IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Breast Cancer Detection Using Support Vector Machine With Principal Component Analysis

[1]Ram Murti Rawat, [2]Shivam Panchal, [3]Vivek Kumar Singh, [4]Yash Panchal

[1]Assistant Professor, Department of Computer Engineering, Delhi Technological University, Delhi
[2]Undergraduate Student, Department of Computer Engineering, Delhi Technological University, Delhi
[3]Undergraduate Student, Department of Computer Engineering, Delhi Technological University, Delhi
[4]Undergraduate Student, Department of Computer Engineering, Delhi Technological University, Delhi

Abstract: Nowadays, Breast Cancer is considered to be a major threat to women which is leading to an increase in death rates of women causing a major concern in the society regarding it. This disease may be a major threat but today medical-science is capable enough to cure such threats without causing any harm to women if detected at early stages. Detecting the cancer and differentiating between the diagnosis that affirms whether the patient has breast cancer or not is considered to be the major challenge. For accurate diagnosis of breast cancer many research paper throws light on Machine Learning Algorithms like Support Vector Machine (SVM) and Artificial Neural Network (ANN). The following paper talks about the diagnosis of breast cancer using Support Vector Machine (SVM) with Principal Component Analysis (PCA) which are supervised machine learning techniques for breast cancer detection. The proposed model is tested on Wisconsin breast cancer diagnosis data set which is selected from UCI machine learning repository. The data was pre processed which was followed by feature extraction of the data set using Principal Component Analysis (PCA), the performance of the model was diagnosed using six different machine learning algorithms namely Naive Bayes, Support Vector Machine, Classification and Regression Trees, Logistic Regression, Linear discriminant Analysis and k-nearest neighbours. The technique proposed in the paper achieved an accuracy of 99.30% by Support Vector Machine with Principal Component Analysis.

Index Terms - **Breast Cancer Detection, Support Vector Machine (SVM), Principal Component Analysis (PCA), Artificial Neural Network (ANN), Machine learning.**

I. INTRODUCTION

Cancer is one of most dangerous health issue caused to people, excess growth of cells in body leads to cancer which results in death of many people suffering from cancer. Breast cancer is a cancer that develops due to excess growth of breast tissue. Breast cancer is considered to be life threatening and is the second highest cause of deaths due to cancer after lung cancer in women. Breast cancer caused 627,000 deaths in 2018 alone with a frequency of 2.1 million as of 2018 worldwide [1]. In women it can be identified by the presence of lumps in the breast. With the growth and advancement of medical-science breast cancer can be cured without any harmful effect to the patient if detected at early stages otherwise it can be life threatening.

The major task that lies is the detection of lumps and tumours in breast, but today machine learning has made it possible for people to design certain models using different techniques in order to detect and diagnose breast cancer using supervised learning by training the model based on the past data of patients of breast cancer. Many people have worked on different machine learning algorithms in the past to develop breast cancer detection model and had achieved remarkable results with good accuracy. Many people have written research papers using several machine learning techniques, Support Vector Machine (SVM) and Artificial Neural Network (ANN) are the techniques that are been used frequently [2].

Several other work on the same field is done using the same data set as used in this model using different machine learning techniques. A study was conducted on breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms [3] which achieved an accuracy of 97.38%. Another study used SVM with Radial Basis (RBF) and Polynomial functions as kernel functions achieved 97.13% [4].
Another study was conducted by the students of University of Akron, support vector machine with linear, polynomial and RBF kernel functions were trained on the same dataset as used in this model [5]. this study was successful and achieved an accuracy of 99.29%. Furthermore, in 2017, a study was done [6] investigating the effects of Correlation based feature selection using Artificial Neural Network and Support Vector Machine which was done to diagnose and differentiate between Benign and Malignant tumour and achieved an accuracy of 97.13% and 96.70% for Support Vector Machine and Artificial Neural Network respectively.

Following this study another study was conducted in 2017 using support vector machine and k-nearest neighbours using 10-fold cross validation to obtain accurate results achieving accuracy of 98.57% and 97.14% for support vector machines and k-nearest neighbours respectively.

The rest of the research paper is structured into various sections. The related work that has been done in the same field highlighted in section II. Theoretical explanation of various machine learning techniques and details regarding the data set of our model is discussed in Section III. Section IV consists of the details of the pre-processing stage and the experimental setup. Result analysis of certain machine learning algorithms and their comparison is illustrated in Section V. Section VI concludes our work.

II. RELATED WORK

In past few decades several different techniques have been evolved to improve the accuracy to of breast cancer detection and diagnosis. The related work has been discussed below briefly.

Dan C. CireSan, Alessandro Giusti, Luca M. Gambardella, Jurgen Schmidhuber proposed a model for detection of mitosis in breast cancer with deep neural networks [7], they've used max pooling convolutional neural network for detecting mitosis in breast. The networks were trained to be capable enough to classify each pixel in the given image, following simple post processing to neural output.

In last decade, computerized tool plays a vital role in diagnosis and detection of breast cancer. R. Nithya and B. Santhi proposed a method for diagnosis in digital mammograms with the help of Grey Level Co-occurrence Matrix (GLCM) features [8]. Mammography is considered to be an important tool which is used for early detection of breast cancer. In this model mammogram images are classified into cancer and normal images and the effectiveness is determined using accuracy, specificity and sensitivity. Accuracy obtained by this model was 96%.

Nor Ashidi Mat Isa, Fadzil Ahmad, Siti Noraini Sulaiman, and Zakaria Hussain successfully proposed a technique called GA-MOO-NN which was used for the detection of breast cancer. This model divided the data set into training, testing and validation parts. The algorithm successfully attained an accuracy of 98.85% in best case and 98.10% in average case. This model also talks about the comparison of this model with other existing models. H. AttyaLafta, N. KdhimAyoob and A. Hussein proposed a technique for detection of breast cancer using a hybrid neuro genetic framework comprising of Training Feed Forward Back Propagation and Genetic Algorithm [9]. The overall accuracy can out to be 97%. Following with a comparison is made by the authors.

III. BACKGROUND

3.1 DATASET

Wisconsin Diagnostic Breast Cancer (WDBC) Dataset from the UCI repository was used by us in our proposed model. This dataset was provided by Dr. William H. Wolberg of the University of Wisconsin in 1995. The dataset has 569 instances in total consisting of 212 malignant and 357 are benign i.e. 37.26% malignant and 62.74% benign. The WDBC dataset has 32 attributes.

For each cell nucleus, 10 aspects (features) were taken into consideration. These 10 features were: radius(average of distances from the center to points on the boundary), texture(variance of gray-scale intensities), perimeter(complete boundary length of cell nucleus), smoothness(local variation in radius lengths), area(number of pixels), compactness(compactness of the cell was measured by combining the perimeter and area -perimeter^2 / area - 1.0), concave points(number of boundary points lying on the concave regions of the boundary), concavity(severity of concave portions of the contour), symmetry, and fractal dimension(used "coastline approximation").

A digitized image of a fine needle aspirate (FNA) of a breast mass was used to compute all these features. These aspects describe the characteristics of cell nuclei present in the image. The, standard error, mean value and extreme value (worst or largest) of each characteristic were computed for each image. This resulted in 30 features for each of the 569 instances. All the 30 features are of float type.

Dataset –WDBC dataset
Provided by- Dr. William H. Wolberg
Number of attributes-32
Number of instances-569 (212 malignant, 357 benign)
Dataset characteristics- Multivariate

3.2 PCA

When many attributes capture the data, most of them are usually redundant. These redundant attributes only add noise to the data and add computational time to the model thereby may reduce the accuracy of our model. Dimensionality reduction help in increasing model interpretability, faster training time, reducing noise, reduction in overfitting models.

PCA or Principal Component Analysis is considered as a dimensionality reduction method. PCA is helps in finding a mapping from inputs in original d-dimensional space to a new k (k<d) dimensional space, with minimum loss of information and data. PCA on the other hand is considered to be a powerful tool for analyzing data of a dataset. PCA helps you to identify patterns in data and expresses data in a way highlighting their similarities and differences. PCA is an unsupervised method used to find the interdependence among a set of variables, it reduces the number of dimensions without much loss of information.

Steps in PCA –

- Get some data
- Subtract mean
- Calculate the covariance matrix
- Calculate Eigenvector and Eigen values of the covariance matrix
- Choose components and form features vectors
- Derive new data set

3.3 SVM

SVM or Support Vector Machine is a supervised machine learning algorithm. It can be used to perform classification tasks as well as regression. However, generally it is used in classification tasks.

A SVM finds a hyperplane in an N-dimensional space that best segregates the classes of data points. Hyperplanes are basically lines or planes which help in classifying the data points, for an n-dimensional space, the hyperplane is (n-1) dimensional. SVM's objective is to find the maximal margin hyperplane. The maximal margin hyperplane is the distance from the **hyperplane** to the closest points in either class. We use support vectors to maximize the margin. Support vectors are the closest data points to the hyperplane. These support vectors or data points influence the orientation and position of hyperplane, thus they play a major role in constructing the maximal margin hyperplane and thus the SVM model.

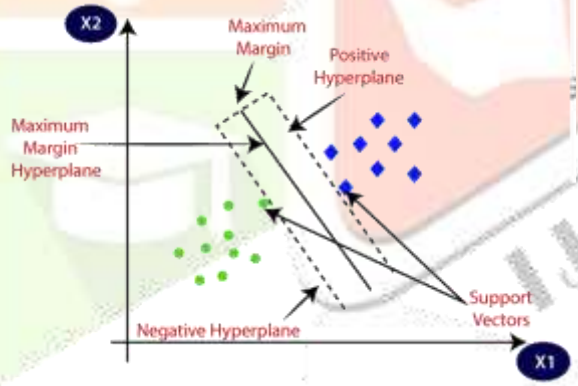

Figure 1 Maximal margin hyperplane in SVM

Advantages of SVM

- Works effectively in high dimensional spaces.
- Memory efficient.
- Since uses different kernel functions, it is versatile.
- Good generalization performance.

Disadvantages

- Use expensive multi-fold cross-validation.
- Feature of duality.
- Not efficient for large data sets.

SVM algorithms use wide range of kernel functions like linear, non linear, polynomial, RBF (radial basis function) and sigmoid. To classify a non linear dataset, the dataset is projected to a higher dimension using kernel functions in which they become linearly separable.

3.4 PERFORMANCE MEASURE INDICES

To measure the performance of the model a confusion matrix is first computed as shown below

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

Figure 2 Confusion matrix

Performance of the system is then measured using the following formulas-

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

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

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

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

IV. EXPERIMENTS

In this section, details of data pre-processing along with the specifications of the experimental setup are provided.

4.1 Data Pre-Processing

Various techniques for data pre-processing are performed. Dataset is pre-processed to get rid of redundancy and defects present. It's necessary to pre-process the dataset before training and testing. For this, some defects within the WDBC dataset are processed and corrected.

The dataset contains 569 records. The dataset does not contain any missing values. The dataset is stored in the excel sheet format that contains 32 e.g. (id, diagnosis, radius_mean, texture_mean, perimeter_worst, area_worst). The 'diagnosis' column tells about the class label. Out of these 32 columns, there is a column whose name is 'id' which is generally the patient id for each patient record in the dataset and it does not serve any purpose in our classification process. So this column has been drop from our dataset. As there are 30 attributes in dataset which are used for classification. The 'diagnosis' column has an object data type that contains the value of either 'M' or 'B' which is converted into integer number using LabelEncode() function provided by Sklearn package that assigns value in the form of '0' and '1' which are of integer data type.

There are so much redundancy exists in this dataset as there are many attributes in this dataset that have some relation with other attributes which is found with the help of finding the correlation between attributes and correlation between attributes have been examined with the help of drawing heat map of correlation between attributes as shown in Figure 3. But we could not eliminate those attributes because they also provide important information for classification process. Therefore Principal Component Analysis (PCA) technique have been applied on the dataset which is generally a statistical procedure to convert the set of input variables into a set of linearly uncorrelated variables using orthogonal transformation called principal components. The number of components used in PCA is 17 which is used after examining the Dataset Explained Variance graph as shown in Figure 4 between variance and number of components in PCA. In this graph after around 16 number of components the variance become approximately same. Therefore number of components which have been used in PCA technique is 17.

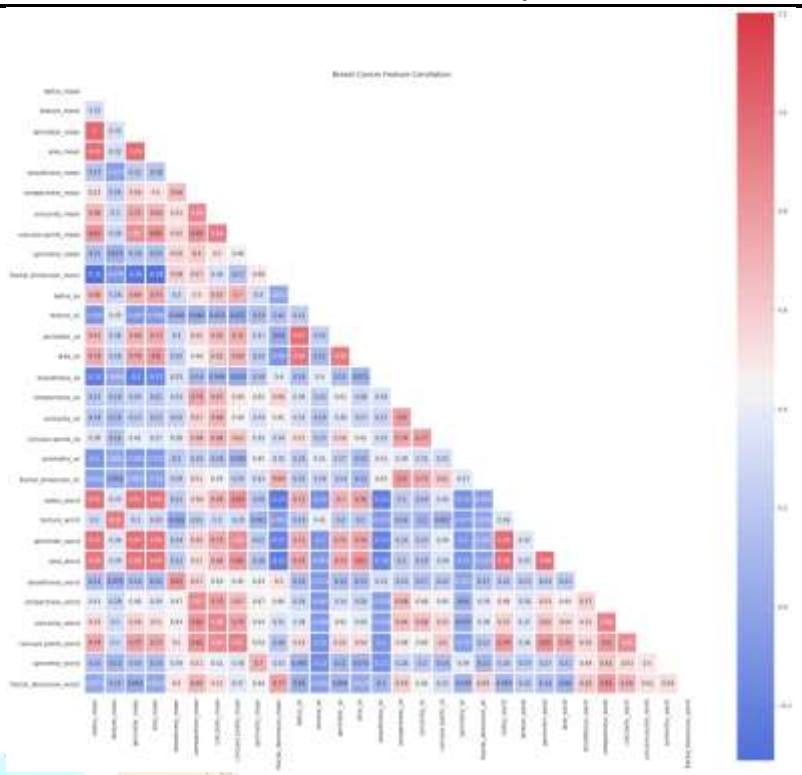

Figure 3 Heatmap of correlation between attributes

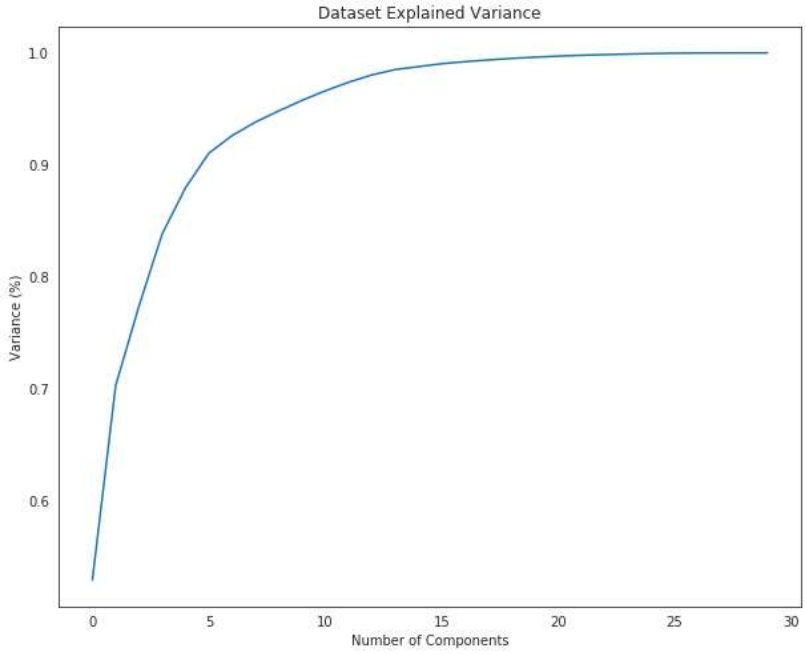

Figure 4 Dataset Explained Variance graph

4.2 Experimental Setup

The technical specifications of the computer 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 6.0.1 on Anaconda 1.9.7 with Python 3.7.4 is used for implementing data pre-processing steps and various machine learning algorithms. Various python libraries are used: pandas, numpy, seaborn, matplotlib and sklearn.

V. RESULTS AND DESCUSSIONS

In this section, results of study carried out are presented. The performance evaluation criterion is based on the values of the accuracy score of different algorithms.

We have first used Principal Component Analysis (PCA) for feature extraction and we have taken 17 principal components for Principal Component Analysis (PCA) after analysing the relation between variance and number of components in Dataset Explained Variance graph as shown in Figure 4. After applying PCA we have used 6 machine learning algorithms, CART (Decision Tree), Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbors and Linear Discriminant Analysis. Out of these six algorithms (CART, Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbors and Linear Discriminant Analysis), Support Vector Machine gave the highest accuracy of 99.30% and highest F1-Score of 0.98877. Table 1 exhibits the value of precision, recall, accuracy and F1-Score of various algorithms.

Table1. (Performance evaluation data for various algorithms)

"ML Algorithm"	Precision	Recall	Accuracy	F1-Score
Naïve Bayes	0.8333	0.8888	0.9090	0.8602
Support Vector Machine	**1.0**	**0.9777**	**0.9930**	**0.9887**
CART	0.8235	0.9333	0.9160	0.8749
Logistic Regression	0.9565	0.9777	0.9790	0.9670
Linear Discriminant Analysis	1.0	0.8888	0.9650	0.9411
KNN	1.0	0.9555	0.9860	0.9772

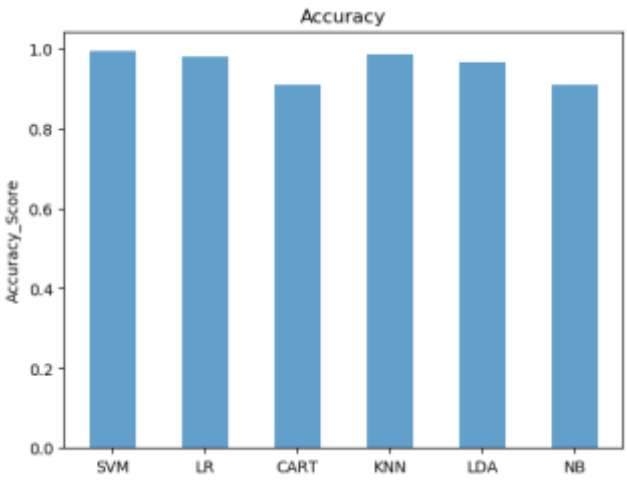

Figure 5 Different algorithm's accuracy comparison)

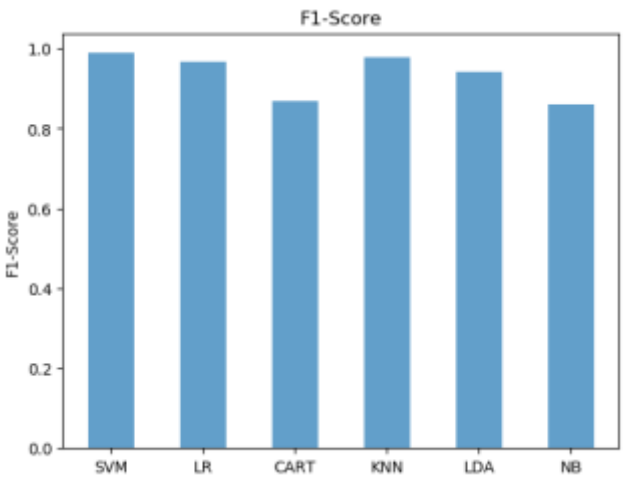

Figure 6 Different algorithm's F1-Score's comparison

Figure 5 and figure 6 exhibits the accuracy comparison and F1-Score comparison respectively that are obtained after applying various machine learning algorithms. It is evident from Fig.5 and Fig.6 that Support Vector Machine has performed best as compared to the other algorithms, that are applied after the data pre-processing and Principal Component Analysis.

VI. CONCLUSION AND RECOMMENDATION

The aim of the study is to implement task for classifying the types of tumours i.e; Malignant and Benign, by use of machine learning techniques to find the best model which can classify the types of tumours accurately. Wisconsin breast cancer diagnosis data set which is selected from UCI machine learning repository. First, the data has been pre-processed and after that Principal Component Analysis (PCA) has been applied with 17 number of components on the dataset. Finally, six machine learning techniques, CART (Decision Tree), Support Vector Machine, Naïve Bayes Logistic Regression, K-Nearest Neighbors and Linear Discriminant Analysis have been used and compared them on the basis of their accuracy. Support vector machine showed the highest accuracy of 99.30%. Further experiments can also be done in future as the size of the current dataset which is available, is small as it has only 569 instances. Therefore, the model shall be evaluated on large dataset for knowing more about the relationship between various attributes. Also the effect of different feature extraction methods could be investigated in future researches.

REFERENCES

[1] World Health Organization 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] Bichen Zheng, Sang Won Yoon, Sarah S. Lam, "Breast Cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms", 2014.

[4] S. V. G. Reddy, K. T. Reddy, V. V. Kumari, and K. V. Varma, "An SVM Based Approach to Breast Cancer Classification using RBF and Polynomial Kernel Functions with Varying Arguments," vol. 5, no. 4, pp. 5901–5904, 2014.

[5] Mihir Sewak, Priyanka Vaidya, Chien-Chung Chan, and Zhing-Hui Duan, "SVM Approach to Breast Cancer Classification", 2007

[6]Reem Alyami, Jinan Alhajjaj, Batool Alnajrani, Ilham Elaalami, Abdullah Alqahtani, Nahier Aldhafferi, Taoreed O. Owolabi , and Sunday O. Olatunji, "Investigating the effect of Correlation based Feature Selection on breast cancer diagnosis using Artificial Neural Network and Support Vector Machines", 2017

[7] Dan C. CireSan, Alessandro Giusti, Luca M. Gambardella, Jurgen Schmidhuber, "Mitosis Detection in Breast Cancer Histology Images with Deep Neural Network", 2013

[8] R. Nithya, B. Santhi, " Classification of Normal and Abnormal Patterns in Digital Mammograms for Diagnosis of Breast Cancer", Aug. 2011

[9] H. AttyaLafta, N. KdhimAyoob and A. A. Hussein, "Breast cancer diagnosis using genetic algorithm for training feed forward back propagation," 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT), Baghdad, pp. 144149, 2017.