IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 8, June 2022
Impact Factor: 6.252

Disease Prediction System Based on Support Vector Machine, Random Forest and Naive Bayes

Ms. Aparna M. Bagde[1], Shreya Wadkar[2], Aditya Patil[3], Amol Nidankar[4]
Assistant Professor, Department of Computer Engineering, NBN Sinhgad School of Engineering, Pune[1]
UG Student, Department of Computer Engineering, NBN Sinhgad School of Engineering, Pune[2,3,4]

Abstract: *The development and application of several leading data-mining techniques in many real-world application areas (e.g., industrial, healthcare, and life sciences) has led to their use in machine learning environments to extract important pieces of information from specified data in health communities, biomedical fields, and so on. Accurate medical database analysis benefits early disease detection, patient care, and community services. Machine learning techniques have been successfully used in a variety of applications, including early-stage disease prediction and diagnosis. This study demonstrates a disease prediction system built with machine learning algorithms such as the Decision Tree classifier, the Random Forest classifier, and the Naive Bayes classifier.*

Keywords: Machine Learning, Data mining, Decision Tree classifier, Random Forest Classifier, Naive Bayes Classifier, etc.

I. INTRODUCTION

Today's healthcare industry is a multibillion-dollar enterprise. The healthcare industry uses and generates a large amount of data that can be used to extract information about a disease for a patient. This healthcare data will be used to develop the most effective and efficient treatments for patients' health. This area also requires some improvement through the use of informative data in healthcare. However, because there is so much data to extract information from, some data mining and machine learning techniques are used.

The expected outcome of this project is to predict the disease in advance, so that the risk of death can be avoided early on, lives can be saved, and treatment costs can be reduced. India should also adopt the non-manual medical treatment system, which is best suited for improving and comprehending human health. The main reason is to apply the concept of machine learning in healthcare to improve patient care. Machine learning has already made identifying and forecasting various diseases much easier. Many machine learning algorithms used in disease prediction analytics help us predict diseases and treat patients effectively. Machine learning disease prediction employs medical histories and health data, as well as data mining and machine learning techniques and algorithms. Malaria, dengue fever, impetigo, diabetes, migraines, jaundice, chickenpox, and other health issues have a significant impact on health and can even result in death. By "evaluating" their massive database, the healthcare industry can make an informed decision. Specifically, the hidden patterns and relationships in the database were extracted. Data mining algorithms such as decision trees, random forests, and naive Bayes can be useful in this situation. As a result, they created an automated system that can discover and extract hidden knowledge associated with diseases from a historical database (disease symptoms) using the respective algorithms' rule set.

Overview:

The dataset under consideration contains 132 symptoms, the combination or permutation of which results in 41 diseases. Based on the 4920 patient records, we hope to create a prediction model that takes the user's symptoms and predicts the disease he is more likely to have.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 8, June 2022
Impact Factor: 6.252

Table I: Diseases and Symptoms

Sr. No.	Diseases	Symptom		Sr. No.	Diseases	Symptom	
1	Back pain	Bloody stool	scurrying	20	Diabetes	Hepatitis D	Hyperthyroidism
2	Constipation	depression	Passage of gases	21	Gastroenteritis	Hepatitis E	Hypoglycemia
3	Abdominal pain	Irritation in anus	Weakness in limbs	22	Bronchial Asthma	Alcoholic hepatitis	Cervical Spondylosis
4	diarrhea	Neck pain	Fast heart rate	23	Hypertension	Tuberculosis	Arthritis
5	Mild fever	dizziness	Internal itching	24	Migraine	Common cold	Osteoarthritis
6	Yellow urine	cramps	Toxic look	25	Paralysis	Pneumonia	
7	Yellowing of eyes	bruising	Palpitations	26	Jaundice	Heart Attack	
8	Acute liver failure	obesity	Painful walking	27	Yellow crust ooze	Swelling joints	Coma
9	Fluid overload	Swollen legs	Prominent veins on calf	28	Loss of smell	Stiff neck	Unsteadiness
10	Swelling of stomach	irritability	Fluid overload	29	Movement stiffness	Muscle weakness	ling lips
11	Swelled lymph nodes malaise	Swollen blood vessels Muscle pain	Excessive hunger Black heads	30	Spinning movements	Red around nose	Weakness of one body side
12	Blurred and distorted vision	Pain in anal region	Pain during bowel movements	31	Bladder discomfort	Foul smell of urine	Continuous feel of urine
13	Fungal Infection	Malaria	Varicose veins	32	Altered sensorium	Red spots over body	Abnormal menstruation
14	Allergy	Chickenpox	Hypothyroidism	33	Dichromic patches	Watering from eyes	Increases appetite
15	Gerd	Dengue	Vertigo	34	Lack of concentration	Visual disturbances	Receiving blood transfusion
16	Chronic cholestasis	Peptic ulcer disease	acne	35	Receiving unsterile injections	Distention of abdomen	History of alcohol consumption
17	Drug reaction	Hepatitis A	Urinary tract infection	36	Puss filled pimples	Blood in sputum	Stomach bleeding
18	Piles	Hepatitis B	Psoriasis	37	Silver like dusting	Small dents in nails	Inflammatory nails
19	AIDS	Hepatitis C	Impetigo	38	blister		

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 8, June 2022
Impact Factor: 6.252

II. IMPLEMENTATION

Inputs (Symptoms)

We designed the model assuming that the user is aware of the symptoms he is experiencing. The developed prediction considers 95 symptoms, among which the user can provide his processing as input.

Data Preparation:

Data pre-processing is a data mining technique that transforms or encodes raw data so that it can be easily interpreted by the algorithm. The following pre-processing techniques are used in the presented work:

Data Cleaning:

Data is cleansed through processes such as filling in missing values, thereby resolving data inconsistencies.

Data Reduction:

When dealing with large databases, analysis becomes difficult. As a result, we eliminate those independent variables (symptoms) that may have little or no effect on the target variable (disease). The current study chose 95 of 132 symptoms that are closely related to diseases.

Models Selected

The system is trained to predict the diseases using three algorithms

Disease Tree Classifier

- Random forest Classifier
- Naïve Bayes Classifier

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

Output(diseases)

Once the system has been trained with the training set using the aforementioned algorithms, a rule set is formed, and when the symptoms are provided by the user as input to the model, those symptoms are processed in accordance with the rule set developed. As a result, classifications are made and the most likely disease is predicted.

III. METHODOLOGY

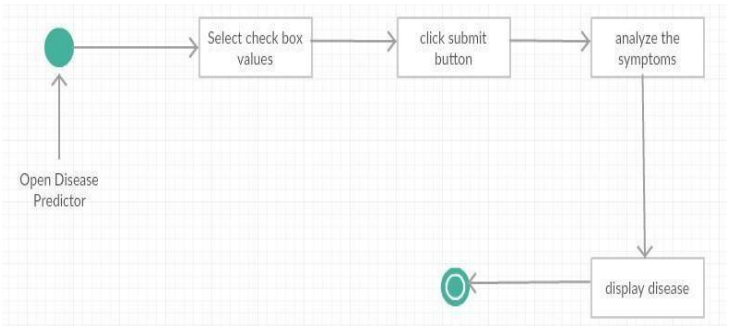


Figure: Work Flow [1]

The disease prediction system utilizes three data mining algorithms: the Decision tree classifier, the Random Forest classifier, and the Naive Bayes classifier. The algorithms' description and operation are provided below.

Decision Tree Classifier

The classification models created by decision tree have a tree-like structure. By learning a series of explicit if-then rules on feature values (in our case, symptoms), it breaks down the dataset into smaller and smaller subsets,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 8, June 2022
Impact Factor: 6.252

resulting in the prediction of a target value (disease). A decision tree is made up of decision nodes and leaf nodes. A decision node is one that has two or more branches. All symptoms are regarded as decision nodes in the work presented.

Decision Node:

A decision node is a node that has two or more branches. All symptoms are considered decision nodes in the study given.

Leaf Node:

The classification, or decision, of any branch is represented by the leaf node. The diseases correspond to the leaf nodes in this diagram.

ID3 Algorithm:

The ID3 algorithm, created by J.R. Quinlan, is one of the most important algorithms we've utilized in our work.ID3 performs a greedy top-down search across the given columns, testing each column (attribute=symptoms) at each node and selecting the best attribute (symptom) for classification of a given set.ID3 uses Entropy and Information Gain to determine which symptom is ideal for building a decision Tree.

Entropy is the measure of randomness or uncertainty. It denotes the predictability of a specific event. To create a decision tree, we must first generate two types of entropy using frequency tables for each attribute: Entropy E(C) utilising a frequency table of one attribute, where C is the current state (current outcomes) and P(h) is the probability of an event h in that state C:

$$E(C) = \sum_{h \in H} -P(h) \log_2 P(h) \qquad (1)$$

Entropy E (C, A) is calculated using a frequency table of two attributes, C and A, where C is the present state with attribute A and A is the considered attribute, and P(h) is the probability of an event H with attribute A.

$$E(C, A) = \sum_{h \in H} [P(h) * E(C)] \qquad (2)$$

The first element E(C) represents the Entropy of the entire set, while the second term E (C, A) represents an attribute A.

Information Gain:

Information gain (also known as Kullback-Leibler divergence) is an effective change in entropy after finalising an attribute A for a state C represented by IG (C, A). It calculates the relative change in entropy in relation to the symptoms, as shown below

$$IG(C, A) = E(C) - E(C, A) \qquad (3)$$

Calculation of E(C):

Step 1:

$$E(C) = \sum_{h \in H} -P(h) \log_2 P(h)$$

$$= -\frac{8}{12} \log_2 \left(\frac{8}{12}\right) - \frac{4}{12} \log_2 \left(\frac{4}{12}\right)$$

$$= 0.91822$$

Step 2: The Root Node should be chosen first when creating a decision tree. The root node is the symptom with the most information gain. Calculate E (C, High fever) and IG (C, High fever) starting with the symptom "High Fever":

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 8, June 2022
Impact Factor: 6.252

Using formula (2) and (3)

$$IG(C, High\ fever) = E(C) - E(C, High\ fever)$$

$$IG(C, High\ fever) = E(C) - \sum_{h \in H} [P(h) * E(C)]$$

The possible values for a symptom are represented by 'h' in P(h). In the considered dataset, the symptom "High Fever" has two possible values: Present (1) and Absent (2). (0). Hence x= (Present, Absent).

$$IG(C, High\ fever) = [E(C) - P(C_{present}) * E(C_{present}) - P(C_{absent}) * * E(C_{absent})]$$

Among 12 examples, we have 7 cases where the "High temperature" symptom is present and 5 cases where it is not.

$$P(C_{present}) = \frac{Number\ of\ present\ events}{Total\ events} = \frac{7}{12}$$

$$P(C_{absent}) = \frac{Number\ of\ present\ events}{Total\ events} = \frac{5}{12}$$

Out of the seven current instances, four lead to Dengue and three to Malaria. So, Entropy for "Present" High Fever Values: $E(C_{present}) = -\frac{4}{7}\log_2\left(\frac{4}{7}\right) - \frac{3}{7}\log_2\left(\frac{3}{7}\right) = 0.98518$

$$IG(C, High\ fever) = E(C) - P(C_{present}) * E(C_{present}) - P(C_{absent}) * E(C_{absent})$$

$$= 0.91822 - \frac{7}{12} * 0.98518 - \frac{5}{12} * 0.721992$$

$$= 0.3483$$

Step 3:

Calculate the information gain for all the symptoms in the similar manner, thus we have:

$$IG(C, High\ fever) = 0.3483$$
$$IG(C, Vomiting) = 0.25155$$
$$IG(C, Shivering) = 0.0102$$
$$IG(C, Muscle\ wasting) = 0.0102$$

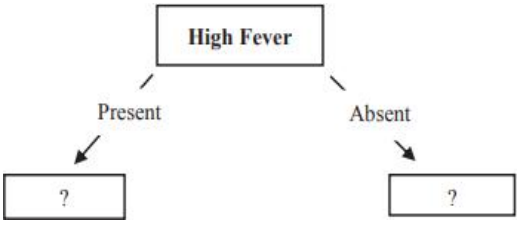


Figure 2: [1]

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 8, June 2022
Impact Factor: 6.252

Vomiting, shivering, and muscular wasting are the three remaining symptoms now that we've employed High fever. The sub-trees are formed by the two potential values of High fever: Present and Absent. Beginning the sub-tree of the Present: High fever is present in all 7 episodes, with 4 leading to Dengue and 3 to Malaria.

Similarly, we calculate the Information gain values w.r.to other symptoms as follows,

$$E(Highfever_{present}) = \sum_{h \in H} -P(h)\log_2 P(h) = -\frac{4}{7}\log_2\left(\frac{4}{7}\right) - \frac{3}{7}\log_2\left(\frac{3}{7}\right) = 0.98514$$

$$IG(Highfever_{present}, Vomiting) = 0.6518$$

$$IG(Highfever_{present}, Shivering) = 0.07718$$

$$IG(Highfever_{present}, Muscle\ wasting) = 0.0772$$

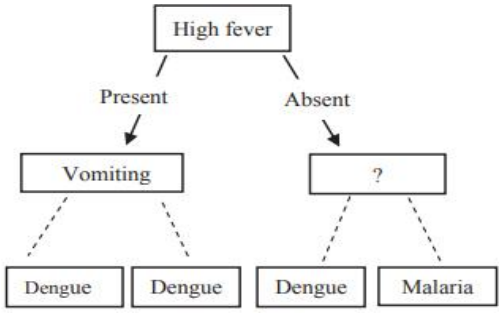


Figure 3: [1]

As a result, by proceeding in the same manner, the full Decision Tree can be constructed, eventually leading to Dengue and Malaria. A branch with entropy 0 is a leaf node, according to ID3.A branch with an entropy value larger than zero must be divided. If achieving zero entropy is not attainable, the decision is determined using the simple majority procedure. The final decision, like in the previous case, will be made by Dengue.

Limitation:

Over fitting occurred when all 132 symptoms from the original dataset were considered instead of 95 symptoms. i.e., the tree appears to remember the given dataset and so fails to classify new data. As a result, just 95 symptoms were considered during the data-cleaning stage, with the optimum ones being chosen.

Naive Bayes:

A type method primarily based totally on Bayes' Theorem with an assumption of independence amongst predictors. In easy terms, a Naive Bayes classifier assumes that the presence of a selected function in a category is unrelated to the presence of every other function.

Bayes' Theorem

Bayes theorem gives a manner to calculate the opportunity of a speculation given our earlier knowledge. In different phrases it unearths the opportunity of an occasion happening given the opportunity of every other occasion that has already occurred. Using Bayes theorem, we are able to discover the opportunity of a happening, for the reason that B has occurred. Here, B is the proof and A is the speculation. The assumption made right here is that the predictors/functions are independent. That is, the presence of 1 specific function does now no longer

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 8, June 2022
Impact Factor: 6.252

have an effect on the difference. Hence, it's far known as naive. Bayes' theorem is said mathematically as the subsequent equation: $P(A|B) = P(A \cap B)/P(B)$.

Wherein A and B are activities and $P(B) \neq 0$. Basically, we're looking for the opportunity of occasion A, given the occasion B is true. Event B is likewise termed as proof (A) is the priori of A (the earlier opportunity, i.e., Probability of occasion earlier than proof is seen). The proof is a characteristic cost of an unknown instance (right here, it's far occasion B). P(A|B) is a posteriori opportunity of B, i.e., opportunity of occasion after proof is seen.

Random Forest

Random Forest is a Supervised Machine Learning Algorithm that is used extensively in Classification and Regression problems. It builds choice bushes on certainly considered one among a type sample and takes their majority vote for class and is not an unusual place in case of regression.

Steps in Random Forest Algorithm:

In Random Forest Various numbers of random statistics are taken having n quantity of statistics. Individual choice bushes are constructed for each sample. Each Decision tree generates Output. Final Output is the Averaging for type and regression respectively.

Advantages of Random Forest:

Diversity- Not all attributes/variables/capabilities are considered at the same time as making a character tree, each tree is specific. Immune to the curse of dimensionality- Since each tree now not continues in thought all the capabilities, the feature space is reduced. Parallelization-Each tree is created independently out of various information and attributes. This way we can make entire use of the CPU to assemble random forests.

Train-Test break up- In a random wooded area we shouldn't segregate the information for teach and check as there will continuously be 30% of the statistics which isn't seen via the choice tree.

Stability- Stability arises because of the reality the give up end result is based mostly on majority vote casting/ averaging.

Hyper parameters In Random Forest:

Hyper parameter is used to boost overall performance and for Predictive energy of fashions or to run fashions faster.

n_estimators– quantity of bushes the set of rules builds earlier than averaging the predictions.

max_features– most quantity of capabilities in a random variable considers splitting a node.

mini_sample_leaf– determines the minimal quantity of leaves required to break up an inner node.

SVM

Support vector machines (SVMs) are effective but bendy supervised system mastering algorithms that are used each for type and regression. But generally, they may be utilized in type problems. An SVM version is largely an illustration of various lessons in a hyper plane in a multidimensional area. The hyper plane could be generated in an iterative way through SVM in order that the mistake may be minimized. The purpose of SVM is to divide the datasets into lessons to discover a most marginal hyperplane (MMH).

SVM Factors:

*H*yperplane - as we will see withinside the above diagram, it's far a choice aircraft or area that is divided among a hard and fast of gadgets having extraordinary lessons.

Support Vectors - Data Points which are closest to the hyperplane are known as aid vectors. Separating line could be described with the assistance of those records factors.

Margin - It can be described as the space among strains at the closet records factors of various lessons. It may be calculated because of the perpendicular distance from the road to the aid vectors. Large margin is taken into consideration as a terrific margin and small margin is taken into consideration as an awful margin.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 8, June 2022
Impact Factor: 6.252

Advantages of SVM classifiers - SVM classifiers give amazing accuracy and paintings nicely with excessive dimensional area. SVM classifiers essentially use a subset of schooling factors subsequently and the end result makes use of very much less memory.
Disadvantages of SVM classifiers - They have excessive schooling time subsequently in exercise now no longer appropriate for big datasets. Another downside is that SVM classifiers no longer paint nicely with overlapping lessons.

IV. IMPLEMENTATION AND RESULTS

The system was trained on the medical records of 4920 patients who were susceptible to 41 diseases as a result of a combination of symptoms. To avoid overfitting, we took into account 95 of the 132 symptoms, on the dataset, we utilised the K fold cross validation technique (K=5) to test the performance of all three algorithms. The accuracy scores over each cross-validation fold (K=5) for each algorithm are shown in the box and whisker plot above. We may conclude from these results that all three methods perform admirably on the dataset. When compared to the other two algorithms, however, Naive Bayes appears to do slightly better.

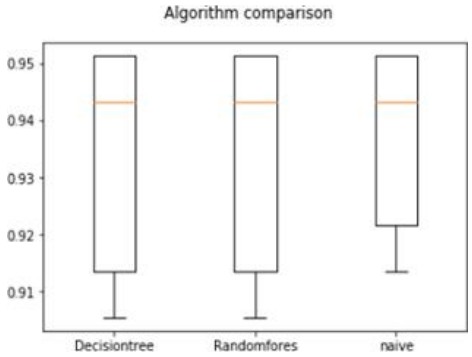


Figure 4: Algorithm Comparison [1]

Following training, the following were the accuracy scores for each algorithm:

Algorithm used	Accuracy score
Decision Tree	0.932927
Random Forest	0.932927
Naïve Bayes	0.936179

Table 2: Accuracy Table

Algorithm used	Accuracy score	Confusion matrix	
		Correctly classified	Incorrectly classified
Decision Tree	0.951219	39	2
Random forest	0.951219	39	2
Naïve Bayes	0.951219	39	2

Table 3: Accuracy and Confusion Matrix

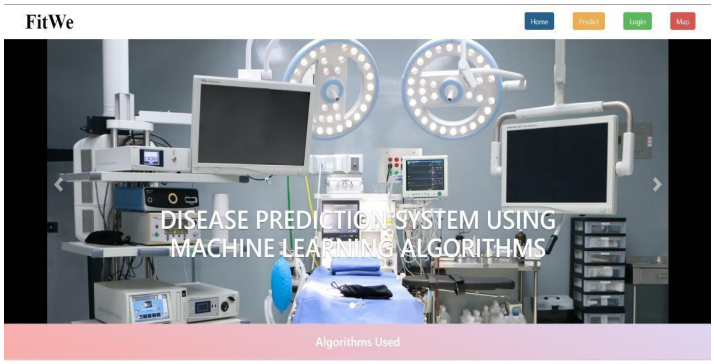
Algorithm Performance on Test Data

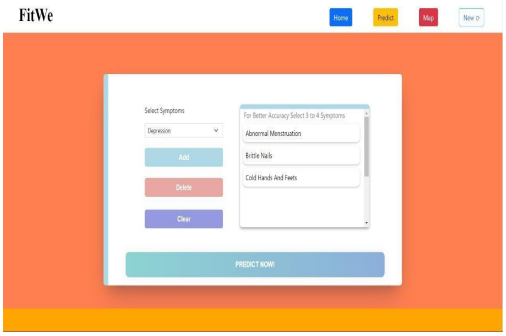
Following training, the system was put to the test on 41 new patient records with 95 symptoms. The confusion matrix and accuracy score are calculated as above:

From the above table, we can infer that all the algorithms have equal accuracy score. The accuracy in terms of percentage: 95.12 percentage.

GUI / Snapshots

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 8, June 2022
Impact Factor: 6.252





V. CONCLUSION

It can be seen from the history of machine learning and its applications in the medical field that techniques and methodologies have arisen that have enabled complex data analyses through the easy and direct usage of machine learning algorithms. This research shows a thorough comparison of the performance of three algorithms on a medical record, each of which provides up to 95% accuracy. The confusion matrix and precision score are used to evaluate the performance. Because of the vast amounts of data produced and saved by current technologies, artificial intelligence will play an ever-bigger role in data analysis in the future.

REFERENCES

[1] Disease Prediction using Machine Learning Algorithms, Sneha Grampurohit, Chetan Sagarnal, Published in:2020 International Conference for Emerging Technology (INCET))

[2] Machine learning equipped web-based disease prediction and recommender system, Harish Rajora, Narinder Singh Pun, Sanjay Kumar Sonbhadra, and Sonali Agarwal, arXiv:2106.02813v2 [cs.CV] 4 Jul 2021March - 2020

[3] Disease Prediction System, Aditya Tomar, International Journal of Advanced Research in computer and Communication Engineering, ISO 3297, July 2016

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 8, June 2022
Impact Factor: 6.252

[4] Disease Prediction System using data mining techniques, Aditya Tomar, International Journal of Advanced Research in computer and Communication Engineering, ISO 3297, July 2016
[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 Detection System”, IJERT, ISSN:22780181, Vol.2 Issue 4, April-2013
[7] Shadab Adam et.al “Prediction system for Heart Disease using NaïveBayes”, International Journal of advanced Computer and Mathematical Sciences, ISSN 2230- 9624, Vol 3, Issue 3,2012, pp 290-294[Accepted- 12/06/2012].
[8] Min Chen, Yixue Hao et.al “Disease Prediction by Machine Learning over big data from Healthcare Communities”, IEEE [Access 2017]
[9] Mr Chintan Shah,Dr. Anjali Jivani, “Comparison Of Data Mining Classification Algorithms for Breast Cancer Prediction”, IEEE-31661
[10] Palli Suryachandra, Prof. Venkata Subba Reddy, “Comparison of Machine Learning algorithms For Breast Cancer”, IEEE.