Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020)
IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4

# *Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning*

Ram MurtiRawat
Department of Computer Science and Engineering
Delhi Technological University,
New Delhi, India

Shivam Panchal
Department of Computer Science and Engineering
Delhi Technological University,
New Delhi, India

Vivek Kumar Singh
Department of Computer Science and Engineering
Delhi Technological University,
New Delhi, India

Yash Panchal
Department of Computer Science and Engineering
Delhi Technological University,
New Delhi, India

***Abstract*— Breast Cancer is one of the most severe diseases that is faced by women leading nowhere other than increased death rates in society and it is considered to be one of the most intense disease in the history of medical science. Looking at the number of deaths caused by Breast cancer, it is considered to be a major threat but today's advancement in medical science has the capability to cure such threat completely if detected at its early stages without causing any harm to the patient. The major challenge arise during the detection of cancer and differentiating between the diagnosis that affirms whether the patient has a benign or malignant type of cancer. Machine Learning Algorithms like K-Nearest Neighbors, Support Vector Machine (SVM) and Artificial Neural Network (ANN) helps us solve this problem by achieving results with high precision and accuracy. The following paper helps in diagnosis of breast cancer using Logistic Regression (LR), K-Nearest Neighbors (KNN) and Ensemble Learning with Principal Component Analysis (PCA) and a comparative study is also made with other papers on the basis of accuracy. The models used here are trained and tested on Wisconsin breast cancer diagnosis data set which is taken from UCI machine learning repository. Pre processing of data was performed followed by feature extraction of data set using Principal Component Analysis (PCA). Various Machine Learning techniques were proposed in the paper which helped us achieve an accuracy of 98.60% using K-Nearest Neighbors, 97.90 using Logistic Regression and 99.30% using Ensemble Learning.**

***Keywords*— *Breast Cancer Diagnosis, Logistic Regression (LR), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Principal Component Analysis (PCA), Ensemble Learning, Machine learning.***

## I. INTRODUCTION

Cancer is considered to the most dangerous health problems leading to severe health conditions, cancer is the result of abnormal excess growth of cells in body which can cause severe health issues and can cause death. Breast cancer is a cancer which is a result of excess growth of breast tissue. Breast cancer is life threatening and is observed to be the second highest cause of deaths after lung cancer in women. About 1 in 8 women suffers from breast cancer once in their lifetime. Breast cancer alone lead to 627,000 deaths in 2018 with a frequency of 2.1 million in 2018 worldwide and more than 1 million cases are observed every year in India [1]. Breast cancer can be identified by observing presence lumps in breast. With the exponential growth and major advancement in the field of medical-science, breast cancer is curable if the threat is detected at early stages otherwise it can cause severe health issues and can cost life of patient.

The major challenge after detection of lumps in breast is to distinguish between a benign or malignant tumour, but today's advance machine learning algorithms allows us to design several advance models by using various machine learning techniques which makes it easier in detection and diagnoses of breast cancer by training the model with the help of previously observed data of patients. Several people has contributed in this field and developed models in the past to detect breast cancer and were successful in achieving remarkable results with good accuracy and precision. Out of various machine learning algorithms, Artificial Neural Network (ANN) and Support Vector Machine (SVM) have proved to be the ones which helped in achieving highest accuracies and are in frequent use [2].

Several people have contributed on the same field working on the WDBC dataset with different machine learning algorithms giving remarkable results. Previously a study was conducted on prediction of breast cancer using K-Nearest Neighbors and Support Vector Machine.[3] This system used 10-fold cross validation for accurate outcomes, the performance of the model was resulted considering specificity, sensitivity, accuracy, false discovery rate, Matthews correlation

coefficient and false omission rate. The technique was able to pull an accuracy of 97.14% and 98.57% with KNN and SVM respectively.

A paper was proposed by the students of University of Dammam, investigating the effect of correlation based feature selection using Artificial Neural Network and Support Vector Machine for breast cancer diagnosis [4], in this breast cancer diagnosis was performed using Artificial Neural Network and Support Vector Machine combined with feature selection. During this study 10-fold cross validation was performed in order to achieve maximum accuracy. Feature selection depends on correlation coefficient against the targeted class where multiple feature subsets were taken. The database used for this model was Wisconsin Diagnostic Breast Cancer database.
Their studies were able to generate an accuracy of 96.71% and 97.14% for Artificial Neural Network and Support Vector Machine respectively.

The below paper is structured into different sections. The related work brief us about the work previously done in the same field mentioned in section II. Section III talks about the Machine Learning Technique used in this paper and also talks about the dataset used. Section IV throws light on pre-processing stage and further performed experimental setup. Comparative analysis of result of various machine learning algorithms is given in Section V. Last section i.e. Section VI sums up our work.

## II. RELATED WORK

Past decades have been a great as several different machine learning techniques have been improved and tested for providing better accuracy for models of breast cancer detection. The below section talks about various related work in this field.

Jurgen Schmidhuber , Alessandro Giusti, Dan C. CireSan, and Luca M. Gambardella successfully proposed a model which helped in detection of the series of mitosis in breast cancer using advanced deep neural networks [5], for detecting mitosis they have used maximum pooling convolution neural network. The neural networks were capable enough to classify each and every pixel in the image, followed by simple post processing the neural output.

In past years, computerized tool were considered to play an essential role in diagnosis of breast cancer. Nithya and Santhi were able to propose a method for diagnosis in digital mammograms with the help of Grey Level Co-occurrence Matrix (GLCM) features [6]. Mammography is an important tool which is often used during early stages for detection of breast cancer. Effectiveness of this model is determined by classifying mammogram images into cancer and non-cancer images. Accuracy of this model was calculated to be 96%.
G. Manikandan, B. Karthikeyan , P. Rajendiran , R. Harish , T. Prathyusha , V. Sethu, worked on Breast cancer Prediction Using Ensemble Techniques [7]. Ensemble algorithm is a blend of combinational rules, in this paper they worked on various standard methodologies and successfully proposed an ensemble model achieving an accuracy of 97.07%. Ensemble Algorithm is considered successful in detection of breast cancer with high accuracy, S. k. Mandal, A. Gupta and Animesh Hazra proposed a paper, Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms [8], a comparative study was conducted between different machine learning techniques also measuring the time complexity of all the techniques used in this paper, their successful work helped the achieve an accuracy of 95.16%, 95.53%, 95.91% for Naive Bayes, Support Vector Machine and Ensemble technique respectively.

## III. BACKGROUND

### A. Dataset

The proposed model uses Wisconsin Diagnostic Breast Cancer (WDBC) Dataset from the UCI repository. This dataset was created and given by Dr. William H. Wolberg of the University of Wisconsin. This dataset was created in 1995. With 212 malignant instances and 357 benign instances, this dataset contains a total of 569 instances i.e 37.26% malignant tumour and 62.74% benign tumour. There are exactly 32 attributes WDBC dataset.

10 features or aspects were considered for each cell nucleus. These 10 features were: radius texture, smoothness, compactness, concavity, perimeter, area, concave points, symmetry, and fractal dimension. These are the features or aspects helped in describing the characteristics of cell nuclei in the image. To compute all these features a digitized image of a fine needle aspirate (FNA) of a breast mass was used.

For each image the mean value, standard error, and extreme value of each characteristic were computed. For each of the 569 instances, 30 features are observed. No missing values are present in this dataset. The dataset is stored in excel sheet format that contains 32 attributes.

The details of each attribute of WDBC dataset is shown in Figure 1 and Figure 2.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 32 columns):
id                       569 non-null int64
diagnosis                569 non-null object
radius_mean              569 non-null float64
texture_mean             569 non-null float64
perimeter_mean           569 non-null float64
area_mean                569 non-null float64
smoothness_mean          569 non-null float64
```

*Fig. 1 Details of first 7 Attributes of Dataset*

```
compactness_mean           569 non-null float64
concavity_mean             569 non-null float64
concave points_mean        569 non-null float64
symmetry_mean              569 non-null float64
fractal_dimension_mean     569 non-null float64
radius_se                  569 non-null float64
texture_se                 569 non-null float64
perimeter_se               569 non-null float64
area_se                    569 non-null float64
smoothness_se              569 non-null float64
compactness_se             569 non-null float64
concavity_se               569 non-null float64
concave points_se          569 non-null float64
symmetry_se                569 non-null float64
fractal_dimension_se       569 non-null float64
radius_worst               569 non-null float64
texture_worst              569 non-null float64
perimeter_worst            569 non-null float64
area_worst                 569 non-null float64
smoothness_worst           569 non-null float64
compactness_worst          569 non-null float64
concavity_worst            569 non-null float64
concave points_worst       569 non-null float64
symmetry_worst             569 non-null float64
fractal_dimension_worst    569 non-null float64
dtypes: float64(30), int64(1), object(1)
```

*Fig.2 Details of rest 25 Attributes of Dataset*

The dataset is divided into 2 parts, one part is for training and one part is for testing. The 75% dataset is used for training and 25% dataset is used for testing. Dataset is split using train_test_split function which is available in sklearn.model_selection package. The number of random_state which are used in train_test_split function is 7.

### *B. PCA*

The redundant attributes in data leads to noise in the data and increase the computational time of the model. Dimensionality reduction help in reduction in overfitting models, reducing noise, increasing model interpretability, reducing training time.

PCA (Principal Component Analysis) is considered to be a technique used for dimensionality reduction. PCA is considered to be an unsupervised method used to find the interdependence between set of variables. PCA helps in reducing the number of dimensions without loss of relevant information. PCA helps in mapping inputs d-dimensional space inputs to k (k<d) dimensional space outputs, with min loss of data and information [10]. In other terms, PCA helps in identifying patterns in data and expresses data highlighting their differences and similarities.

PCA is a useful statistical technique. It has applications in wide range of fields such as image compression and face recognition, and it is also a common technique for finding patterns in data of high dimension.

### *C. Logistic Regression*

In terms of statistics, the logistic regression is used for solving binary classification problem to model events or class probabilistically. Logistic Regression is statistical model used for modelling binary classification problem using logistic function and many more complex extensions are exist for logistic regression. Logistic regression is supervised classification algorithm. Logistic regression is basically a regression model which uses the regression model to predict the probability that a given data object or entry belong to given category [11]. Logistic Regression is based on the assumption that the linear function is followed by the data. Logistic Regression uses sigmoid function for modelling the data.

Equation of Linear Regression is as follows:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn \quad \ldots\ldots (1)$$

Where y is the dependent variable and x1,x2…xn are explanatory variables.

Sigmoid Function:

$$p = 1/1 + e^{-y} \quad \ldots\ldots (2)$$

After applying sigmoid function in equation 2 on equation 1,

$$p = 1/1 + e^{-(\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta nXn)} \quad \ldots\ldots (3)$$

The graph of sigmoid function is shown in Fig 3.

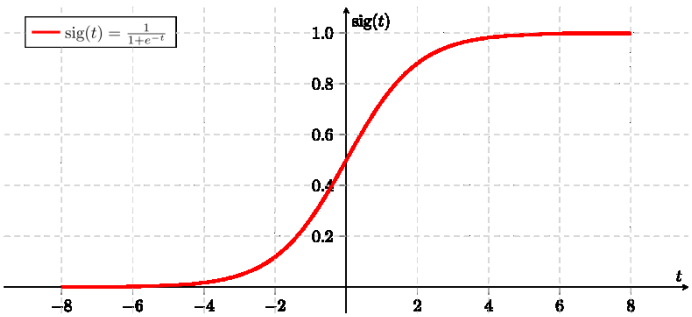



*Fig. 3 Sigmoid Function Graph*

The threshold is important aspect which make logistic regression a classification technique. In logistic regression, the setting of the value of threshold is very important aspect and it

is dependent on classification problem itself. Precision and Recall affects the value of threshold and ideally it is required that the values of precision and recall should be 1, but this seldom is the case.

### *D. Ensemble Learning*

In machine learning, ensemble methods collectively uses different algorithms in order to get better results and better performance than obtained from any alone standing algorithms. Statistical ensemble is usually infinite, a machine learning ensemble consists of finite set of alternate models, allowing more flexible structure to exist in these alternate methods [12].

A Voting Classifier is defined as a machine learning model that trains an ensemble of various models and helps in predicting more accurate output [13].
It aggregates the findings which are passed into Voting Classifier and helps in predicting the output based on voting and goes with the majority. The idea in ensemble learning is that instead of creating separate models and calculating accuracy separately, a model which are trained by these individual models are developed and delivers the output based on voting majority for each output. There are two types of voting classifiers, hard voting and soft voting classifiers. Since ensemble learning is a combined model of different machine learning techniques, the machine learning techniques used to train the ensemble model in this paper are Logistic Regression, K-Nearest Neighbor, Linear Discriminant Analysis, Support Vector Classifier and Random Forest Classifier.

### *E. K-Nearest Neighbor*

KNN (K-Nearest Neighbor) is one of the simplest forms of classification technique. In this technique, k training samples, whose attributes are relatively similar (closest) to the test samples, are found. These training samples are known as Nearest-Neighbors. The KNN algorithm assumes that similar things exist in close proximity i.e. similar things are near to each other. The proximity between similar things can be computed by computing the distance between them [14].
The class labels of k neighbors of the test sample decide the classification of the test sample. In this algorithm majority voting is used for classifying the test sample.
The value of K is crucial and thus has to be chosen wisely because if K is too small, then the classifier maybe susceptible to overfitting because of the noise in training data and if K is too large, then the classifier may misclassify the test sample as the data points that lie far away from the neighbourhood of the test sample may also get included in the number of nearest neighbors.

Few characteristics of KNN-

- It is instance based
- It is rote learner
- It is a non-parametric technique
- It doesn't require any training time
- It is a lazy learner.

### *F. Performance Measure Indices*

The performance of the model a confusion matrix is first computed as shown below [15]

| | Actually Positive (1) | Actually Negative (0) |
|---|---|---|
| Predicted Positive (1) | True Positives (TPs) | False Positives (FPs) |
| Predicted Negative (0) | False Negatives (FNs) | True Negatives (TNs) |

Performance is later measured using the formulas stated below-

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

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

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

$$\text{F} - \text{measure} = \frac{2 \times \text{Recall} \times \text{Precison}}{\text{Recall} + \text{Precision}}$$

## IV. EXPERIMENTS

In this section, details regarding data pre-processing on the dataset and the experimental setup are illustrated.

### *A. Data Pre-Processing*

Data is pre-processed using various techniques. To get rid of defects and redundancy present in the data, pre-processing on the dataset has been done. It is an important to pre-process the dataset and remove the redundancy before training and

testing the model. For our model, the defects in the WDBC dataset have been pre-processed and corrected.
The 'diagnosis' column give information about the class label. Out of the 32 columns in the dataset, the column named 'id' is the patient id for each patient in the dataset which do not serve any purpose in our classification process, therefore 'id' column is dropped from the dataset. There are in total of 30 attributes in the dataset used for classification. An important column i.e. 'diagnosis' column in the dataset has object data type that has the value 'M' or 'B', the values in this column are converted into integer number using LabelEncode() function available in Sklearn package that helps in assigning integer type values i.e. '0' and '1'.

Due to many attributes, a lot of redundancy dataset is observed, these attributes possess relation with other attributes which is observed by the correlation between attributes and this correlation is examined properly with the help of correlation heat map as shown in Figure 4. Eliminating these attributes causes problem as these attributes provide important information for classification process, so Principal Component Analysis (PCA) is applied to the dataset to avoid these problems.

The Variance graph as shown in Figure 5 is between variance and number of components in PCA has been examined for choosing the number of components in PCA. In this graph the variance become approximately same after around 17 number of components. Therefore 17 number of components have been used in PCA technique.

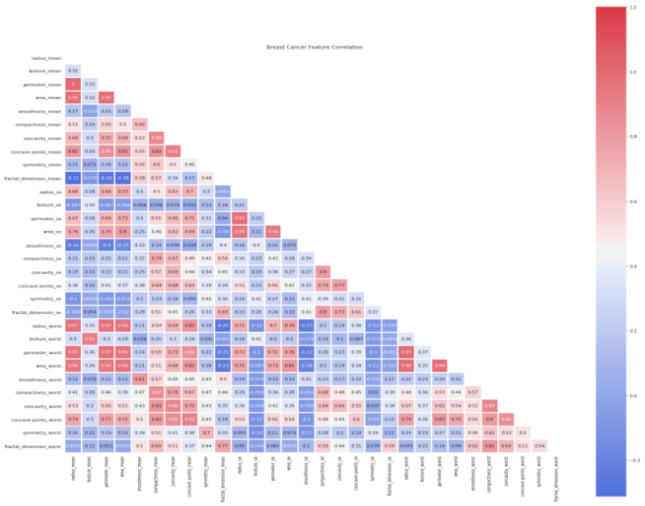


*Fig. 4 Heatmap for correlation of attributes*

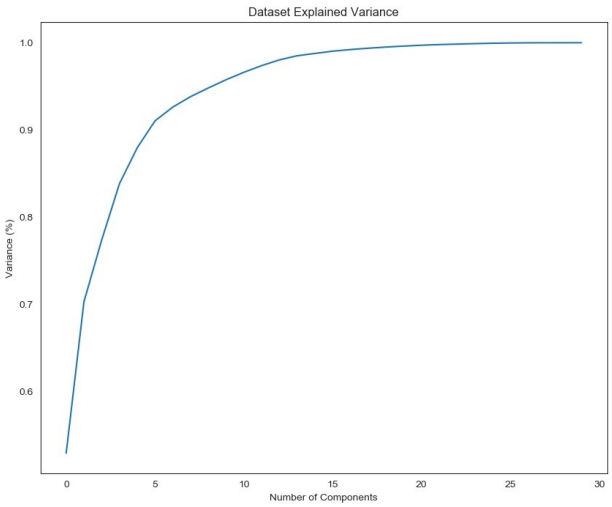



*Fig.5 Variance graph of Dataset*

## B. Experimental Setup

The specifications of the system which has been used in the implementation:

| | |
|---|---|
| CPU: | "Intel(R) Core (TM) i5-6200U CPU @ 2.30GHz" |
| RAM: | 4 GB |
| OS: | "Windows 10 64-bit" |
| GPU: | "Intel HD 520" |

Jupyter Notebook with version 6.0.1 in Anaconda version 1.9.7 is used for carrying out the experiment on Windows 10 64-bit Operating System. Specification of CPU is "Intel(R) Core (TM) i5-6200U CPU @ 2.30GHz" and specification of GPU is "Intel HD 520". Python with version 3.7.4 is used.
Various python libraries have been used such as pandas, numpy, seaborn, matplotlib and sklearn for implementing data pre-processing tasks and various machine learning algorithms.

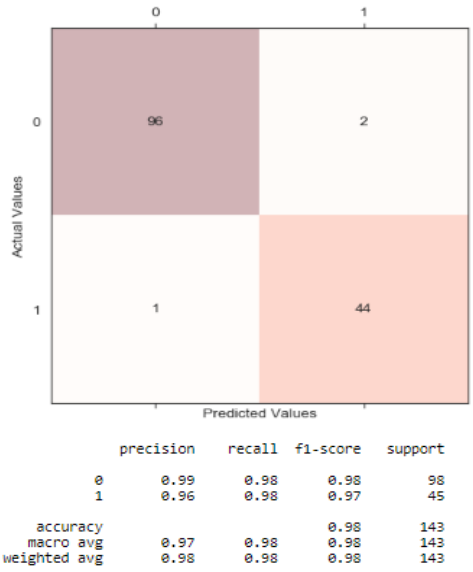
## V. RESULTS AND DISCUSSIONS

In this section, the results are evaluated of various techniques that are used in this paper- K- Nearest Neighbor, Logistic Regression and Ensemble Learning. The results are evaluated using confusion matrix. First, Principal Component Analysis (PCA) is applied on the dataset for feature extraction and the number of components used in PCA is 17 after analysing dataset explained variance graph as shown in Figure 6.

After applying PCA, 3 models (Logistic Regression, K-Nearest Neighbor and Ensemble Learning) are used for classification of tumour type i.e; benign or malignant. There are 5 machine learning algorithms that are used in Ensemble Learning are Logistic Regression, K-Nearest Neighbor, Linear Discriminant Analysis, Support Vector Classifier and Random Forest Classifier. Hard voting is used in Ensemble Learning which predict the result based on the highest majority of votes.

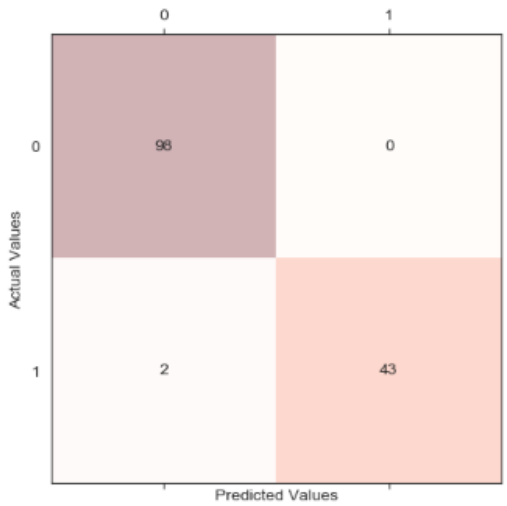
The dataset is divided into 2 parts, 75% dataset is used for training and 25% dataset is used for testing. Dataset is split using train_test_split function which is available in sklearn.model_selection package. The number of random_state which are used in train_test_split function is 7.

After analysing the results using confusion matrix, Logistic Regression gave 97.90% accuracy and confusion matrix of Logistic Regression is shown in Figure 6. K-Nearest Neighbor gave 98.60% accuracy and its confusion matrix is shown in Figure 7. Ensemble Learning technique of 5 machine learning algorithms (Logistic Regression, K-Nearest Neighbor, Linear Discriminant Analysis, Support Vector Classifier and Random Forest Classifier) gave the accuracy of 99.30%. Figure 8 shows the confusion matrix of Ensemble Learning.



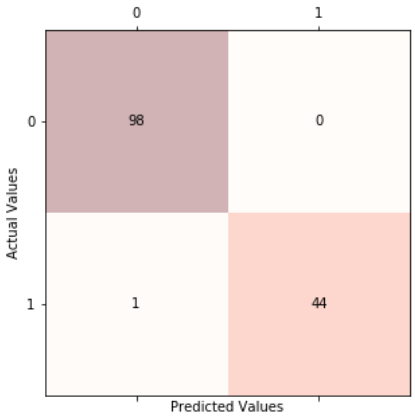

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 0.98 | 0.98 | 98 |
| 1 | 0.96 | 0.98 | 0.97 | 45 |
| accuracy | | | 0.98 | 143 |
| macro avg | 0.97 | 0.98 | 0.98 | 143 |
| weighted avg | 0.98 | 0.98 | 0.98 | 143 |

*Fig. 6 Confusion Matrix of Logistic Regression*




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 1.00 | 0.99 | 98 |
| 1 | 1.00 | 0.96 | 0.98 | 45 |
| accuracy | | | 0.99 | 143 |
| macro avg | 0.99 | 0.98 | 0.98 | 143 |
| weighted avg | 0.99 | 0.99 | 0.99 | 143 |

*Fig. 7 Confusion Matrix of K-Nearest Neighbor*




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 1.00 | 0.99 | 98 |
| 1 | 1.00 | 0.98 | 0.99 | 45 |
| accuracy | | | 0.99 | 143 |
| macro avg | 0.99 | 0.99 | 0.99 | 143 |
| weighted avg | 0.99 | 0.99 | 0.99 | 143 |

*Fig. 8 Confusion Matrix of Ensemble Learning*

Table. 1 talks about the values of various algorithms in terms of precision, recall, accuracy and F1-Score.

| "ML Algorithm" | Precision | Recall | Accuracy | F1-Score |
|---|---|---|---|---|
| **Logistic Regression** | 0.9565 | 0.9777 | 0.9790 | 0.9670 |
| **KNN** | 1.0 | 0.9555 | 0.9860 | 0.9772 |
| **Ensemble Learning** | 1.0 | 0.9777 | 0.9930 | 0.9887 |

*Table1. Performance evaluation for different algorithms*

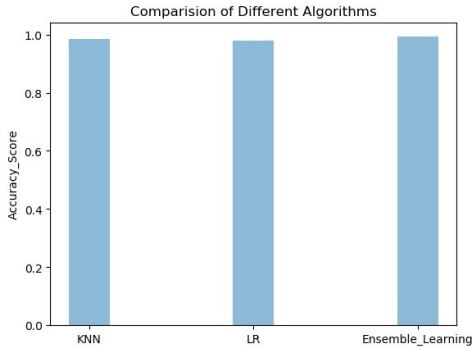



*Fig. 9. Accuracy chart of different algorithms*

Figure 9 exhibits the accuracy comparison that is obtained after applying different ML algorithms on the dataset. IT can be observed from figure 9 that Ensemble Learning has performed best and has achieved the highest accuracy as compared to the other algorithms.

## VI. CONCLUSIONS AND FUTURE SCOPE

The aim of this paper is to classify the types of tumour i.e. Benign tumour and Malignant tumour, various machine learning techniques allows us to find the most suitable model which is capable enough classify the tumour with high accuracy. Wisconsin breast cancer diagnosis (WDBC) data set is taken from UCI machine learning repository. Initially the data is pre-processed followed by applying Principal Component Analysis (PCA) with 17 components on the dataset. After which, different machine learning techniques like K-Nearest Neighbor, Logistic Regression and Ensemble Learning are applied and their results are evaluated using confusion matrix. The dataset consist of 569 instances and in future more data would be added to the database which would increase help in better training of machine learning models and would work more accurately, which will also brief us about the relationship among various attributes.

## REFERENCES

[1] WHO (World Health Organisation) Fact Sheet https://www.who.int/en/news-room/fact-sheets/detail/cancer
[2] C. Chen, "Curriculum Assessment Using Artificial Neural Network and Support Vector Machine Modeling Approaches: A Case Study," Jan, 2010.
[3] Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, and Md. Kamrul Hasan, "Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors ", 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dec 2017.
[4] Jinan Alhajjaj, Reem Alyami, Abdullah Alqahtani, Batool Alnajrani, Ilham Elaalami, Taoreed O. Owolabi , Nahier Aldhafferi, and Sunday O. Olatunji, "Investigating the effect of Correlation based Feature Selection on breast cancer diagnosis using Artificial Neural Network and Support Vector Machines", IEEE, 2017
[5] Dan C. CireSan, Alessandro Giusti, Luca M. Gambardella, Jurgen Schmidhuber," Mitosis Detection in Breast Cancer Histology Images and Deep Neural Network", International Conference on Medical Image Computing and Computer-Assisted Intervention 2013, pp 411-418
[6] B. Santhi, R. Nithya, " Classification of Normal and Abnormal Patterns in Digital Mammograms for Diagnosis of Breast Cancer", Aug. 2011
[7] G. Manikandan, B. Karthikeyan , P. Rajendiran , R. Harish , T. Prathyusha , V. Sethu, " Breast Cancer Prediction Using Ensemble Techniques", Indian Journal of Public Health Research & Development, July 2019, Vol.10, No. 7
[8] S. k. Mandal, A. Gupta and Animesh Hazra, "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms", IJCA (0975 – 8887) Volume 145 – No.2, July 2016
[9] UCI Machine Leaning Repository, Center of Machine Learning and Intelligent Systems, https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
[10] Ram Murti Rawat, Shivam Panchal, Vivek Kumar Singh, Yash Panchal, "Breast Cancer Detection Using Support Vector Machine with Principal Component Analysis", IJCRT, March 2020, Volume 8, Issue 3, ISSN: 2320-2882.
[11] H. Yusuff, N. Mohamad, U.K. Ngah, A.S.Yahaya, "Breast Cancer Analysis using Logistic Regression", International Journal of Recent Research and Applied Studies, Jan 2012, Volume 10, Issue 1, pp 14-22.
[12]Ms. Steffi Thomas, Mr. Atharva Joshi, Ms. Rutu Kalhapure, Mr. Deepraj Bhosale, Prof. Deepali Sonawane, 2020, Bionic ARM for Prosthetist, International Journal Of Engineering Research & Technology (IJERT) ICSITS – 2020 (Volume 8 – Issue 05)
[13] Voting Classifier using Sklearn https://www.geeksforgeeks.org/ml-voting-classifier-using-sklearn/
[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 eSystems Engineering (DeSE), Liverpool, 2016, pp. 35-39.
[15]Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Zanasi, A. (1998). Discovering Data Mining: From Concept to Implementation, Upper Saddle River, N.J., Prentice Hall.