IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Drug Recommendation System forNon-Communicable Disease

Kundan Kumar Sah and Dr. B V A N S S Prabhakar Rao
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, India

***Abstract:*** The second cause of death is cancer in this world. In 2019, around 9 million cancer-related patients died. Breast cancer is the primary reason behind the death of women. It is the leading cause of women's deaths and is the most prevalent type of cancer in the world. Many kinds of studies on early detection of breast cancer have been carried out to start treatment and increase the likelihood of survival. It is crucial to find different, cheaper and safer data sets which can make forecasts more secure and easier for the implementation and working of alternative methods. This article provides a mixed model of several machine learning algorithms, including the Logistic Regression, Nearest Neighbors and Decision Tree, for the detection of breast cancer. In this review, data sets used to identify and diagnose breast cancer are also discussed.

***Index Terms*** **- Breast Cancer, NCD, Machine Learning.**

## I. INTRODUCTION

Non-Communicable Diseases (NCDs) is a disease which has a long-standing and slow progression in human bodies. Over 70% of deaths worldwide are caused by NCDs. Chronic NCDs are common among people under the age of 45. This category covers diseases such as cancer, diabetes, cardiovascular and chronic Lung diseases. From which second leading cause of death for women is breast cancer (after lung cancer). In recent years, there are a large number in the women with an estimated 40,450 women who are to be diagnosed as new cases of invasive breast cancer. Breast cancer represents approximately 12.5% of all new cancers and 25.4% [2] of all cancers in women. Breast cancer is a breast cancer that spreads to other parts of the body. Cancer increases when cells start to develop out of control. Cells of breast cancer typically form a tumour that is often seen or felt in x-rays. Breast cancer can spread when cells enter the blood or lymph system and reach other body areas. The cause of breast cancer includes changes in DNA and mutations. There are many algorithms for classifying breast cancer outcomes. Side effects of breast cancer include: tiredness, headache, pain and numbness (peripheral neuropathy), bone loss, and osteoporosis. The classification and prediction of outcomes of breast cancer are several algorithms available.

This paper compares the performance of four classifiers: Logistic Regression, KNN, Naïve Bayes and Decision Tree, among the most influential Machine Learning algorithms. [14] In order to prevent cancer from spreading, patients must undergo breast cancer surgery, chemotherapy, radiotherapy and endocrine surgery. The study aims to define and classify patients with malignant and benign diseases and to decide how our classification methods can be parameterized in order to achieve high precision.

## II. RELATED WORK

In the application of imprecise and ambiguous information, soft methods of computing play a complex role in judging. The implementation of soft calculation disciplines in medical applications is the creation of the interpretation and forecast of the enemy. In the numerous soft computing methods unclear qualified system utilises a furious system; data is meant as a set of obvious philological laws. Breast cancer research questions about ambiguity and fuzziness associated with correct input action and inadequacy of expectation results [7]. There are, however, many technology-oriented studies mentioned for the study of breast cancer, with few studies for the prediction of breast cancer. [1] To further endorse the procedure of breast cancer prediction, an unpredictable expert system for breast cancer forecast. This is done to capture confusing and imprecise details in the classification of breast cancer.

The paper used a prematurely interpretable ambiguous reasoning model to diagnose the accuracy of the system by an average of 93 percent, which demonstrates the system's superiority Compared with other similar work in the prediction method. The study and forecast of breast cancer were two medical requests that put the researchers as a great test [13]. The use of machine learning and data processing techniques has changed the entire practice of diagnosing and predicting breast cancer.

Cancer of the breast Diagnose determines the specification of Diagnose and Predicted breast lump and breast cancer design. The Breast Cancer Prediction says that it is likely that breast cancer will return in patients who have their tumours removed. [8] These two issues were also mainly within the framework of the problems of the company to raise the diagnosis & outlook for breast cancer. The purpose of our study is to clarify the link between judgments, physical observations and research centers such as Image Processing Experts, Radiologists, to automatic breast cancer detection.

.

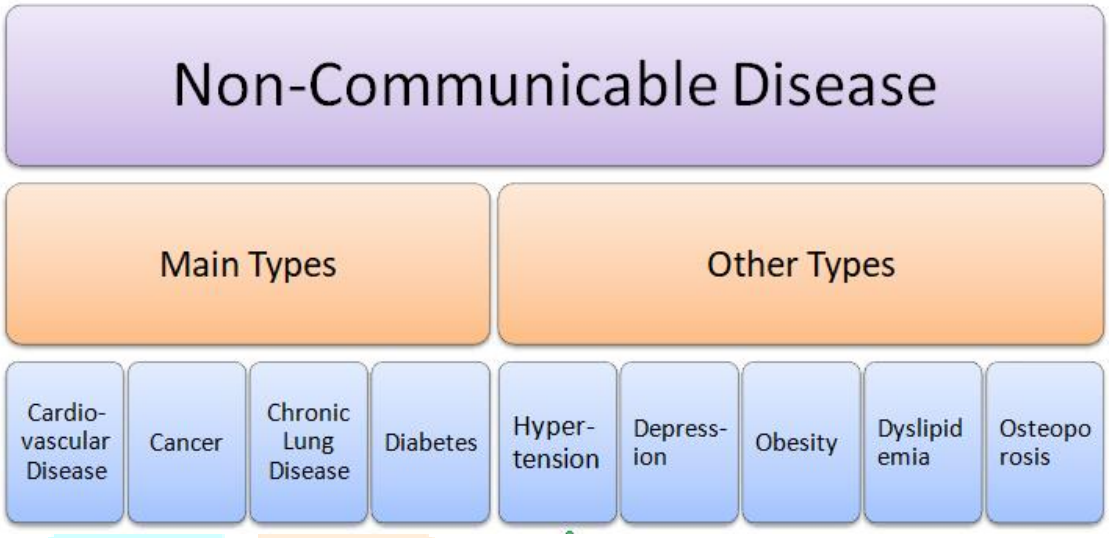


Fig.(i) Types of NCDs

## III. Research Methodology

**NUMPY** - Numpy is a commonly used cluster handling package that provides a high-quality multidimensional exhibit object as well as software for working with it. Numpy is a vital Python package for logical processing. It includes numerous highlights, including the following major ones:

- C/C++ and Fortarn code integration platform.
- An important entity of the N-dimensional sequence.
- Sophisticated feature (broadcasting).

**MATPLOTLIB** - Matplotlib is the two dimensional library of Python that provides high quality images and intuitive conditions across stages in different version classes. Python content, the Python and Python shell, the Jupyter notepad, the desktop application server and the four graphical user interface toolboxes are supported by MatPlotlib. The aim of Matplot is to make basic things simple and complicated things possible for simple thongs.

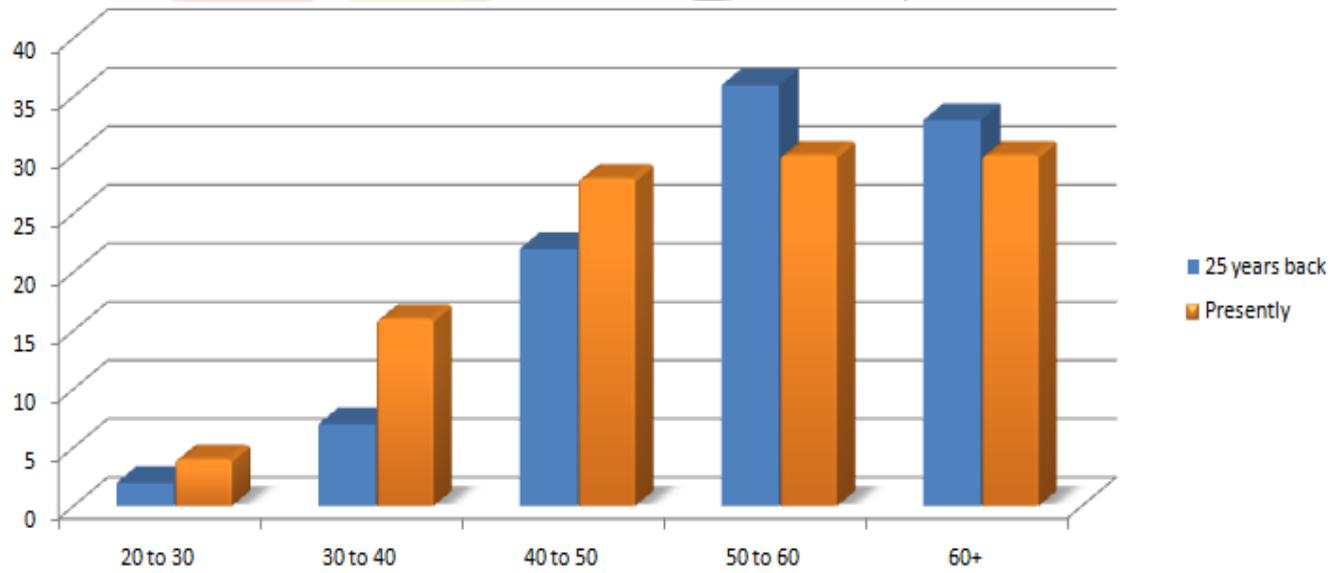


Fig.(ii)Non-Communicable Disease in India

With only a few lines of coding, you can create graphs, histograms, power spectra, bar outlines, error diagrams, scatter plots, and other graphics.

**PANDAS** - The panda is an open source Python library that uses its excellent data structures to provide superior data collection and analysis capabilities. For code theft and readiness, Python is commonly used. It was unconcerned with data processing. Pandas made the decision to take on the issue. Five traditional projects in data handling and investigation are planned, tracked, modelled, and analysed using this, with no respect for the cause of information overload. Python and Pandas are used in a wide range of areas, including academia, industry, and finance.

**SCIKIT-LEARN** – Scikit-learn provides a predictable Python interface for handling and unsupervised learning calculations. In several Linux distributions, it is agreed with a lenient rearranged BCD. Academic and industrial implementations are also possible for this licenced and id appropriated product. The scikit-learn library is built on top of the Scipy (logical Python) programming language, which must be learned before it can be used.

**STEPS INVOLVED IN PROJECT**

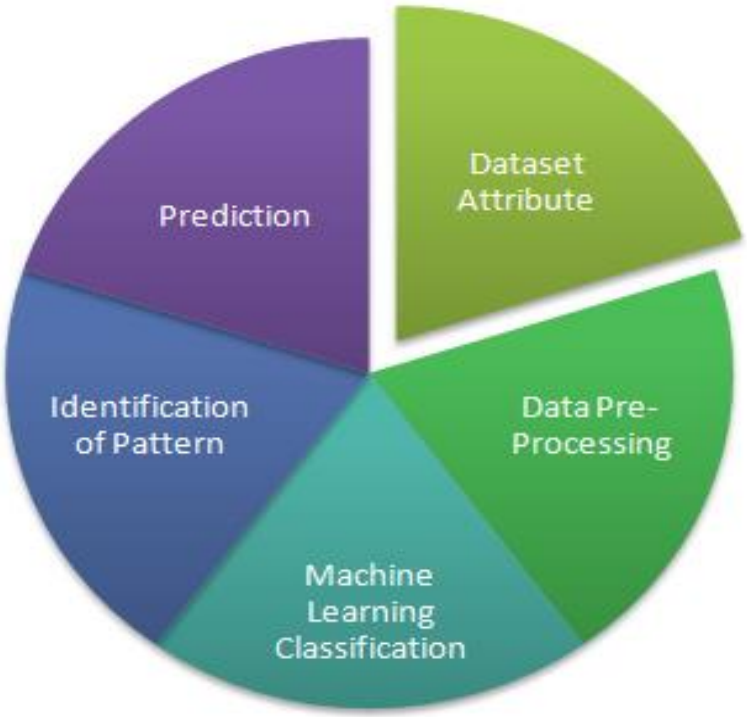


Fig.(iii) Steps involved in recommendation system

***Dataset-*** The cancer data in the dataset has a number of characteristics. In accordance with the dataset, the ML algorithm produces precise results.

***Pre-processing on Data-*** The first step is to compile and apply algorithms for the analysis of the data that we're interested in. The translation of raw data into a functional format is referred to as "data pre-processing."

***Preparing Data-*** Loading and storing data for use in computer science is what data preparation means. We'd start by gathering all of our results, and then arrange it in a random order.

***Classification Techniques-*** Methods of classification for machine teaching such as Logistic Regression, KNN, Nave Bayes, Decision Tree, and other techniques were employed in this project. To assess which algorithm had the highest classification accuracy, we measured the performance and efficiency of these algorithms.

## IV. DATASET

Any quest needs a data collection to work properly. We analysed the different forms of cancer that occur around the world using data sets from various sources and articles in this research report. The medical world is becoming more supportive of cancer. Breast cancer is also the most prevalent cancer in women. Breast cancer is the leading cause of death in women, according to medically confirmed evidence.

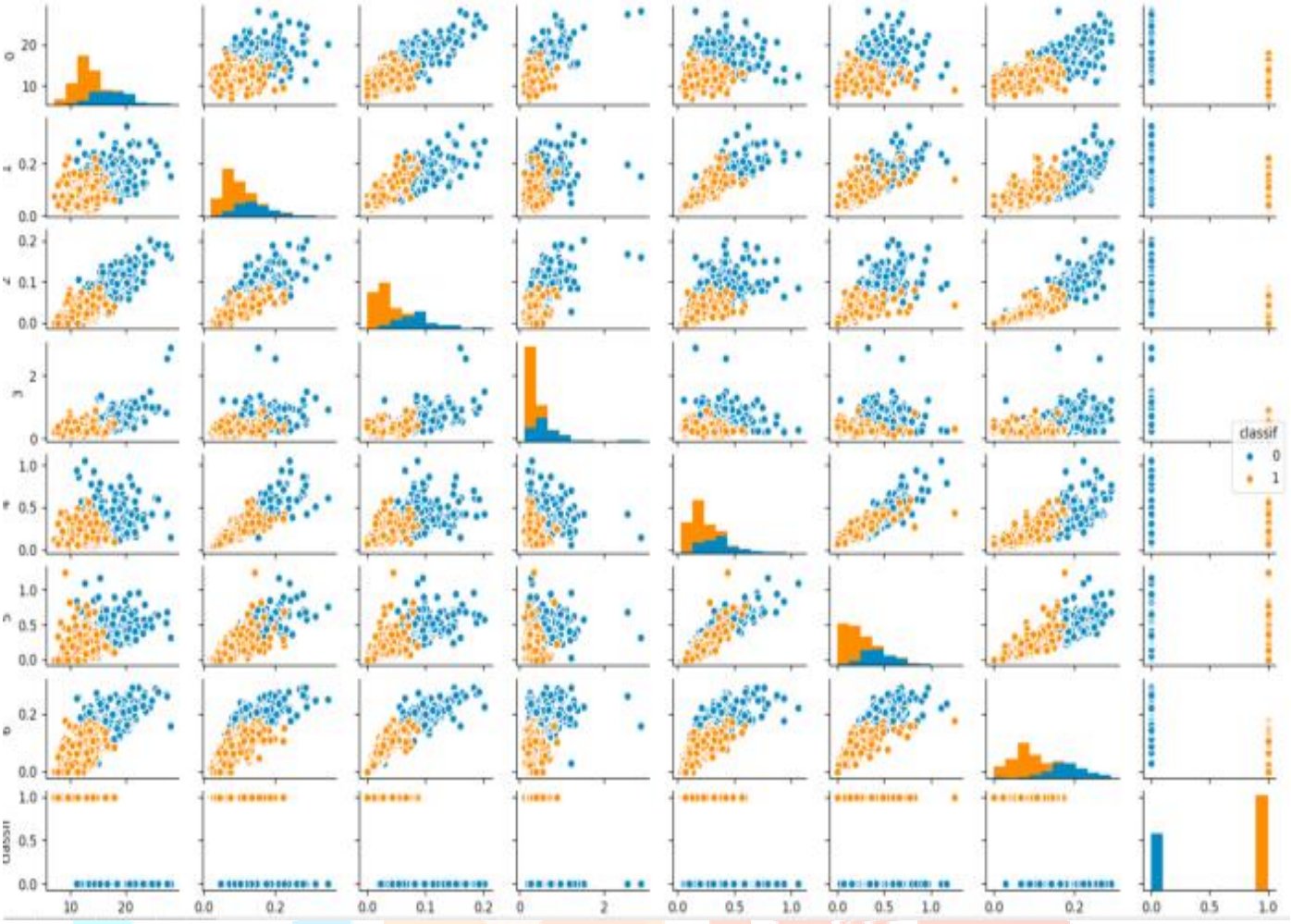


Fig.(iv) Identifying Malignant - Benign

## V. MACHINE LEARNING ALGORITHMS

### A. Logistic Regression

Logistic regression is one of the most common machine learning algorithms under the supervised learning approach. It is used to use a given set of independent variables to forecast the categorical dependent variable. It predicts a categorical dependent variable's output.

Therefore, a categorical or discrete value must be the product. It can be either Yes or No, 0 or 1, Incorrect and so on. So it gives the probabilistic values that lie between 0 and 1 instead of giving the same value as 0 and 1.

### B. K-Nearest Neighbor (KNN)

K-Nearest Neighbour is one of the most straightforward machine learning algorithms based on supervised learning methods. The K-NN algorithm assumes that the new case/data is similar to existing cases and places the new case in the most similar category. This algorithm stores all available data and classifies a new data point according to its similarity. This means that it can be conveniently categorised into a well-suited group using the K-NN algorithm as new data emerges. It can be used for both regression and classification, but it is often used for problems with classification.

### C. Naïve Bayes (NB)

The Naïve Bayes algorithm is a supervised theorem-based learning algorithm used to solve classification problems. It is mainly used in text classification that requires a large-scale preparation data set. This classification system is one of the most simple and potent algorithms that help generate fast machine learning models that can predict fast. It is a classifier that calculates probability on the basis of the probability of an object. Spam filtration, sentimental analysis, and article classification are some common examples of the Naïve Bayes Algorithm..

**ARCHITECTURE DIAGRAM**

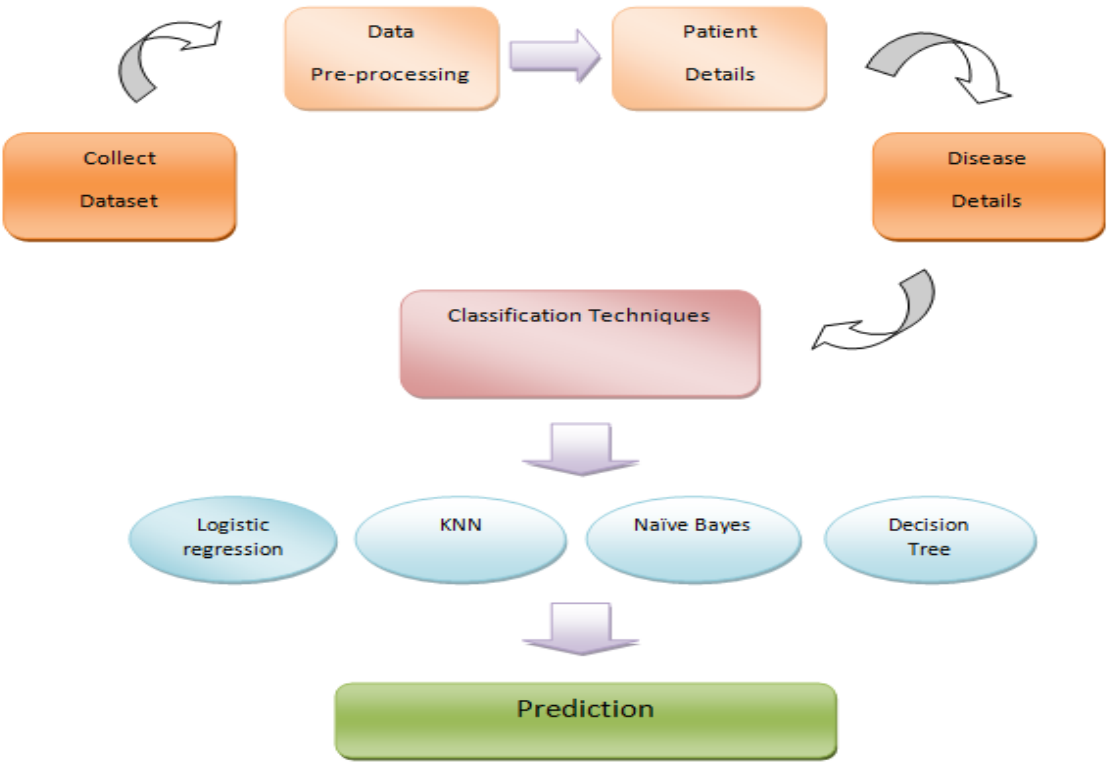


Fig.(v) Drug Recommendation system for Non-Communicable Disease.

**D. Decision Tree (DT)**

Decision Tree is a supervised learning approach that can be used for classification and regression challenges but is often preferred to solve classification problems. It is a classifier organised by a tree, with internal nodes defining the attributes of a dataset, branches represent the rules of judgement and the outcome is represented by each leaf node. There are two nodes in the Decision Tree, which are the Decision Node and the Leaf Node. Decision nodes are used to make any decisions and have several branches, while the performance of those decisions is Leaf nodes and there are no additional branches. The decisions or experiments are carried out on the basis of the features of the dataset in question.

**E. Support Vector Machine (SVM)**

Support for the Vector Machine or SVM is one of the most common supervised learning algorithms used for problems of classification and regression. However, it is primarily used for classification problems in machine learning. The aim of the SVM algorithm is to create the best line or decision line to divide n-dimensional space in classes so as to place the new data point conveniently into the right category in the future.

This limit is called a hyperplane of the best decision. SVM selects extreme points/vectors that assist in the building of the hyperplane. These extreme cases are known as vectors of support, and so the algorithm is known as the support vector machine.

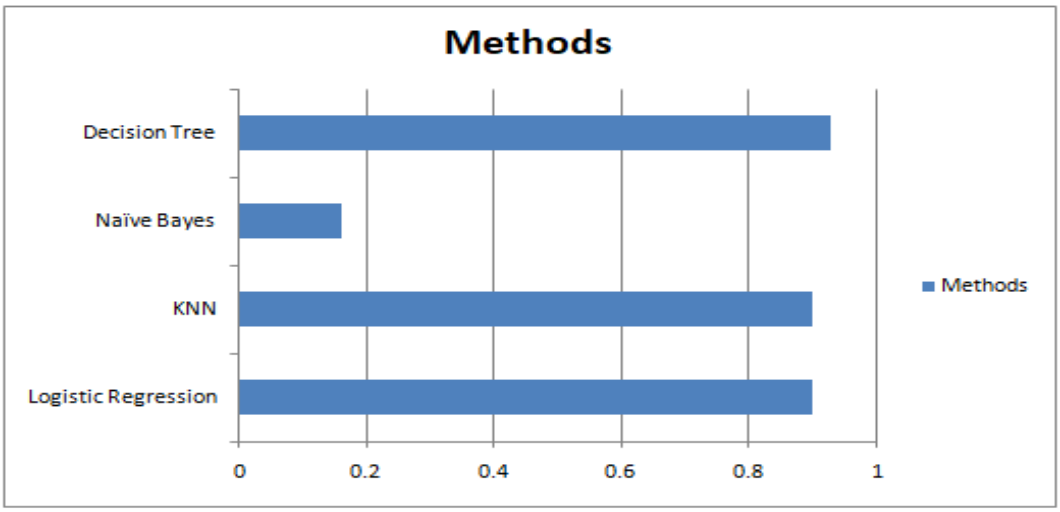


Fig.(vi) Using machine learning methods in cancer detection

### F. Random Forest (RF)

Random Forest is a popular algorithm for machine learning, part of the controlled learning system. It can be used for classification and regression problems in Machine Learning. It is based on the concept of ensemble learning, which combines many classifiers to resolve a complex problem and improve the effectiveness of the model. "Random Forest is a classifier that comprises a number of decision trees on different subsets of the given dataset and takes the average to enhance the predictive accuracy of that dataset," as the name implies. Instead of leaning on a decision tree the random forest takes the forecast from every tree and is based on the majority vote of forecasts, which forecasts the final output. The larger number of woodland plants leads to higher accuracy and prevent the issue of overfitting.

## VI. IMPLEMENTATION

A heat map is a graphical representation of data with colour-coded values.

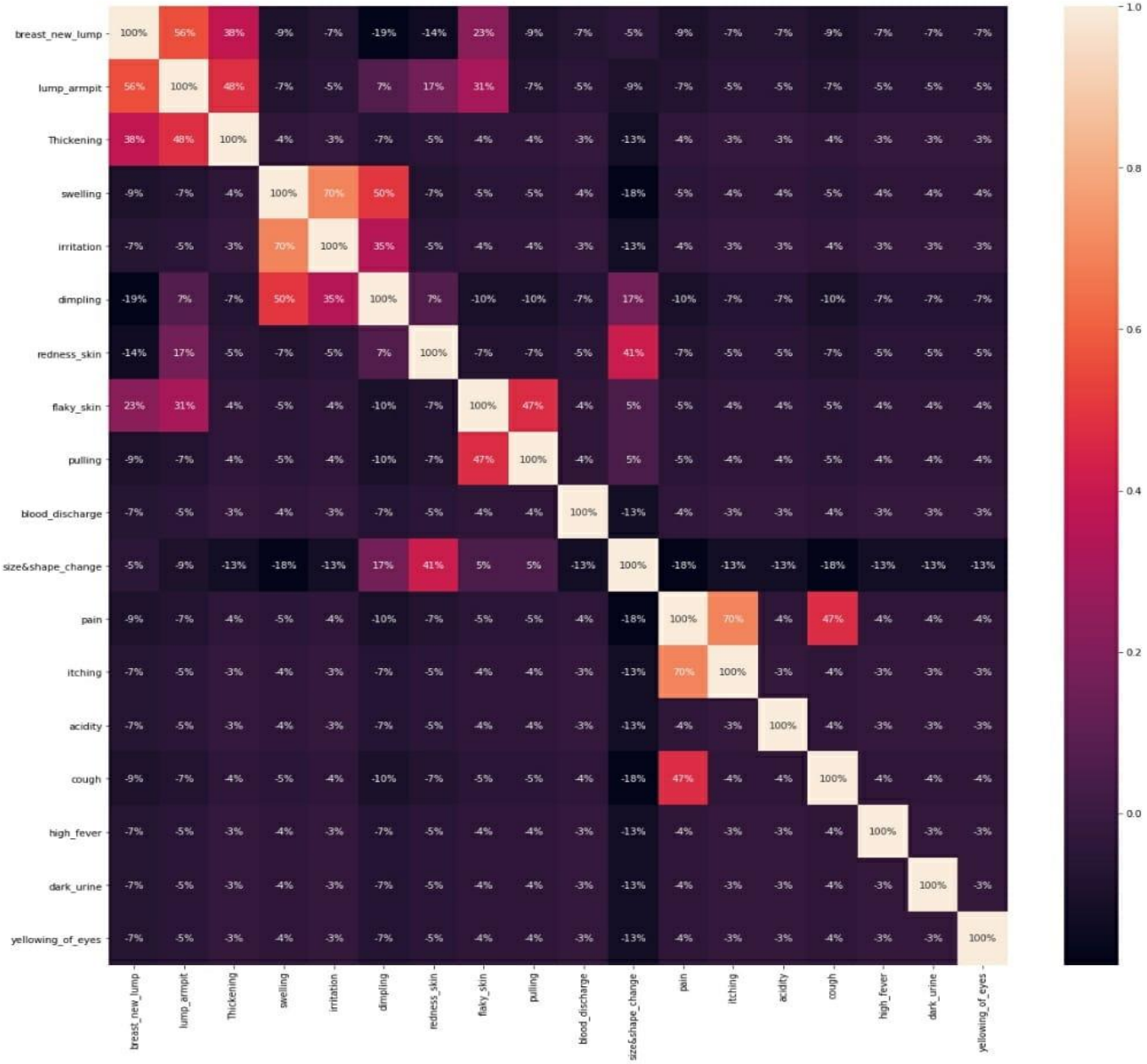


Fig.(vii) Heat map for complicated visualisation

Heat maps make visualising complicated details simple and knowing it at a glance: The data on the left and right are similar, but the data on the left is much easier to comprehend.This technique portrays the magnitude of a phenomenon in two dimensions as colour. The colour variation can be by hue or intensity, providing the reader with obvious visual signs of how the phenomenon is clustered or differs over space.

### 6.1. ACCURACY

```
[0]Logistic Regression Training Accuracy: 0.9
[1]K Nearest Neighbor Training Accuracy: 0.9
[2]Gaussian Naive Bayes Training Accuracy: 0.16666666666666666
[3]Decision Tree Classifier Training Accuracy: 0.9333333333333333
```

Above figure is showing accuracy result of particular algorithms according to their datasets.

```
[[10  0]
 [ 1  0]]
Model[0] Testing Accuracy = "0.9090909090909091!"

[[10  0]
 [ 1  0]]
Model[1] Testing Accuracy = "0.9090909090909091!"

[[ 0 10]
 [ 0  1]]
Model[2] Testing Accuracy = "0.09090909090909091!"

[[10  0]
 [ 1  0]]
Model[3] Testing Accuracy = "0.9090909090909091!"
```

```
#Using Logistic Regression
from sklearn.linear_model import LogisticRegression
log = LogisticRegression(random_state = 0)
log.fit(X_train, Y_train)

#Using KNeighborsClassifier
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors = 5,
                           metric = 'minkowski', p = 2)
knn.fit(X_train, Y_train)

#Using GaussianNB
from sklearn.naive_bayes import GaussianNB
gauss = GaussianNB()
gauss.fit(X_train, Y_train)

#Using DecisionTreeClassifier
from sklearn.tree import DecisionTreeClassifier
tree = DecisionTreeClassifier(criterion = 'entropy',
                              random_state = 0)
tree.fit(X_train, Y_train)
```

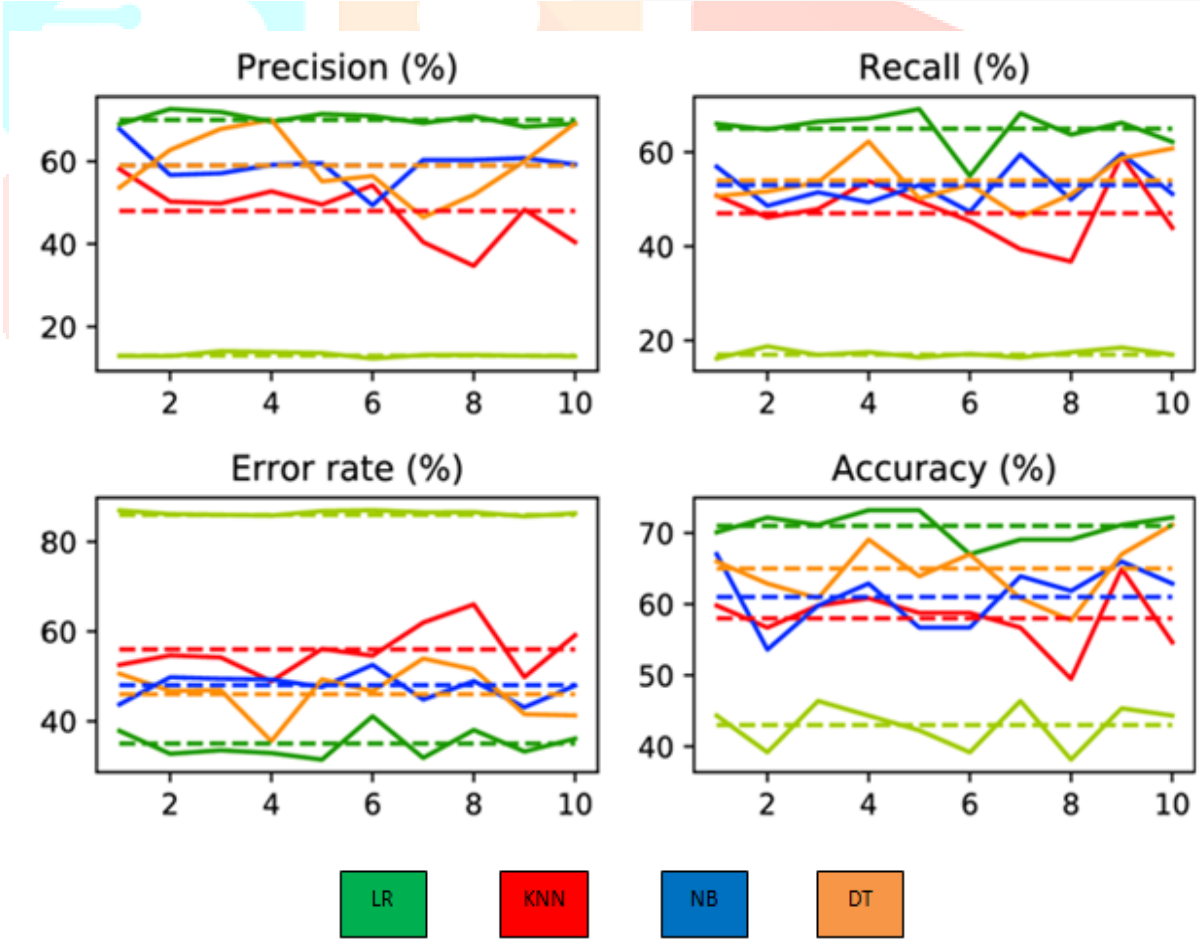


Fig.(viii) Accuracy and error rate graph

The above graph gives the overview in the reviewed documentary report of the use of ML methods and algorithms used for diagnosing breast cancer. The most widely used approach is Decision Tree. It is observed. In this figure, the outcomes of diagnosis of breast cancer using ML methods are mentioned above.

Outcomes Confusion matrix is a N x N matrix used to calculate the classification model's achievement, where N indicates the number of objective classes. The matrix compares the real objectives to the projections of the machine learning model. The value expected for the destination variable is shown in rows.

## VII. CONCLUSION

Various techniques of machine learning are available for the study of medical data. The building of reliable and computer-efficient classifiers for medical applications is a big role in machine learning. In this study, we used 4 primary algorithms: logistical regression, K- Nearest Neighbors, Naïve Bayes, and Breast Cancer Decision Tree (original). In order to find the best classification accuracy, we have attempted to compare the performance and efficiency of these algorithms with precision, sensitivity and specificity. Decision Tree's precision exceeds and thereby overcomes all other algorithms by 93.0 percent. Decision Tree has proved its efficacy in prediction and diagnosis of breast cancer and achieved the highest accuracy and low error rate results.

## REFERENCES

**[1]** Sri Hari Nallamala, Pragnyaban Mishra, Suvarna Vani Koneru, " Breast Cancer Detection using Machine Learning Way ", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S3, July 2019.

**[2]** M. Tahmooresi, A. Afshar, B. Bashari Rad1, K. B. Nowshath and M. A. Bamiah, "Early Detection of Breast Cancer Using Machine Learning Techniques", Asia Pacific University of Technology and Innovation (APU), University of Malaya, Malaysia. e-ISSN: 2289-8131 Vol. 10 No. 3-2 21, 2018.

**[3]** R. Kirubakaran, T. C. Jia, and N. M. Aris, "Awareness of Breast Cancer among Surgical Patients in a Tertiary Hospital in Malaysia," Asian Pacific Journal of Cancer Prevention, 2017, vol. 18, no. 1, pp. 115–120.

**[4]** K. Vanisree, S. Jyothi, "Decision Support System for Congenital Breast Cancer Diagnosis based on Signs and Symptoms using Neural Networks", International Journal of Computer Applications vol.19, issue.6, pp.6 – 12, 2011.

**[5]** S.F. Weng, J. Reps, J. Kai, J.M. Garibaldi, N. Qureshi, "Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data", vol.1, issue.12, pp. e0174944, 2017.

**[6]** M. Thiyagaraj, G. Suseendran, "Survey on Breast Cancer prediction system based on data mining techniques", Indian Journal of Innovations and Developments vol.6 issue.1, pp.1-9, 2017.

**[7]** C.S. Dangare, S.S. Apte, "Improved Study of Breast Cancer Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications vol.47, issue.10, pp. 44-48, 2012.

**[8]** S. Palaniappan, R. Awang, "Intelligent disease prediction system using data mining techniques", In 2008 IEEE/ACS international conference on computer systems and applications, pp. 108-115, 2008.

**[9]** Sri Hari Nallamala, Siva Kumar Pathuri, Dr Suvarna Vani Koneru, "A Literature Survey on Data Mining Approach to Effectively Handle Cancer Treatment", International Journal of Engineering & Technology (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7 (2018), SI 7, P. 729 – 732.

**[10]** Sri Hari Nallamala, Dr. Pragnyaban Mishra and Dr. Suvarna Vani Koneru, "Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems", International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), ISSN (ONLINE): 2278 – 3091, Vol. 8 No. 2 (2019), P. 259 – 264.

**[11]** Sri Hari Nallamala, Siva Kumar Pathuri, Dr Suvarna Vani Koneru, "An Appraisal on Recurrent Pattern Analysis Algorithm from the Net Monitor Records", International Journal of Engineering & Technology (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7 (2018), SI 7, P. 542 – 545.

**[12]** Y. Tsehay et al., "Biopsy-guided learning with deep convolutional neural networks for Prostate Cancer detection on multiparametric MRI," 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, VIC, 2017, P. 642-645.

**[13]** S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017, P. 13-14. doi: 10.1109/CEM.2017.7991863.

**[14]** M. R. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," 2016 9th International Conference on Developments in Systems Engineering (DeSE), Liverpool, 2016, P. 35-39.

**[15]** C. Deng and M. Perkowski, "A Novel Weighted Hierarchical Adaptive Voting Ensemble Machine Learning Method for Breast Cancer Detection," 2015 IEEE International Symposium on Multiple-Valued Logic, Waterloo, ON, 2015, P. 115-120.

**[16]** H. R. Mhaske and D. A. Phalke, "Melanoma skin cancer detection and classification based on supervised and unsupervised learning," 2013 International conference on Circuits, Controls and Communications (CCUBE), Bglore, 2013, P. 1-5.

**[17]** Datafloq - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining algorithmsdemystified/1144. Accessed December 29,2015.

**[18]** Chaurasia and S. Pal, "Data Mining Techniques : To Predict and Resolve Breast Cancer Survivability," vol. 3, no. 1, pp. 10–22, 2014.

**[19]** B V A N S S Prabhakar Rao, AI based E-Healthcare in Rural Areas, International Journal of Innovative Technology and Exploring Engineering Volume-9 Issue-3, 2020, Pp. 3098-3104.

**[20]** Djebbari, A., Liu, Z., Phan, S., AND Famili, F. International journal of computational biology and drug design (ijcbdd). 21stAnnual Conference on Neural Information Processing Systems (2008).

**[21]** S. Aruna and L. V Nandakishore, "Knowledge Based Analysis of Various Statistical Tools in Detecting Breast Cancer," pp. 37–45, 2011.

**[22]** Y, "An Empirical Comparison of Data Mining Classification Methods," vol. 3, no. 2, pp. 24–28, 2011.

**[23]** B V A N S S Prabhakar Rao, Disruptive Intelligent System in Engineering Education for Sustainable Development, Procedia Computer Science, Volume 172, 2020, Pages 1059-1065.

**[24]** Pradesh, "Analysis of Feature Selection with Classification : Breast Cancer Datasets," Indian J. Comput. Sci. Eng., vol. 2, no.5, pp. 756–763, 2011.

**[25]** Thorsten J. Transductive Inference for Text Classification Using Support Vector Machines. Icml. 1999;99:200-209. doi:10.4218/etrij.10.0109.0425.

**[26]** L. Ya-qin, W. Cheng, and Z. Lu, "Decision tree based predictive models for breast cancer survivability on imbalanced data," pp. 1–4, 2009.

**[27]** D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," Artif. Intell. Med., vol. 34, pp. 113–127, 2005.

**[28]** W. Version, "Machine Learning with WEKA," 2004.

**[29]** "UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set." [Online]. Available https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29. [Accessed: 29-Dec-2015]

**[30]** Manikandan and B V A N S S Prabhakar Rao, Knowledge Based Advisory System To Analyze Search Through Different Page Ranking Implementations, School of Computer Science and Engineering, VIT Chennai, Volume 8, Issue 6 June 2020 | ISSN: 2320-28820

**[31]** SUGI 31 Statistics and Data Analysis Receiver Operating Characteristic ( ROC ) Curves Mithat Gönen , Memorial Sloan-Kettering Cancer Center SUGI 31 Statistics and Data Analysis FN + FP," pp. 1–18, 2001.