# Multiple Disease Prediction using MachineLearning Algorithms

BARRI SAI KUMAR [1], KILLAMSETTI JABILI [2],KARANAM KURMA RAO [3] ,
RAYALA PRANAVI[4] ·SMT.J.SWATHI [5]

[1,2,3,4]*B.Tech UG Students, Dept. Of ECE, Aditya Institute Of Technology And Management,Tekkali,AP,India.*

**5Guide, Assistant Professor, Dept. Of ECE, Aditya Institute Of Technology And Management, Tekkali, AP, India*

## ABSTRACT

*Now a days for small problems, the users have to go personally to the hospital for check-up which is more time consuming. Also handling the telephonic calls for appointments is quite hectic. Such a problem can be solved by using disease prediction applications by giving proper guidance regarding healthy living. Over the past decade, the use of the specific disease prediction tools along with the concerning health has been increased due to a variety of diseases and less doctor-patient ratio. Thus, in this system, we are concentrating on providing immediate and accurate disease prediction to the users about the symptoms they enter along with the severity of disease predicted. For prediction of diseases, different machine learning algorithms are used to ensure quick and accurate predictions. In one channel, the symptoms entered will be cross checked with the database. Further, it will be preserved in the database if the symptom is new which its primary work is and the other channel will provide severity of disease predicted.*

**Keywords**: *Machine Learning, Support vector machine, Random Forest Algorithm,Naive Bayes Algorithm.*

## 1. INTRODUCTION

At present, when one suffers from a particular disease, then the person has to visit a doctor which is time consuming and costly too. Also if the user is out of reach of doctors and hospitals it may be difficult for the user as the disease can not be identified. So, if the above process can be completed using an automated program which can save time as well as money, it could be easier to the patient which can make the process easier.The Disease Prediction system has data sets collected from different health related sites. With the help of Disease Predictor the user will be able to know the probability of the disease with the given symptoms.As the use of the internet is growing every day, people are always curious to know different new things. People always try to refer to the internet if any problem arises. People have access to the internet more than hospitals and doctors. People do not have immediate options when they suffer with a particular disease. So, this system can be helpful to the people as they have access to the internet 24 hours.

## 2. LITERATURE SURVEY

The study for the best medical diagnosis mining technique was performed by K.M. Al-Aidaroos, A.A. Bakar, and Z. Othman. For this study, the authors compared Naive Bayes to five other classifiers: LR, KStar (K*), Decision Tree (DT), Neural Network (NN), and a basic rule-based algorithm (ZeroR). The efficiency of all algorithms was evaluated using 15 real-world medical problems from the UCI machine learning repository (Asuncion and Newman, 2007). In

the experiment, NB outperformed the other algorithms in 8 of the 15 data sets, leading to the conclusion that the predictive accuracy results in Naive Bayes are superior to other techniques. Darcy A. Davis, Nitesh V. Chawla, Nicholas Blumm, Nicholas Christakis, and Albert-Laszlo Barabasi discovered that treating chronic illness at a global level is neither time nor cost effective. As a result, the authors performed this study in order to forecast potential disease risk. CARE (which uses only a patient's medical history and ICD-9-CM codes to predict possible disease risks) was used for this. Based on their own medical history and that of similar patients, CARE incorporates collective filtering approaches with clustering to predict each patient's greatest disease risks. ICARE, an iterative version that integrates ensemble principles for improved efficiency, has also been defined by the authors.

These cutting-edge systems don't need any advanced knowledge and can predict a wide range of medical conditions in a single run. ICARE's remarkable potential risk coverage means more precise early alerts for thousands of illnesses, several years ahead of time. When used to its full extent, the CARE system can be used to investigate a wider range of disease backgrounds, raise previously unconsidered questions, and facilitate discussions regarding early detection and prevention.

This research paper was written by JyotiSoni, Ujma Ansari, Dipesh Sharma, and SunitaSoni to provide a survey of existing techniques of information discovery in databases using data mining techniques that are used in today's medical research, specifically in Heart Disease Prediction. A number of experiments have been carried out to compare the performance of predictive data mining.

## 3. DATA SET

The dataset we have considered consists of 132 symptoms, the combination or permutations of which leads to 41 diseases. Based on the 4920 records of patients, we aim to develop a prediction model that takes in the symptoms from the user and predicts the disease he is more likely to have.

Diseases we considered are :

**TABLE I. DISEASES**

| DISEASES | | |
|---|---|---|
| FUNGAL INFECTION | MALARIA | VARICOSE VEINS |
| ALLERGY | CHICKENPOX | HYPOTHYROIDISM |
| GERD | DENGUE | VERTIGO |
| CHRONIC CHOLESTASIS | PEPTIC ULCER DISEASE | ACNE |
| DRUG REACTION | HEPATITIS A | URINARY TRACT INFECTION |
| PILES | HEPATITIS B | PSORIASIS |
| AIDS | HEPATITIS C | IMPETIGO |
| DIABETES | HEPATITIS D | HYPERTHYROIDISM |
| GASTROENTERITIS | HEPATITIS E | HYPOGLYCEMIA |
| BRONCHIAL ASTHMA | ALCOHOLICHEPATITIS | CERVICAL SPONDYLOSIS |
| HYPERTENSION | TUBERCULOSIS | ARTHRITIS |
| MIGRAINE | COMMON COLD | OSTEOARTHRITIS |

| | | |
|---|---|---|
| PARALYSIS | PNEUMONIA | TYPHOID |

The considered symptoms are:

**TABLE II. SYMPTOMS**

| Symptoms | | |
|---|---|---|
| Back Pain | Bloody Stool | Scurrying |
| Constipation | Depression | Passage Of Gases |
| Abdominal Pain | Irritation In Anus | Weakness In Limbs |
| Diarrhea | Neck Pain | Fast Heart Rate |
| Mild Fever | Dizziness | Internal Itching |
| Yellow Urine | Cramps | Toxic Look |
| Yellowing Of Eyes | Bruising | Palpitations |
| Acute Liver Failure | Obesity | Painful Walking |
| Fluid Overload | Swollen Legs | Prominent Veins On Calf |
| Swelling Of Stomach | Irritability | Fluid Overload |
| Swelled Lymph Nodes | Swollen Blood Vessels | Excessive Hunger |
| Malaise | Muscle Pain | Black Heads |
| Blurred And Distorted Vision | Pain In Anal Region | Pain During Bowel Movements |
| Phlegm | Brittle Nails | Rusty Sputum |
| Throat Irritation | Belly Pain | Mucoid Sputum |
| Redness Of Eyes | Enlarged Thyroid | Puffy Face And Eyes |
| Sinus Pressure | Slurred Speech | Hip Joint Pain |
| Runny Nose | Knee Pain | Polyuria |
| Congestion | Skin Peeling | Family History |
| Chest Pain | Extra Marital Contacts | Swollen Extremities |

| | | |
|---|---|---|
| Yellow crust ooze | Swelling joints | Coma |
| Loss of smell | Stiff neck | Unsteadiness |
| Movementstiffness | Muscle weakness | Drying and tinglinglips |
| Spinning movements | Red sore aroundnose | Weakness of one body side |
| Bladder discomfort | Foul smell ofurine | Continuous feel ofurine |
| Altered sensorium | Red spots overbody | Abnormal menstruation |
| Dyschromicpatches | Watering fromeyes | Increases appetite |
| Lack of concentration | Visual disturbances | Receiving blood transfusion |
| Receiving unsterile injections | Distention ofabdomen | History of alcohol consumption |

| | | |
|---|---|---|
| Puss filledpimples | Blood in sputum | Stomach bleeding |
| Silver likedusting | Small dents innails | Inflammatory nails |

## 4. PROPOSED WORK

**4.1** The flow chat mentioned below explains the work flow of our project

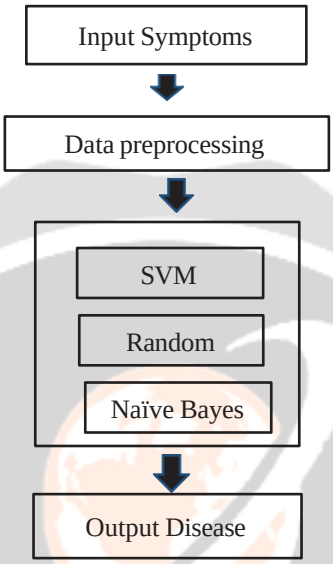



**Fig. 1**: Flow of the model

### 4.1.1 Input (Symptoms):

While designing the model we have assumed that user has a clear idea about the symptoms he is experiencing. The Prediction developed considers 95 symptoms amidst which the user can give the symptoms his processing as the input.

### 4.1.2 Data preprocessing:

The data mining technique that transforms the raw data or encodes the data to a form which can be easily interpreted by the algorithm is called data preprocessing. The preprocessing techniques used in the presented work are:

- **Data Cleaning:** Data is cleansed through processes such as filling in missing value, thus resolving the inconsistencies in the data.
- **Data Reduction:** The analysis becomes hard when dealing with huge database. Hence, we eliminate those independent variables(symptoms) which might have less or no impacton the target variable(disease). In the present work, 95 of 132symptoms closely related to the diseases are selected**.**

### 4.1.3 Models selected:

The system is trained to predict the diseases using three algorithms

- Support vector machine
- Random forest Classifier
- Naïve Bayes Classifier

A comparative study is presented at the end of work, thus analyzing the performance of each algorithm of the considered database.

#### 4.1.4 .Output(diseases):

Once the system is trained with the training set using the mentioned algorithms a rule set is formed and when the user the symptoms are given as an input to the model, those symptoms are processed predicting the most likely disease.

## 5. METHODOLOGY

The disease prediction system is implemented using the three data mining algorithms i.e. Support vector classifier, Random forest classifier and Naïve Bayes classifier. The description and working of the algorithms are given below.

### 5.1 Support vector machine:

The aim of a support vector machine algorithm is to find the best possible line, or *decision boundary*, that separates the data points of different data classes. This boundary is called a *hyperplane* when working in high-dimensional feature spaces. The idea is to maximize the margin, which is the distance between the hyperplane and the closest data points of each category, thus making it easy to distinguish data classes.

SVMs are useful for analyzing complex data that can't be separated by a simple straight line. Called *nonlinear SMVs*, they do this by using a mathematical trick that transforms data into higher- dimensional space, where it is easier to find a boundary.

**Types of support vector machines**

Support vector machines have different types and variants that provide specific functionalities and address specific problem scenarios. Here are two types of SVMs and their significance:

1. **Linear SVM.** Linear SVMs use a linear kernel to create a straight-line decision boundary that separates different classes. They are effective when the data is linearly separable or when a linear approximation is sufficient. Linear SVMs are computationally efficient and have good interpretability, as the decision boundary is a hyperplane in the input feature space.
2. **Nonlinear SVM.** Nonlinear SVMs address scenarios where the data cannot be separated by a straight line in the input feature space. They achieve this by using kernel functions that implicitly map the data into a higher-dimensional feature space, where a linear decision boundary can be found. Popular kernel functions used in this type of SVM include the polynomial kernel, Gaussian (RBF) kernel and sigmoid kernel. Nonlinear SVMs can capture complex patterns and achieve higher classification accuracy when compared to linear SVMs.

### 5.2 Random Forest Classifier:

Random forest is a flexible, easy to use machine learningalgorithm that provides exceptional results most of the time even without hyper-tuning. As mentioned in the Decision tree, the major limitation of decision tree algorithm is overfitting**.** It appears as if the tree has memorized the data.

Random Forest prevents this problem: It is a version of ensemble learning. Ensemble learning refers to usingmultiple algorithms or same algorithm multiple times. Random forest is a team of Decision trees. And greater the number of these decision trees in Random forest, the better the generalization.

- More precisely, Random forest works as follows:
- Selects k symptoms from dataset (medical record) with a total of m symptoms randomly (where k<<m). Then, it builds a decision tree from those k symptoms.
- Repeats **n** times so that we have **n** decision trees built from different random combinations of k

symptoms (or a different random sample of the data, called *bootstrap sample*)

- Takes each of the **n**-built decision trees and passes a random variable to predict the Disease. Stores the predicted Disease, so that we have a total of **n** Diseases predicted from **n** Decision trees.
- Calculates the votes for each predicted Disease and takes the mode (most frequent Disease predicted) as the final prediction from the random forest algorithm.

### 5.3 Naive Bayes Classifier:

The fundamental Naïve Bayes assumption is that each feature makes an:

- Independent
- Equal

Contribution to the outcome. Its advantage is that it works fast even on a large dataset as it requires less computational power.

Bayes theorem
Naïve Byes algorithm is based on Bayes theorem given by: P(s/h)=
(P(h/s)P(s) )/P(h)
Where

P(s/h) = posterior probability P(h/s) = likelihood
P(s) = Class Prior Probability P(h) = Predictor Prior Probability

In the formula above 's' denotes class and 'h' denotes features. In *P(h)*, the denominator consists the only term that is a function of data(features)- it is not a function of the class we are currently dealing with. Thus, it will be same for all the classes. Traditionally in naïve Bayes Classification, we ignore this denominator as it does not affect the result of the classifier in order to make the prediction.

Key Terms:

- *Prior probability* is the proportion of Disease in the considered data set.
- *Likelihood* is the probability of classification a disease in presence of some other symptoms.
- *Marginal Likelihood* is the proportion of symptoms in the considered dataset.

### 5.4 Column Distribution:

Column distribution shows you the distribution of the data within the column and the counts of distinct and unique values. Distinct values are all values in a column, including duplicates and null values while unique values do not include duplicates or nulls.

### 5.5 Confusion Matrix:

The confusion matrix helps assess classification model performance in machine learning by comparing predicted values against actual values for a dataset. A confusion matrix (or, error matrix) is a visualization method for classifier algorithm results.

## 6. RESULTS

**TABLE 6.1 CLASSIFICATION REPORT**

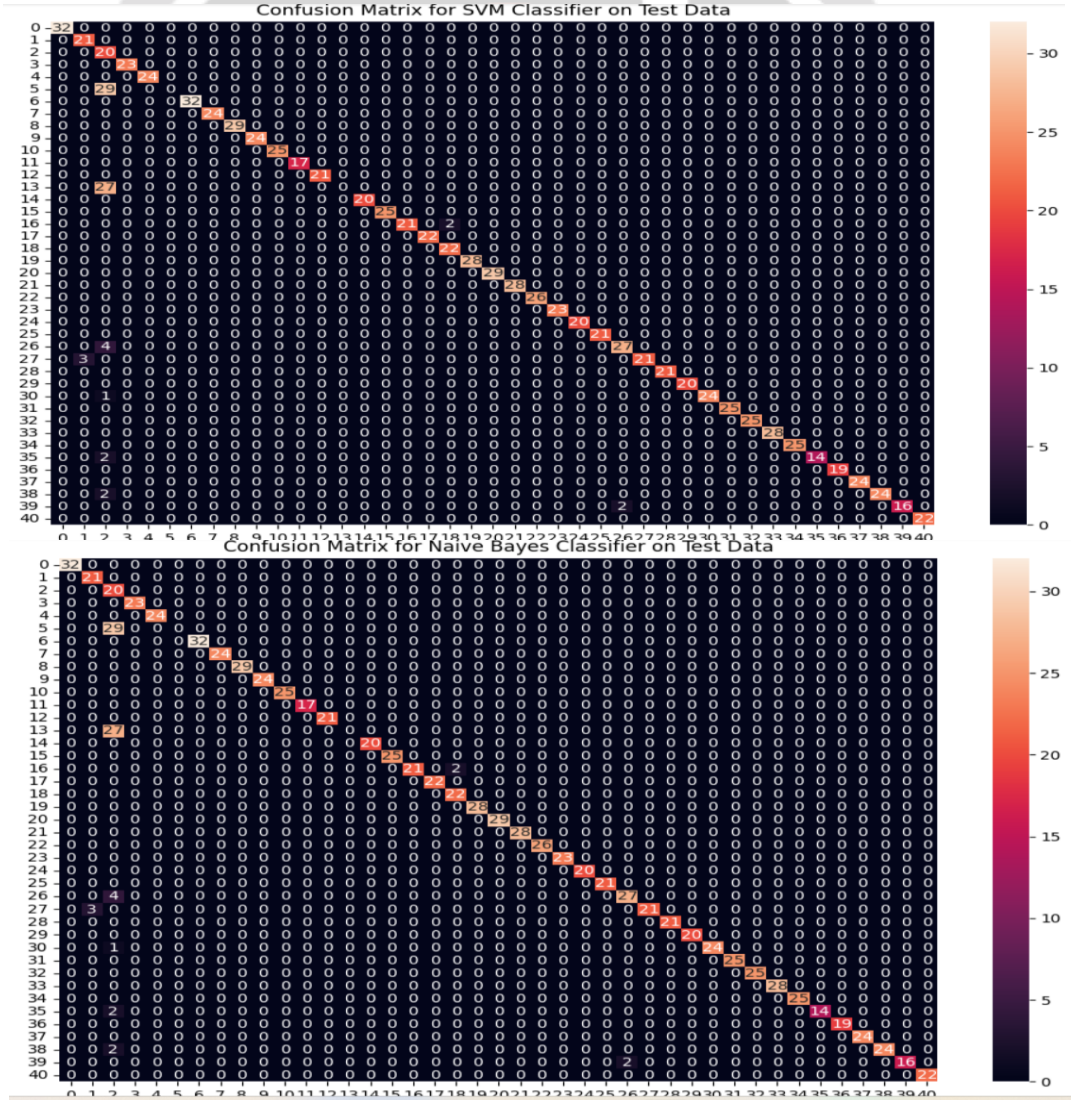
| Prognosis | precision | recall | f1-score | support |
|---|---|---|---|---|
| (vertigo) Paroymsal Positional Vertigo | 1.00 | 1.00 | 1.00 | 18 |
| AIDS | 1.00 | 1.00 | 1.00 | 30 |
| Acne | 0.00 | 0.00 | 0.00 | 24 |
| Alcoholic hepatitis | 1.00 | 1.00 | 1.00 | 25 |
| Allergy | 1.00 | 1.00 | 1.00 | 24 |
| Arthritis | 0.28 | 1.00 | 0.44 | 23 |
| Bronchial Asthma | 1.00 | 1.00 | 1.00 | 33 |
| Cervical spondylosis | 1.00 | 1.00 | 1.00 | 23 |
| Chicken pox | 1.00 | 1.00 | 1.00 | 21 |
| Chronic cholestasis | 1.00 | 1.00 | 1.00 | 15 |
| Common Cold | 1.00 | 1.00 | 1.00 | 23 |
| Dengue | 1.00 | 1.00 | 1.00 | 26 |
| Diabetes | 1.00 | 1.00 | 1.00 | 21 |
| Dimorphic hemmorhoids(piles) | 0.00 | 0.00 | 0.00 | 29 |
| Drug Reaction | 1.00 | 1.00 | 1.00 | 24 |
| Fungal infection | 1.00 | 1.00 | 1.00 | 19 |
| GERD | 1.00 | 0.93 | 0.96 | 28 |
| Gastroenteritis | 1.00 | 1.00 | 1.00 | 25 |
| Heart attack | 0.92 | 1.00 | 0.96 | 23 |
| Hepatitis B | 1.00 | 1.00 | 1.00 | 27 |
| Hepatitis C | 1.00 | 1.00 | 1.00 | 26 |
| Hepatitis D | 1.00 | 1.00 | 1.00 | 23 |
| Hepatitis E | 1.00 | 1.00 | 1.00 | 29 |
| Hypertension | 1.00 | 1.00 | 1.00 | 25 |
| Hyperthyroidism | 1.00 | 1.00 | 1.00 | 24 |
| Hypoglycemia | 1.00 | 1.00 | 1.00 | 26 |
| Hypothyroidism | 0.95 | 0.90 | 0.93 | 21 |
| Impetigo | 1.00 | 1.00 | 1.00 | 24 |
| Jaundice | 1.00 | 1.00 | 1.00 | 19 |
| Malaria | 1.00 | 1.00 | 1.00 | 22 |
| Migraine | 1.00 | 1.00 | 1.00 | 25 |
| Osteoarthristis | 1.00 | 1.00 | 1.00 | 22 |
| Paralysis (brain hemorrhage) | 1.00 | 1.00 | 1.00 | 24 |
| Peptic ulcer diseae | 1.00 | 1.00 | 1.00 | 17 |
| Pneumonia | 1.00 | 1.00 | 1.00 | 28 |
| Psoriasis | 1.00 | 0.86 | 0.93 | 22 |
| Tuberculosis | 1.00 | 1.00 | 1.00 | 25 |
| Typhoid | 1.00 | 1.00 | 1.00 | 19 |
| Urinary tract infection | 1.00 | 1.00 | 1.00 | 26 |
| Varicose veins | 1.00 | 0.95 | 0.98 | 22 |
| hepatitis A | 1.00 | 1.00 | 1.00 | 34 |
| accuracy | | | 0.94 | 984 |
| macro avg | 0.93 | 0.94 | 0.93 | 984 |
| weighted avg | 0.93 | 0.94 | 0.93 | 984 |

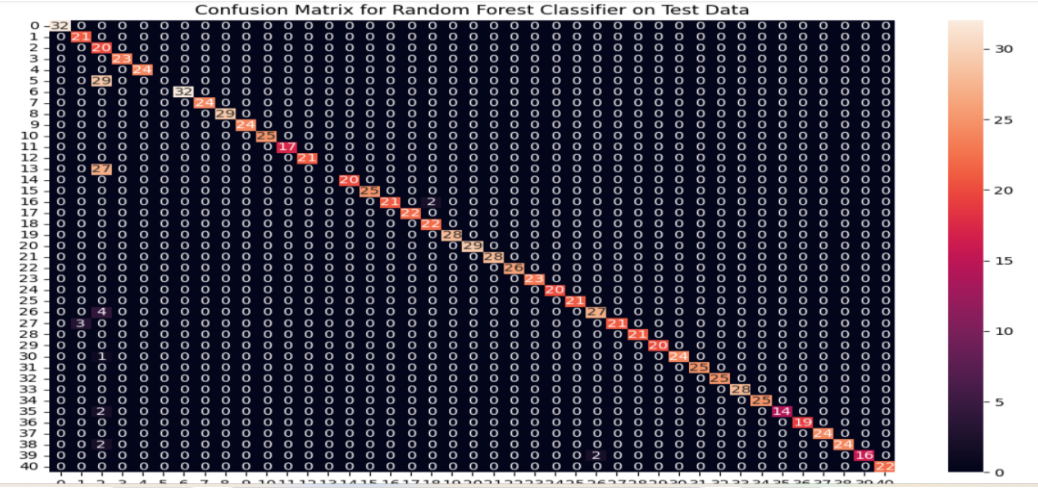
### 6.1 ACCURACY RESULTS

**TABLE 6.1 ACCURACY TABLE**

| Algorithm used | Accuracy score |
|---|---|
| Support vector machine | 0.9380 |
| Random Forest | 0.9380 |
| Naïve Bayes | 0.9380 |

### 6.2 Confusion matrix






$$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1-score} = 2 * \frac{precision*Recall}{Precision+Recall}$$

TP is the Instances that are actually positive and are correctly predicted as positive.
TN is Instances that are actually negative and are correctly predicted as negative.
FP is the Instances that are actually negative but are incorrectly predicted as positive.
FN is the Instances that are actually positive but are incorrectly predicted as negative

### 6.3 Prediction Results

**Fig 6.3.1 PREDICTED RESULT-1**

```
_skin'],symptoms_dict['nausea'],symptoms_dict['loss_of_appetite'],symptoms_dict['abdominal_pain'],symptoms_dict['yellowing_of_eyes']]] = 1

array(['Chronic cholestasis'], dtype=object)
```

**Fig 6.3.2 PREDICTED RESULT-2**

```
input_vector = np.zeros(len(symptoms_dict))
input_vector[[symptoms_dict['itching'], symptoms_dict['skin_rash'],symptoms_dict['nodal_skin_eruptions']]] = 1
rf_clf.predict_proba([input_vector])
rf_clf.predict([input_vector])

array(['Fungal infection'], dtype=object)
```

**Fig 4.5 PREDICTED RESULT-3**

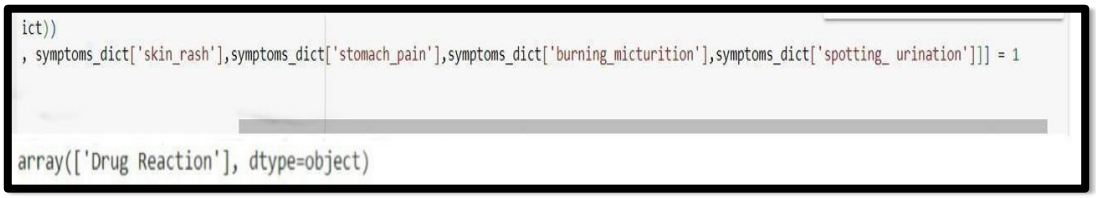



**Fig 4.6 PREDICTED RESULT-4**


**Fig 4.7 PREDICTED RESULT-5**


## 7. CONCLUSION

We set out to create a system which can predict disease on the basis of symptoms given to it. Such a system can decrease the rush at OPDs of hospitals and reduce the workload on medical staff. We were successful in creating such a system and use three different algorithms to do so. On an average we achieved accuracy of
~95%. Such a system can be largely reliable to do the job. Creating this system we also added a way to store the data entered by the user in the database which can be used in future to help in creating a better version of such a system. Our system also has an easy to use interface. It also has various visual representation of data collected and results achieved.

**FURTHER SCOPE**

- Facility for modifying user details.
- More interactive user interface.
- Facilities for Backup creation.
- Can be implemented as a Web page.
- Can be implemented as a Mobile Application.
- More Details and Latest Diseases.

## 8. REFERENCE

[1] Jaymin Patel, Prof.Tejal Upadhyay, Dr. Samir Patel "Heart disease prediction using Machine learning and Data Mining Technique" Volume 7.Number1 Sept 2015March 2016.

[2] "Disease Prediction Using Machine Learning Over Big Data" Vinitha S, Sweetlin S,Vinusha H and Sajini S (2018).

[3] "Multi Disease Prediction Using Data Mining Techniques" K.Gomathi , Dr. D.Shanmuga Priyaa (2017).

[4] Implementing WEKA for medical data classification and early disease prediction. "3rd IEEE International Conference on "Computational Intelligence and Communication Technology" (IEEE-CICT 2017)".

[5] Kaveeshwar, S.A., and Cornwall, J., 2014, "The current state of disease mellitus in India".AMJ, 7(1), pp. 45-48.

[6]Dean, L., McEntyre, J., 2004, "The Genetic Landscape of Disease [Internet].Bethesda (MD): National Center for Biotechnology Information (US); Chapter 1, Introduction to Disease. 2004 Jul 7.

[7]Disease Prediction and Doctor Recommendation System by https://www.irjet.net

[8] GDPS - General Disease Prediction System by https://www.irjet.net

[9]Machine Learning Methods used in Disease by https://en.wikipedia.org/wiki/Machine_learning

[10] https://ieeexplore.ieee.org/document/8819782

[11] http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html

[12] https://ieeexplore.ieee.org/document/8819782