyr	No	Yes	No	No	28	Female	Normal	Normal	Negative	
11	Hyperthyr	No	Yes	No	No	28	Female	Normal	Normal	Negative	
12	Asthma	Yes	No	No	Yes	28	Male	High	Normal	Positive	
13	Allergic Rh	No	Yes	Yes	No	29	Female	Normal	Low	Negative	
14	Anxiety Di	No	Yes	No	No	29	Female	Normal	High	Negative	
15	Common (	No	No	No	No	29	Female	Low	Normal	Negative	
16	Diabetes	No	No	No	No	29	Male	Low	Normal	Negative	
17	Gastroent	No	Yes	No	No	29	Female	Normal	Normal	Negative	
18	Pancreatit	Yes	No	No	No	29	Female	High	Normal	Negative	

	A	B	C	D	E	F	G	H	I	J
18	Pancreatit	Yes	No	No	No	29	Female	High	Normal	Negative
19	Rheumatc	No	Yes	Yes	Yes	29	Female	High	High	Negative
20	Depressio	Yes	Yes	Yes	Yes	29	Male	High	Normal	Positive
21	Liver Canc	Yes	Yes	Yes	Yes	29	Female	Normal	Normal	Positive
22	Stroke	Yes	Yes	Yes	Yes	29	Female	Normal	Normal	Positive
23	Urinary Tr	Yes	Yes	Yes	No	29	Male	High	High	Positive
24	Dengue Fe	Yes	No	Yes	No	30	Female	Normal	Normal	Negative
25	Dengue Fe	Yes	No	Yes	No	30	Female	Normal	Normal	Negative
26	Eczema	No	Yes	Yes	No	30	Male	High	High	Negative
27	Gastroent	Yes	Yes	Yes	No	30	Male	High	High	Negative
28	Hepatitis	Yes	Yes	Yes	Yes	30	Male	High	Normal	Negative
29	Kidney Ca	No	No	Yes	No	30	Male	Normal	Normal	Negative
30	Migraine	Yes	No	No	No	30	Female	Normal	Normal	Negative
31	Migraine	No	Yes	Yes	No	30	Female	Normal	Normal	Negative
32	Muscular	No	No	Yes	No	30	Male	High	High	Negative
33	Sinusitis	No	Yes	Yes	No	30	Male	Normal	Normal	Negative
34	Ulcerative	Yes	Yes	No	No	30	Female	Normal	Normal	Negative
35	Ulcerative	No	Yes	Yes	No	30	Female	Normal	Normal	Negative

Methodology:

Machine Learning Algorithms

Computational techniques known as machine learning algorithms allow computers to learn and make decisions without needing to be explicitly programmed. Artificial Intelligence (AI) is largely dependent on these algorithms, which are widely used in a variety of fields such as healthcare, finance, image recognition, natural language processing, and many more. An overview of some of the most popular machine learning algorithms can be found here:

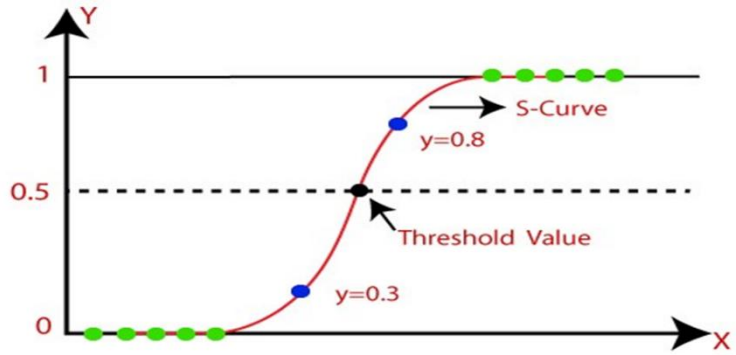
1. SUPERVISED LEARNING ALGORITHMS:

Machine learning algorithms frequently involve intricate mathematical formulas, and it is impractical to provide all of the formulas for any algorithm in a concise manner. To assist you in understanding the concepts, I can provide you with a brief overview of some fundamental machine learning methods and simplified equations.

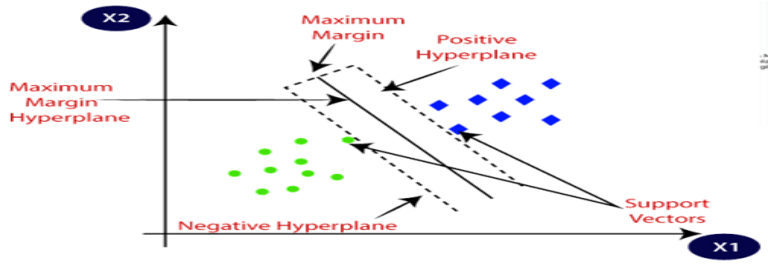
A popular statistical modelling technique called linear regression seeks to establish a linear

2.LOGISTIC REGRESSION

The purpose of the powerful statistical method known as logistic regression is to predict the probability of a binary outcome, usually represented as 'y' being equal to 1, based on one or more independent variables 'x'. It is mostly used to solve binary classification problems. " The logistic function, which converts a linear combination of these variables, represented as "mx + b," into a probability, is the source of logistic regression. In this formula,'m' denotes the coefficients, which represent the weight and significance of each independent variable, and 'b' is the intercept term. The logistic function, denoted as 1 / (1 + e^-(mx + b)), represents the likelihood that 'y' will equal 1 given the values, by mapping the linear combination into a probability value that ranges from 0 to 1.



3.SUPPORT VECTOR MACHINE



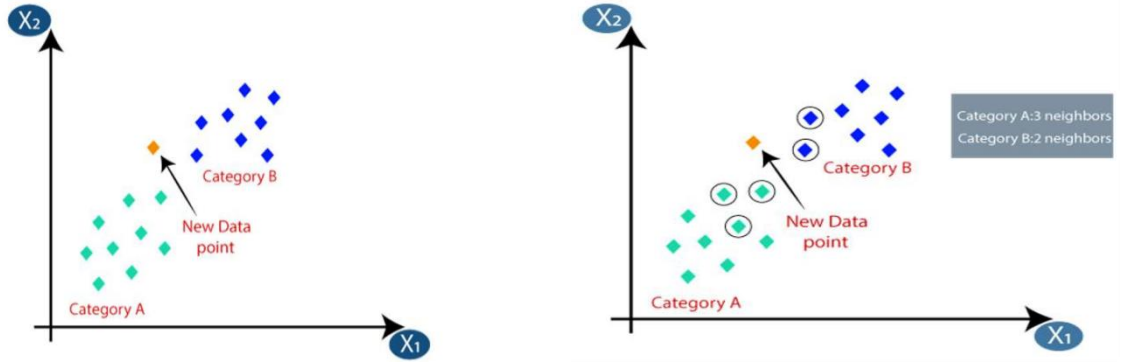
The decision boundary equation is more complex since it depends on the kernel that is used (such as linear, polynomial, or radial basis function). The equation is similar to a linear combination of support vectors for linear SVM.

4. DECISION TREE:

Decision trees use a tree-like structure to make decisions. There isn't a single equation, but the process involves recursive decision-making based on feature values.

5. K-NEARSET NEIGHBOUR (K-NN):

Not represented by a specific equation, but it works by finding the k-nearest data points in the training set and classifying the test point based on the majority class among its neighbors.



The K-NN working can be explained on the basis of the below algorithm:
Step-1: Select the number K of the neighbors
Step-2: Calculate the Euclidean distance of K number of neighbors
Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
Step-4: Among these k neighbors, count the number of the data points in each category.
Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready

6. NEURAL NETWORKS:

Neural networks consist of multiple interconnected layers with weighted connections. The forward pass involves a series of matrix multiplications and activation functions at each layer.

7. RANDOM FOREST:

Random forests use multiple decision trees and aggregate their outputs. The final output is a combination of the outputs from individual trees. +

EVALUATIONS:

This section describes the performance measures used in reference literature. Performance indicators, including accuracy, precision, recall, and F1 score, are widely employed in disease diagnosis. For example, lung cancer can be categorized as true positive (TP) or true-negative (TN) if individuals are diagnosed correctly, while it can be categorized as false positive (FP) or false negative (FN) if misdiagnosed. The most widely used metrics are described below

1. Predictive accuracy:

TP+TN/TP+FP+FN+TN

2. Sensitivity:

TP/TP+FN

3. Specificity:

TN/TN+FP

4. Positive predictive value:

TP/TP+FP

5. Negative predictive value:

TN/TN+FN

6. F1score:

2×(Senstivity×Positivepredictivevalue)/(Senstivity+Positivepredictivevalue)

8.PRECEPTRON:

A perceptron is a basic machine learning model used for binary classification tasks. You can think of it as a simplified version of a neural network. It has four main components: input values, weights, bias, and an activation function.

1. Input Values: These are the data points you want to classify. In a sense, they represent the information you're working with.

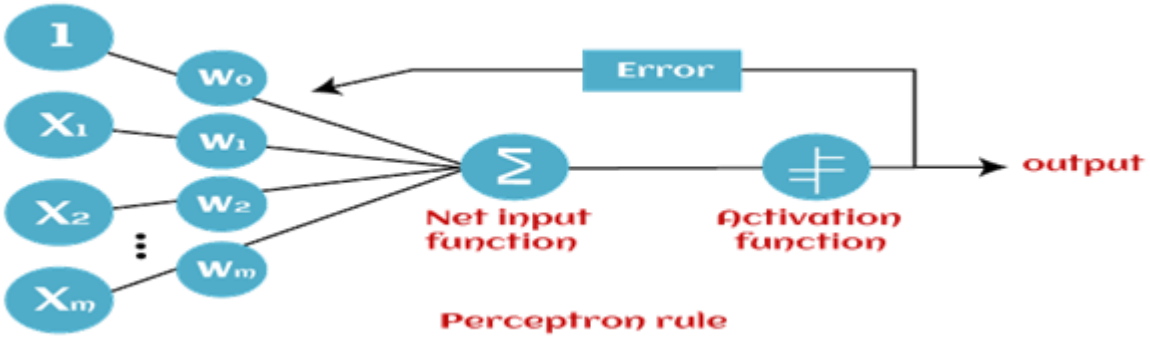
2. Weights: Each input value is associated with a weight, which represents the importance of that input in making a decision. The model multiplies each input by its corresponding weight.

3. Bias: The bias is like an extra input that's always set to 1. It allows the model to make adjustments, even when all the other inputs are zero.

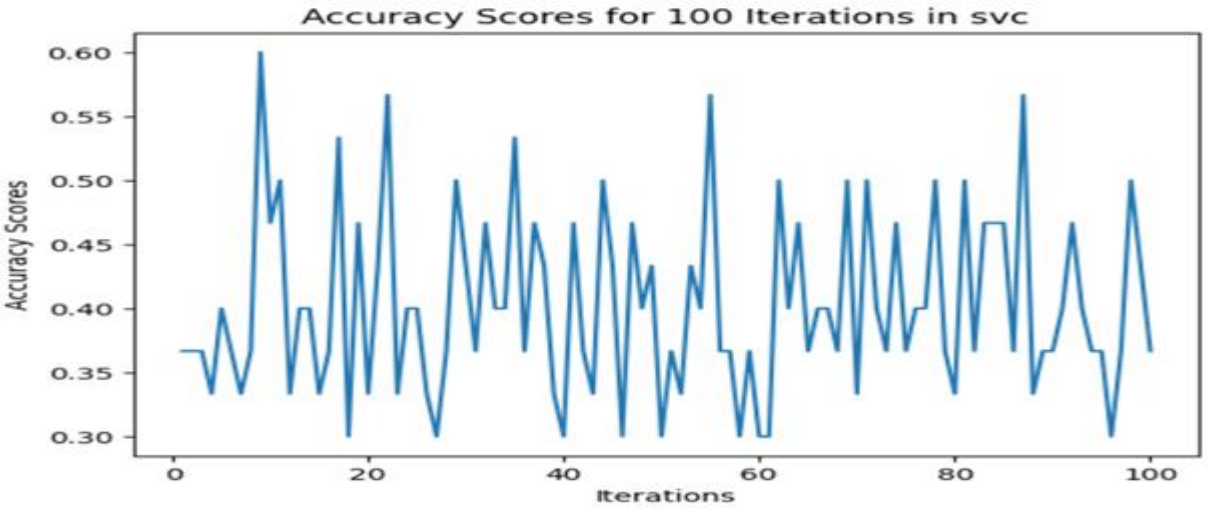
4. Net Sum: The model multiplies the input values by their weights and adds them

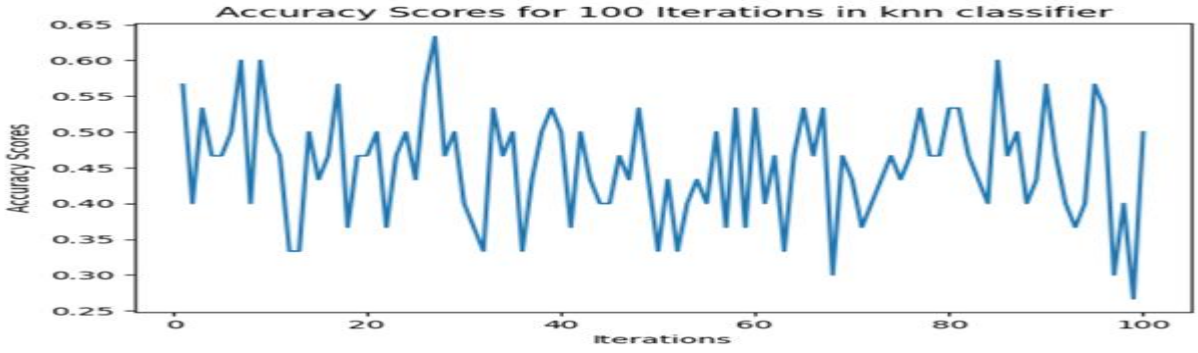
up. This results in a weighted sum, which is a measure of the overall input's impact.

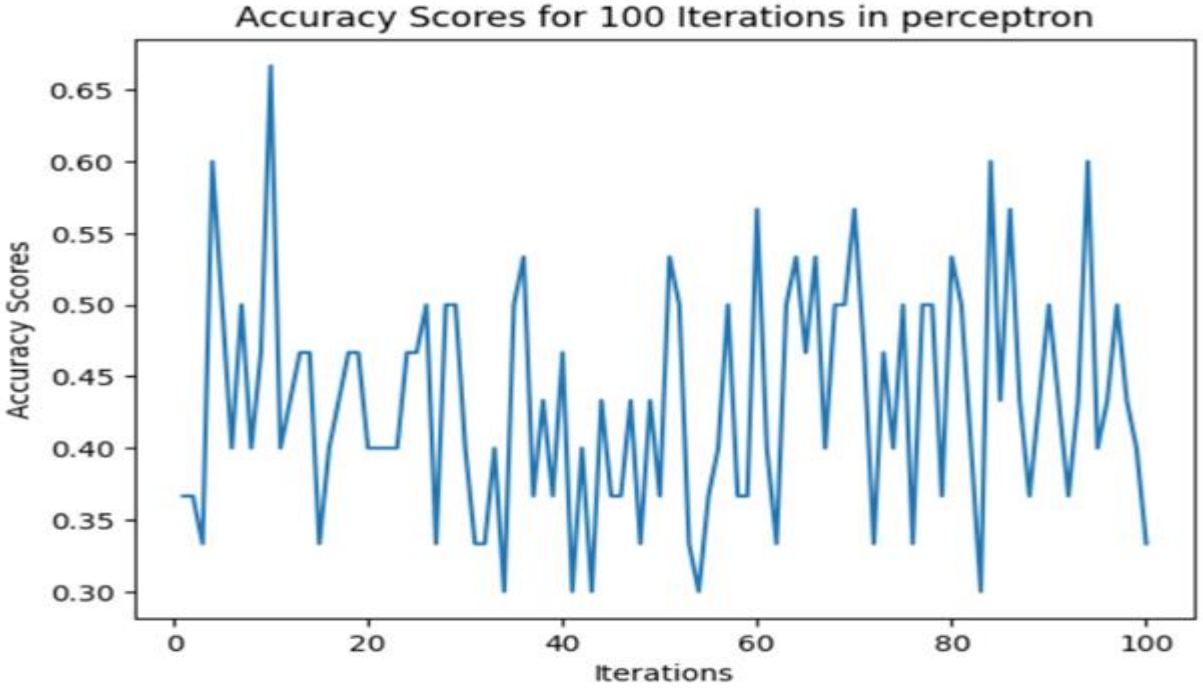
5. Activation Function: The weighted sum is then passed through an activation function. This function determines the output of the perceptron. In the case of the perceptron, the activation function is often a step function, which makes a binary decision .e.g., if the weighted sum is greater than a certain threshold, output 1; otherwise, output



9.BOOTSTRAP:

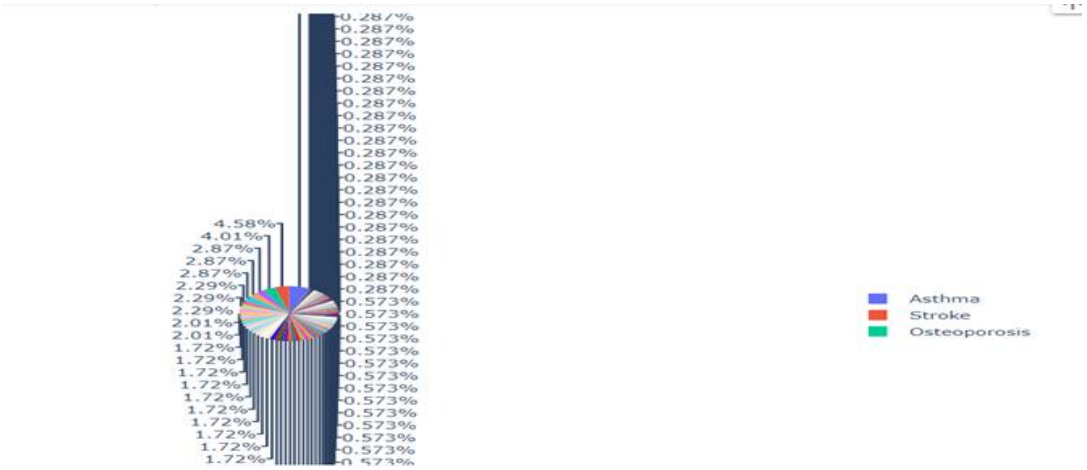


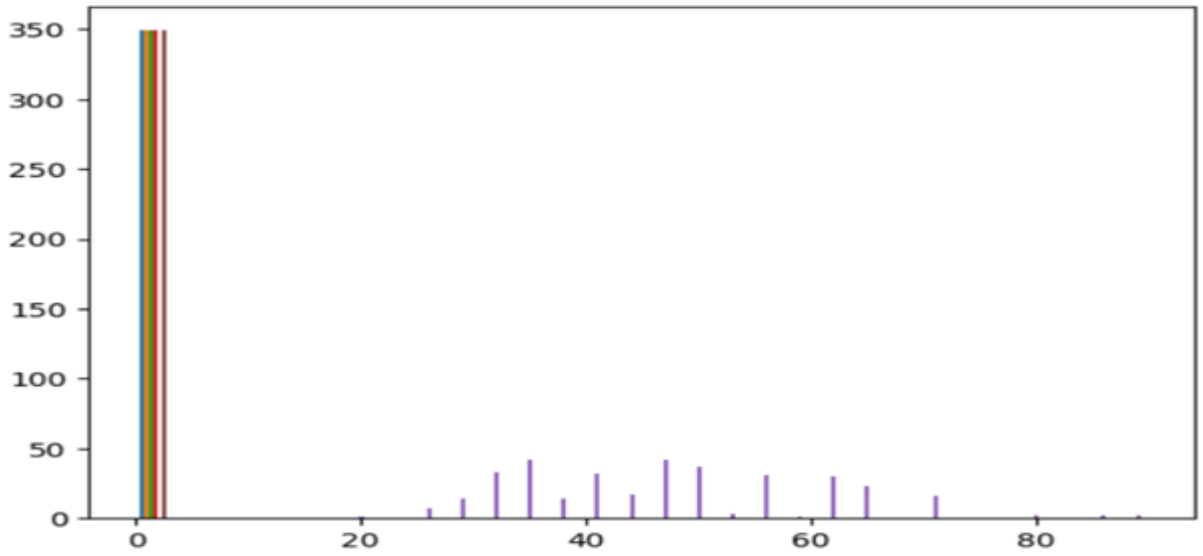


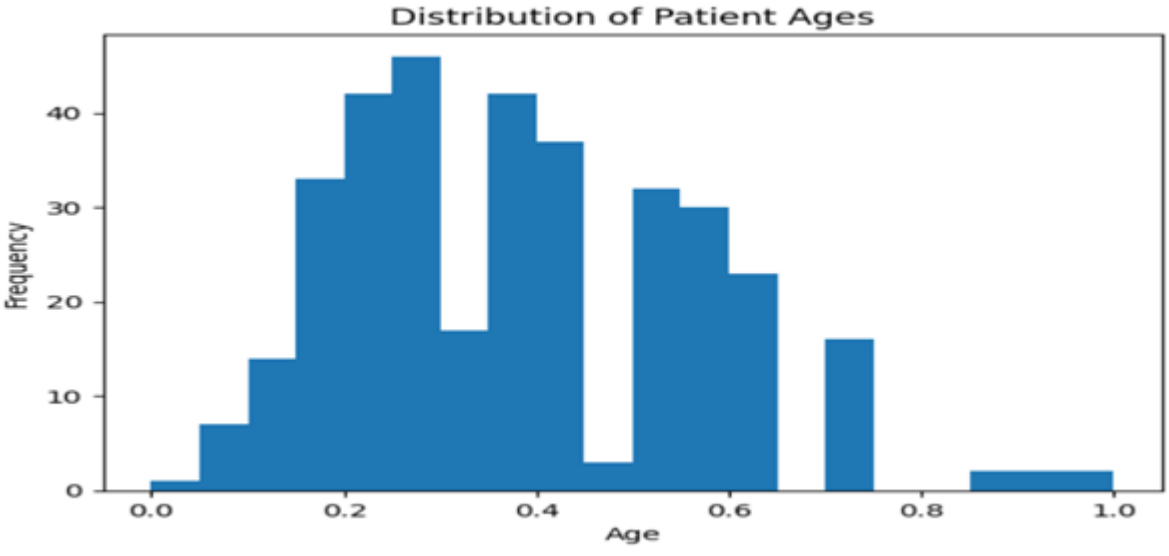

Accuracy Scores for 100 Iterations in perceptron
Accuracy Scores
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0
20
40
60
80
100
Iterations

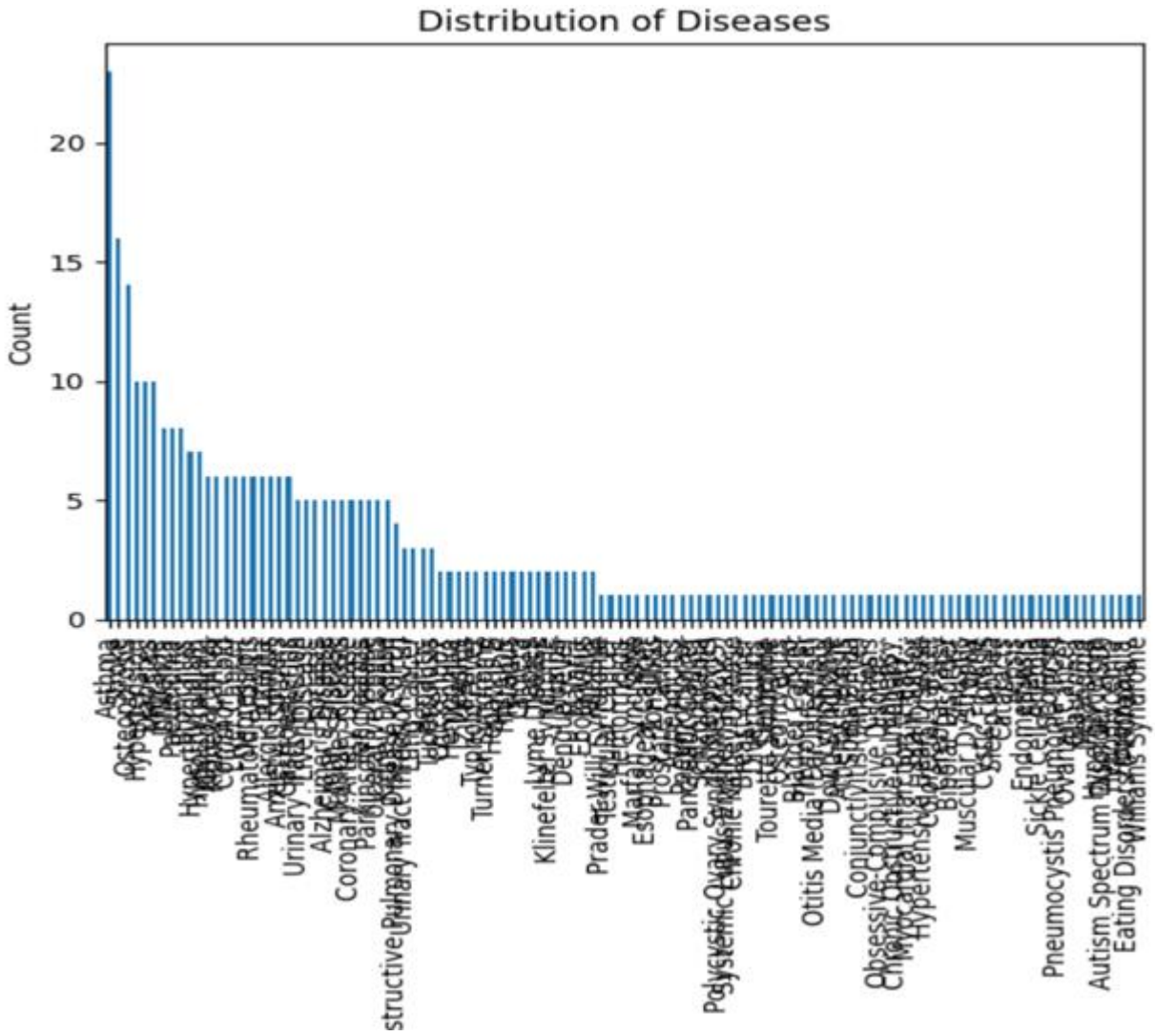

Accuracy Scores for 100 Iterations in logistic regression
Accuracy Scores
0.55
0.50
0.45
0.40
0.35
0.30
0
20
40
60
80
100
Iterations

10.GRAPHS:


Asthma
Stroke
Osteoporosis


350
300
250
200
150
100
50
0
0
20
40
60
80


Distribution of Patient Ages
Frequency
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1.0
Age



Perceptron Model:

95.0 confidence interval 30.0 and 60.0

Mean Accuracy => 0.43

Standard Deviation => 0.08

Logistic Regression:

95.0 confidence interval 28.3 and 53.5

Mean Accuracy => 0.40

Standard Deviation => 0.07

SVM Classifier:

95.0 confidence interval 30.0 and 56.7

Mean Accuracy => 0.40

Standard Deviation => 0.07

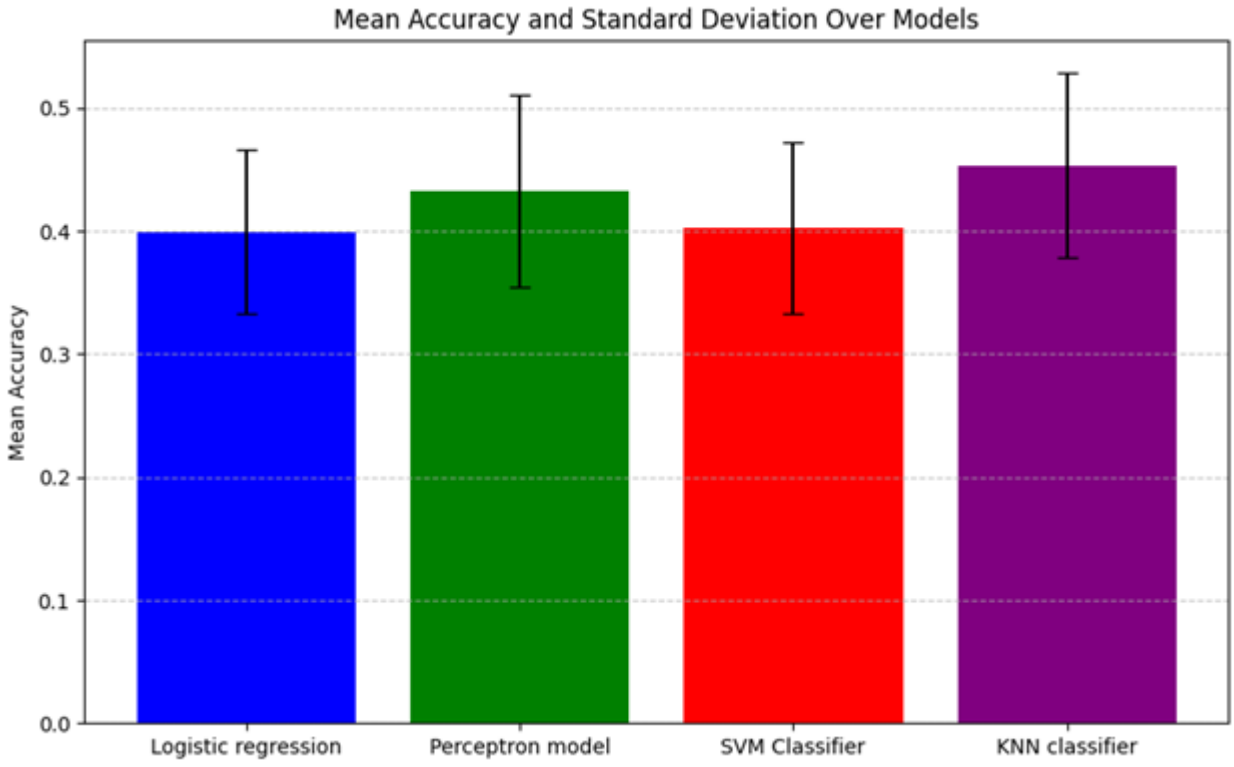
KNN Classifier:

95.0 confidence interval 31.6 and 60.0

Mean Accuracy => 0.45

Standard Deviation => 0.07

DISCUSSION:



The annotated literature has strengthened in this study as machine learning (ML) and deep learning (DL) have become more prominent in disease detection over the past ten years. The review commenced with targeted research inquiries and endeavoured to address them through the reference literature. The whole research indicates that CNN is one of the most advanced algorithms, surpassing all other machine learning algorithms because of its strong performance with both image and tabular data.

A crucial use of data science and health care technology is disease symptom prediction, which has the potential to enhance early detection and treatment of a variety of medical illnesses. Here, we will succinctly go over the meaning and context of disease symptom prediction.

Early Detection and Prevention:

Disease symptom prediction uses a variety of data sources, including patient reports, wearable devices, and electronic health records, to analyse patterns and forecast the likelihood of specific diseases. For many illnesses, early detection is essential because it can result in timely intervention and better treatment outcomes.

Data-Driven Approach:

This approach relies on the vast amount of healthcare data available today, including historical patient records, genetics, and environmental factors. Advanced machine learning and artificial intelligence techniques can process this data to identify hidden patterns, correlations, and risk factors.

Personalized Medicine:

Disease symptom prediction enables a shift towards personalized medicine. By analyzing an individual's health data, including genetic predispositions and lifestyle factors, healthcare providers can tailor treatment plans to the unique needs of each patient.

Reducing Healthcare Costs:

Early diagnosis and intervention can help reduce healthcare costs associated with prolonged treatments, hospitalizations, and the management of advanced-stage diseases. Predictive models can help allocate resources more efficiently and focus on preventive care.

Research Challenges

Although machine learning-based applications have been widely used in disease detection, researchers and practitioners continue to face several obstacles when using them in the real world of healthcare. The following is a summary of the key challenges associated with machine learning in disease diagnosis in this section:

Data Related Challenges

• 1.Data scarcity: Despite the fact that numerous hospitals and health care providers have recorded patient data, real-world data is frequently unavailable for international research purposes because of the Data Privacy Act.

2.Algorithm Related Challenges

• 1. Supervised vs unsupervised: With the labelled data, the majority of ML models (logistic regression and linear regression) performed admirably. However, the performance of similar algorithms was much worse when using the unlabeled data. However, well-known algorithms that exhibit good performance with unlabeled data—like K-means clustering, support vector machines, and kernel density networks—also underperformed when dealing with multidimensional data.

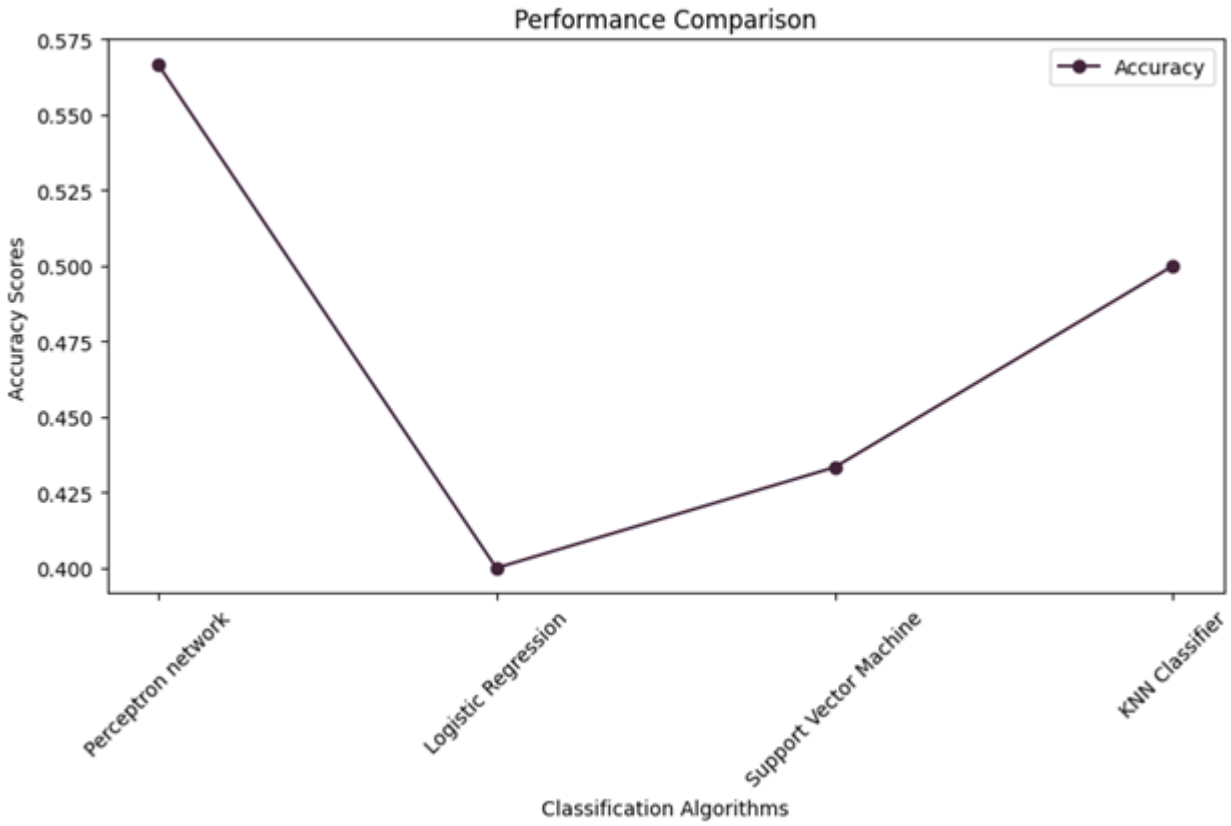
Results:

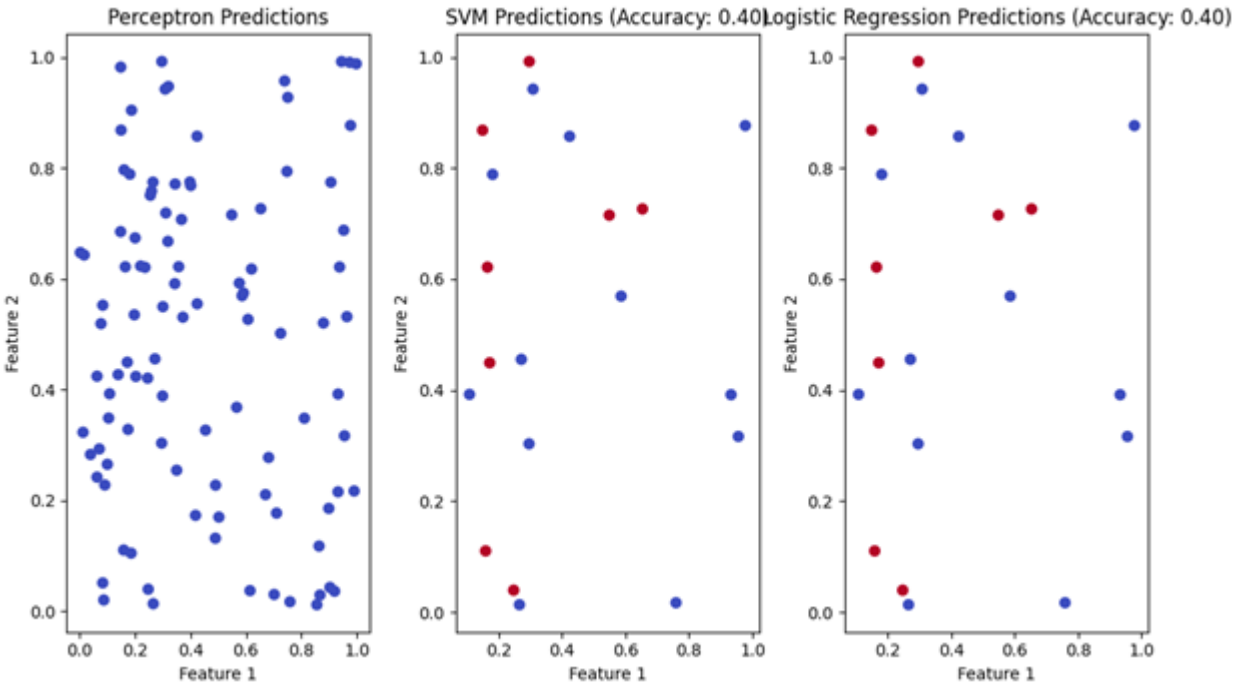
Accuracy: 0.5571428571428572

Precision: 0.5673469387755102

Recall: 0.5571428571428572

F1-score: 0.5594139194139195





Conclusions :

Most Fluctuating Model: Perceptron model

Most Consistent Model: Logistic regression

In summary, machine learning-based disease symptom prediction presents a huge opportunity for the healthcare sector. Through the analysis of extensive datasets containing patient data and medical records, machine learning algorithms can significantly aid in the timely and accurate diagnosis of a variety of diseases. This technology has the ability to revolutionise healthcare by improving patient outcomes, lowering costs, and maximising the effectiveness of medical professionals.

Nonetheless, it is imperative to recognise the difficulties and constraints linked to this methodology. The accuracy of forecasts is highly dependent on the calibre and volume of available data. Patient data privacy concerns and ethical considerations need to be carefully addressed. Moreover, in order to ensure their reliability, machine learning models need undergo ongoing validation and updates.

As a result, even if machine learning is a powerful tool for disease symptom prediction, it should be carefully integrated with other clinical practises and not be seen as a replacement for healthcare professionals. Applying it carefully and responsibly can have a big impact on early detection and personalised treatment, which will ultimately improve overall health care quality.

IMPLEMENTATION OF CODE: