# Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms

Animesh Hazra
Computer Science & Engineering Department, Jalpaiguri Govt. Engg. College , Jalpaiguri, West Bengal ,India

Subrata Kumar Mandal
Information Technology Department , Jalpaiguri Govt. Engg. College , Jalpaiguri, West Bengal ,India

Amit Gupta
Information Technology Department, Jalpaiguri Govt. Engg. College , Jalpaiguri, West Bengal ,India

## ABSTRACT

Breast cancer is one of the second leading causes of cancerdeath in women. Despite the fact that cancer is preventable and curable in primary stages, the huge number of patients are diagnosed with cancer very late. Conventional methods of detecting and diagnosing cancer mainly depend on skilled physicians, with the help of medical imaging, to detect certain symptoms that usually appear in the later stages of cancer [1]. The objective of this paper is to find the smallest subset of features that can ensure highly accurate classification of breast cancer as either benign or malignant. Then a comparative study on different cancer classification approaches viz. Naïve Bayes, Support Vector Machine and Ensemble classifiers is conducted where the time complexity of each of the classifier is also measured. Here, Naïve Bayes classifier is concluded as the best classifier with lowest time complexity as compared to the other two classifiers.

## General Terms

Breast Cancer, Classification Accuracy, Feature Selection, Feature Extraction.

## Keywords

Supervised machine learning, benign, cancer classification, malignant.

## 1. INTRODUCTION

### 1.1 Preliminaries

In breast cancer, cancer cells form in the tissues of the breast of the woman [2]. The breast is made up of lobes containing 15 to 20 sections and ducts. The most common type of breast cancer begins in the cells of the ducts. Cancer that starts in the lobes or lobules found in both breasts are other types of breast cancer. Warm, red, and swollen breast is an indicator for breast cancer. Age and health history can affect the risk of developing breast cancer [3]. For detecting the different stages of the breast cancer, Chest X-ray, CT scan, Bone scan and PET scans are widely used. The number of breast cancer diagnosis is calculated to be 1.2 million among women every year according to projections by the World Health Organization. In the year 2006 an estimate of 214,460 new cancer diagnosis was made and total death of at least 41,000 occurred within the US [4]. Since the early years of cancer research, biologists have used the traditional microscopic technique to assess tumor behavior for breast cancer patients [5]. For the diagnosis and treatment of cancer, precise prediction of tumors is critically important. Latest machine learning techniques are increasingly being used by biologists to obtain proper tumor information from the databases. Among the existing techniques, supervised machine learning methods are the most popular in cancer diagnosis.

### 1.2 Basic Concepts used in Cancer Cell Detection

In this research paper Principal Component Analysis for feature extraction, Pearson Correlation Coefficient for feature selection and Naïve Bayes, Support Vector Machine (SVM) and Ensemble classifiers are used for cancer classification. These concepts are discussed as follows:

#### *1.2.1 Principal Component Analysis (PCA)*

It is a feature extraction technique which takes an orthogonal transformation to convert a set of observations of possibly correlated parameters into a set of values of linearly uncorrelated parameters called principal components [6].

#### *1.2.2 Pearson's Correlation Coefficient*

It is a feature selection procedure which is used to measure the strength of a linear association between two variables, where the value of correlation coefficient r = 1 implies a perfect positive correlation and the value r = -1 implies a perfect negative correlation. Correlation between sets of data is a measure of how perfect they are related to each other. The most common measurement of correlation in Statistics is the Pearson Correlation Coefficient. The coefficient value lies between -1 and 1 [7].

#### *1.2.3 Naïve Bayes Classifier*

Bayesian classifiers are the example of statistical classifiers. They can predict class membership probabilities such that the probability of a given tuple falls into a particular class [8].Bayes' theorem is the very basic of Bayesian classification.

#### *1.2.4 Support Vector Machine (SVM) Classifier*

A method for the classification of both linear and nonlinear data. In a brief, an SVM is an algorithm that works as follows. SVM transform the original training data into a higher dimension using nonlinear mapping. Within this new dimension, it searches for the linear optimum separating hyper-plane to differentiate the tuples among the sets. With an appropriate nonlinear mapping to an adequate high dimension, data from two sets can always be separated by a hyper-plane. The SVM finds this hyper-plane with the help of support vectors ("essential" training tuples) and margins (defined by the support vectors) [9].An unlimited number of separating lines that could be drawn here. The target is to identify the "best" one which will have the minimum classification error on preceding unseen tuples.

#### *1.2.5 Ensemble Classifier*

An ensemble classifier combines a series of k learned models (or base classifiers), $M_1$, $M_2$,...,$M_k$, with the aim of designing an improved hybrid classification model, M*. D is the given data set which is used to create k training sets, $D_1, D_2, ..., D_k$, where $D_i$ ($1 \leq i \leq k-1$) is also used to generate the classifier $M_i$.Given a new data tuple to classify, each of the base classifiers vote by returning a class prediction. Based on the votes of the base classifiers an ensemble returns a class prediction. An ensemble classifier can predict more accurate result than its base classifiers [10].

## 2. LITERATURE SURVEY

In the paper [11] by Sau Loong Ang et al. attempts were made to improve the Naive Bayes by introducing links or associations between the features such as the Tree Augmented Naive Bayes (TAN). In this study, they had shown the accuracy of a General Bayesian Network (GBN) applied with the hill-climbing learning approach, which did not impose any restrictions on the structure and represented the dataset in a better way. To measure the performance of GBN against the Naive Bayes and TAN, they used seven nominal datasets with the absence of missing values for comparative purposes. These nominal datasets were taken from the UCI Machine Learning Repository (Lichman, 2013) and they were fed into the Naive Bayes, GBN and TAN for classification with ten-fold cross validation in WEKA software using 286 instances each containing 10 attributes. Naïve Bayes model gave an accuracy of 71.68% followed by 69.58% for TAN and 74.47% for GBN.

In the paper [12] by K. Shivakami breast cancer prediction was done using DT-SVM Hybrid Model. This study was performed using the Wisconsin Breast Cancer Dataset (WBCD) taken as input from UCI machine learning repository (UCI Repository of Machine Learning Databases). The dataset contained 699 instances taken from needle aspirates from patients' breasts, of which 458 cases belonged to benign class and the remaining 241 cases belonged to malignant class. It should be noted that there were 16 instances which had missing values. In this study all the missing values were replaced by the mean of the attributes. Each record in the database had nine attributes. These nine attributes were found to differ significantly between benign and malignant samples. In case of DT-SVM the accuracy obtained was 91% with an error rate of 2.58%. Other classification algorithms had also been applied like IBL, SMO and Naïve Bayes. For IBL the accuracy obtained was 85.23% with an error rate of 12.63%. For SMO the accuracy was 72.56% with an error rate of 5.96%. For Naïve Bayes the accuracy obtained was 89.48% with an error rate of 9.89%. So this comparative study revealed that DT-SVM performed well in classifying the breast cancer data compared to all other algorithms.

In the paper [13] by Shweta Kharya et al. the core objective was to develop a probabilistic breast cancer prediction system using Naive Bayes Classifiers which can be used in making expert decision with highest accuracy. The system may be implemented in remote areas like countryside or rural regions, to imitate like human diagnostic expertise for treatment of cancer disease. The system is user friendly and reliable as model was already developed. For training Wisconsin Datasets containing 699 records with 9 medical attributes was used. For Testing 200 records were taken. This dataset had almost 65.5% benign cases and remaining 34.5% malignant cases. The accuracy was found to be 93%.

In the paper [14] by G. Ravi Kumar et al. the data set consisted of 699 patient's records of which 499 were considered for training and 200 for testing purposes. Among them, 241 or 34.5% were reported to have breast cancers while the remaining 458 or 65.5% were non-cancerous. In order to validate the prediction results of the six popular data mining techniques the 10-fold crossover validation was used. The k-fold crossover validation was usually used to reduce the error coming from random sampling to compare the accuracies of a number of prediction models. The entire set of data was randomly divided into k folds with the same number of instances in each fold. The training and testing were performed for k times and one fold was selected for further testing while the rest were selected for further training. The present knowledge distributes the data into 10 folds where 1 fold was used for testing and 9 folds were used for training purpose in the 10-fold crossover validation. Here by applying Naïve Bayes algorithm on testing data an accuracy of 94.5% had been obtained. Same result had been obtained for SVM.

In the paper [15] by C.D. Katsis et al. the proposed methodology used a Correlation Feature Selection (CFS) procedure to rank the extracted different features and an Artificial Immune Recognition System (AIRS) classifier in order to support breast cancer diagnosis. To evaluate the methodology, data had been gathered arising from 53 subjects out of 4726 cases. The specific topics expressed lesions that were not highly suggestive of benignity or malignancy when evaluated on all modality used. In every case biopsy was conducted and the biopsy results were used as golden standard to validate the methodology. The constructed dataset consisted of the features as well as the biopsy results (malignancy or benignity) for all 53 subjects. In the University Hospital of Ioannina, Greece, all data were collected. SVM technique gave an accuracy of 70.00+6.33 % considering the full set of features and an accuracy of 68.92+6.97 % considering the subset of CFS selected features.

This paper [16] by Gouda I. Salama et al. presented a comparison among the different classifiers decision tree (J48), Naive Bayes (NB), Multi-Layer Perception (MLP), Sequential Minimal Optimization (SMO) and Instance Based for K-Nearest neighbor (IBK) on three very popular different databases of breast cancer (Wisconsin Breast Cancer (WBC),Wisconsin Prognosis Breast Cancer (WPBC) and Wisconsin Diagnosis Breast Cancer (WDBC)) by using confusion matrix and classification accuracy based on 10-fold cross validation method. They introduced a fusion at classification level between these classifiers to get the most appropriate multi-classifier method for each data set. The experimental results showed that in the classification using fusion of J48 and MLP with the PCA was superior to the other classifiers using WBC data set. The PCA was used in WBC dataset as a features reduction transformation method which combined a set of correlated features. An accuracy of 92.97% was achieved using Naïve Bayes as classifier.

In the paper [17] by Kim W et al. SVM technique was used on breast cancer dataset consisting of 679 records. The types of data were clinical, pathologic and epidemiologic. The accuracy obtained was 99% considering the feature local invasion of tumour.

In the paper [18] by Mehmet Fatih Akay SVM with feature selection was used to diagnose the breast cancer. For training and testing experiments the WDBC dataset has been taken from the University of California at Irvine (UCI) machine learning repository .It was spotted that the proposed method produced the highest classification accuracies (99.51%, 99.02% and 98.53% for 80–20% of training-test partition, 70–30% of training-test partition and 50–50% of training-test partition respectively) for a subset that carried five features.

Also, other measures such as the sensitivity, specificity, confusion matrix, negative predictive value and positive predictive value and ROC curves were used to show the performance of SVM with feature selection.

In this paper [19] by Diana Dumitru the Naive Bayes classifier was applied to the Wisconsin Prognostic Breast Cancer (WPBC) dataset, containing a number of 198 patients and a binary decision class: non-recurrent-events having 151 instances and recurrent-events having 47 instances. The testing diagnosing accuracy, that was the main performance measure of the classifier, was about 74.24%, in compliance with the performance of other well-known machine learning techniques.

In this paper [5] by Daniele Soria et al. a comparison of three different classifiers in machine learning was presented, namely the Naive Bayes algorithm, the Multilayer Perceptron function and the C4.5 decision tree. C4.5 algorithm developed by Ross Quinlan,is used to generate a decision tree. C4.5 is an extension of Quinlan’s earlier ID3 algorithm. The decision trees created by C4.5 can be used for classification purpose and for this reason C4.5 is often referred to as a statistical classifier [20]. A Multilayer Perceptron is a feed forward artificial neural network model which maps sets of input data onto a set of proper output. It is a moderation of the standard linear perceptron where it uses three or more layers of neurons i.e., nodes with nonlinear activation functions and is more powerful than the perceptron in which it can differentiate data that is not linearly separable or separable by a hyper plane. The study was motivated by the necessity to detect an automated and robust method to validate their previous classification of breast cancer markers. They had, in fact, obtained six classes using agreement between different clustering algorithms. Starting from these groups they wanted to replicate the classification keeping into account the high non-normality of used data. For this reason they started using the C4.5 and the Multilayer Perceptron classifiers and then they compared results with the Naïve Bayes. Surprisingly, it was found that when the dataset was reduced to ten markers, the Naive Bayes classifier performed better than the C4.5. The number of instances taken was 663. An accuracy of 93.1% was obtained using 10 markers and this accuracy became 86.9% using 25 markers.

The objective of this paper [4] by Haowen You et al. was to provide a comparative analysis on the utilized potential classification tools (back-propagation neural network, linear programming, Bayesian network and support vector machine) on the problem by a benchmark dataset which consisted of numeric cellular shape features extracted from pre-processed Fine Needle Aspiration biopsy image of cell slides. The benchmark dataset in this research was obtained from the UCI machine learning repository classified data as malignant (M) or benign (B). The dataset was composed of a total of 569 observations with benign and malignant cases being 357 and 212 observations respectively. Each of the dataset in the observation was composed of 30 variables and 10 of the featured variables were related to the aforementioned characteristics. Here Naïve Bayes classifier gave an accuracy of 89.55%.

## 3. PROPOSED METHODOLOGY

Today's real-world databases are highly vulnerable to noisy, missing and inconsistent data due to their typically massive size and their likely origin from multiple, miscellaneous sources. Hence data preprocessing is a necessary phase for classification purposes. Data preprocessing includes data cleaning, data dimensionality reduction, data transformation (data normalization, data binning) followed by classification.

Here WDBC breast cancer dataset has been taken from UCI machine learning repository [21] as an input data. This WDBC dataset contains 569 instances and 32 attributes of which 300 instances have been taken for training purpose and 269 instances for testing purpose. These testing data are applied over three classification methods which detect whether the cell is malignant or benign.

Here, the data cleaning technique includes removing the missing values if present, with the mean of the attributes. Data normalization brings the range of all attribute values between 0 and 1.The following workflow diagram represents breast cancer cell detection using Pearson Correlation Coefficient as a feature selection technique.

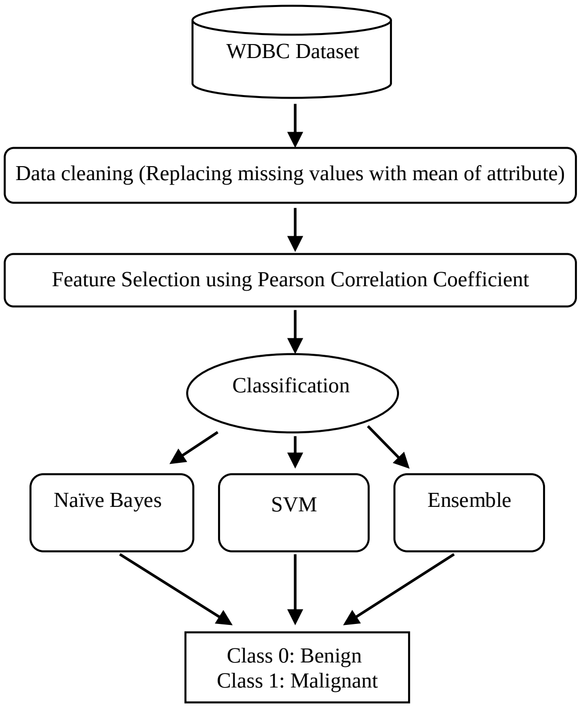
WDBC Dataset
Data cleaning (Replacing missing values with mean of attribute)
Feature Selection using Pearson Correlation Coefficient
Classification
Naïve Bayes
SVM
Ensemble
Class 0: Benign
Class 1: Malignant

**Fig 1: Workflow diagram for breast cancer cell detection using Pearson Correlation Coefficient.**

In Figure 1 at first data cleaning technique has been applied. After that a feature selection technique was implemented on the normalized dataset using Pearson Correlation Coefficient which reveals how much the attributes of the dataset are related to the class attribute and based on that a ranking of the features has been obtained. Here, five features are considered according to the descending order of their ranks considering the threshold value of correlation as 0.74. These five features for 569 instances have been taken and applied over three classification techniques viz. Naïve Bayes, Support Vector Machine and Ensemble.

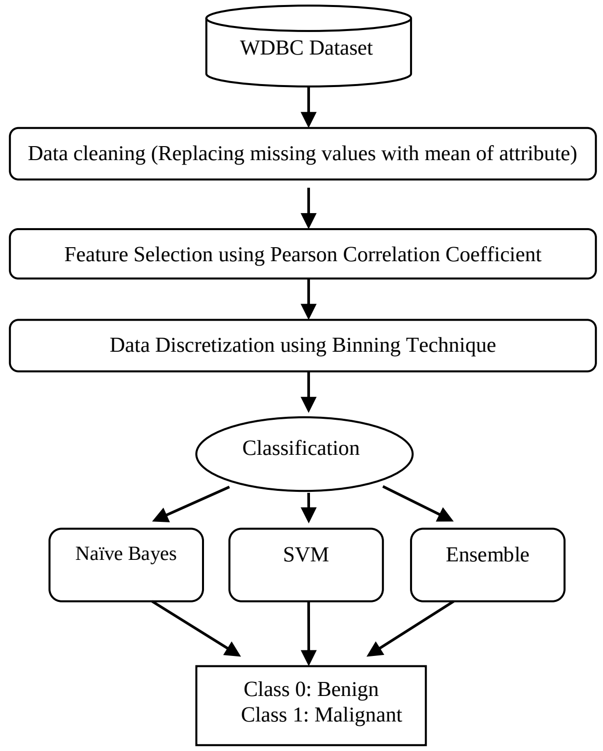
WDBC Dataset
Data cleaning (Replacing missing values with mean of attribute)
Feature Selection using Pearson Correlation Coefficient
Data Discretization using Binning Technique
Classification
Naïve Bayes
SVM
Ensemble
Class 0: Benign
Class 1: Malignant

**Fig 2: Workflow diagram for breast cancer cell detection using Pearson Correlation Coefficient with binning concept.**

In Figure 2 one more technique before classification phase is considered which is binning. The preprocessed data undergoes binning where the entire range of values of each attribute is divided into three bins. To prove how much this binning technique is appropriate, the concept of entropy for each attribute is implemented, conditional entropy of decision attribute for a given attribute and a metric named level of consistency (LOC). Here, the value of LOC is 0.98 at the first level of discretization which is almost close to one. It implies that no more bins are required for any attribute. So the procedure is stopped here. Now the resultant dataset after binning is applied to the classification phase. The result of each classification techniques are observed by assigning lower bounds of each bin to the data values corresponding to that bin and also by assigning upper bounds of each bin to the data values similarly.

Now in Figure 3, the classification accuracy with feature extraction instead of feature selection is observed. PCA is used for feature extraction and mapped the data into a lower dimensional space (here five dimensional space have been taken). Now the result of PCA is used as an input for three classification techniques i.e., Naïve Bayes, Support Vector Machine and Ensemble.

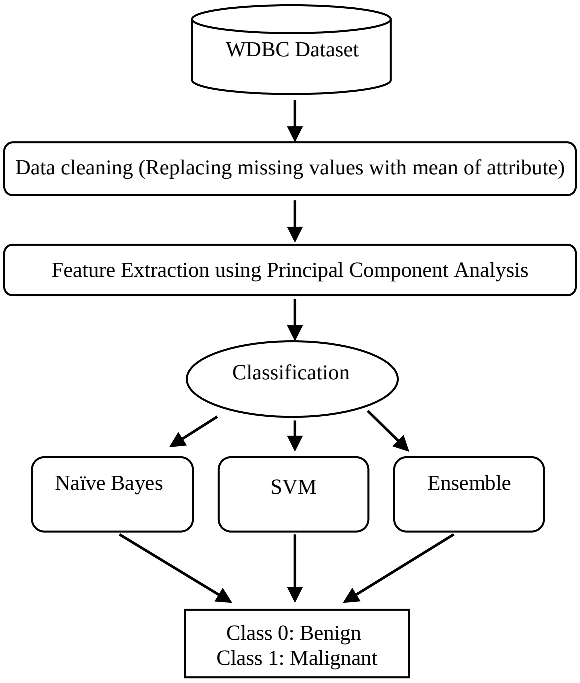
WDBC Dataset
Data cleaning (Replacing missing values with mean of attribute)
Feature Extraction using Principal Component Analysis
Classification
Naïve Bayes
SVM
Ensemble
Class 0: Benign
Class 1: Malignant

**Fig 3: Workflow diagram for breast cancer cell detection using Principal Component Analysis.**

## 4. RESULT AND DISCUSSIONS

In this paper a comprehensive study on different classification techniques have been conducted and provided a basis for comparison among them in terms of accuracy percentage and time complexity. The level of effectiveness of the classification model is calculated by using confusion matrix.

Figure 4 represents a plot of number of combined features at each step Vs. classification accuracy using Naïve Bayes classifier, taking the preprocessed data after binning. Here five most dominant features have been considered which is obtained using Pearson Correlation Coefficient concept and the results are observed by taking first dominant feature at first, then taking first two dominant features, after those first three dominant features and continue this procedure until the combination of five features. The maximum classification accuracy percentage obtained over here is 97.3978%.

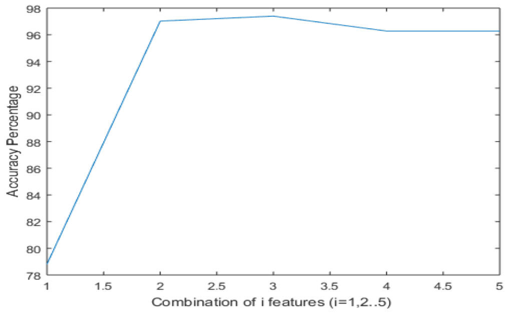

**Fig 4: Classification accuracies by taking five most dominant features after binning and applying Naïve Bayes algorithm to features combined increasingly at each step.**

Figure 5 represents a plot of number of combined features at each step Vs. classification accuracy using Support Vector Machine taking the preprocessed data after binning. Here, five most dominant features are considered and the results are observed by taking first dominant feature at first, then taking first two dominant features, after those first three dominant features and so on up to combination of five features. The maximum classification accuracy percentage obtained over here is 97.3978%.

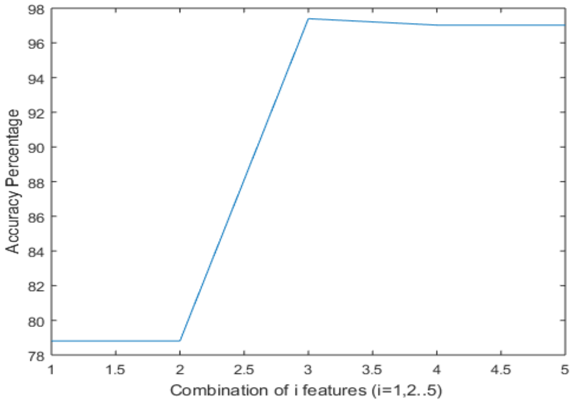

**Fig 5: Classification accuracies by taking five most dominant features after binning and applying Support Vector Machine to features combined increasingly at each step.**

Figure 6 represents a plot of number of combined features at each step Vs. classification accuracy using Ensemble classifier taking the preprocessed data after binning. Here, five most dominant features are considered and the results are observed by taking first dominant feature at first, then taking first two dominant features, after those first three dominant features and continue this procedure until the combination of five features. The maximum classification accuracy percentage obtained over here is 97.3978%.

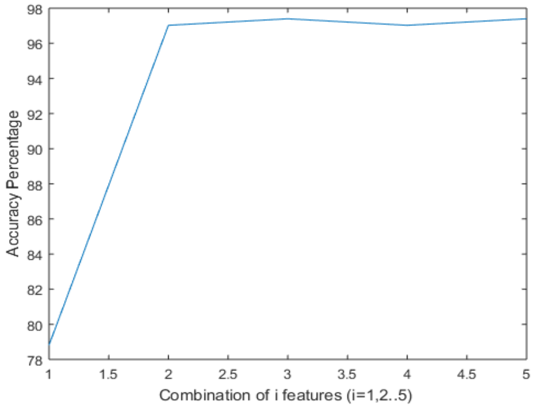

**Fig 6: Classification accuracies by taking five most dominant features after binning and applying Ensemble algorithm to features combined increasingly at each step.**

Figure 7 represents a plot of number of combined features at each step Vs. classification accuracy. Here, 19 most dominant features are considered and the results are observed by taking first dominant feature at first, then taking first two dominant features, after those first three dominant features and so on up to combination of 19 features. Using these features, SVM is applied for classifying the dataset without binning and again it is applied after data binning.

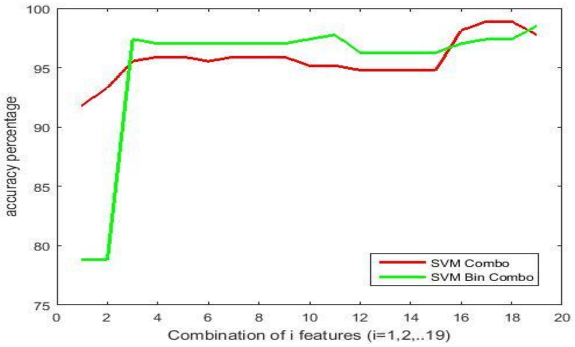

**Fig 7: Classification accuracies by taking nineteen most dominant features with binning and without binning and applying Support Vector Machine to features combined increasingly at each step.**

The following Figure 8 represents a plot of number of combined features at each step Vs. classification accuracy using Naïve Bayes, Support Vector Machine and Ensemble classifiers. It takes the preprocessed data obtained by applying Principal Component Analysis. The classification accuracies obtained from Naïve Bayes, SVM and Ensemble algorithms are 95.1673%, 95.5390% and 95.9108% respectively. These results are shown in Table 1.

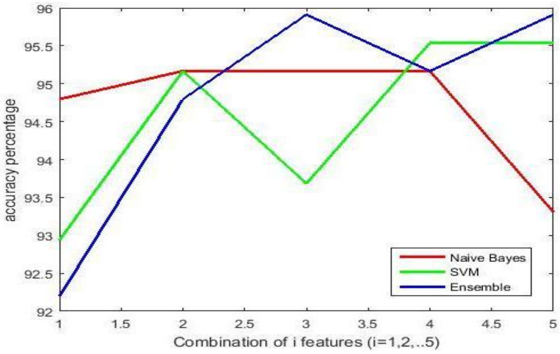

**Fig 8: Classification accuracies after mapping the dataset into a five dimensional space using Principal Component Analysis and applying the result to Naïve Bayes, Support Vector Machine and Ensemble classifiers.**

**Table 1. Classification accuracies using the result of PCA**

| Sl. No. | Name of Algorithm | Classification Accuracy (in %) |
|---|---|---|
| 1. | Naïve Bayes | 95.1673 |
| 2. | Support Vector Machine | 95.5390 |
| 3. | Ensemble | 95.9108 |

Here confusion matrix is used which is a table that is often applied to describe the performance of a "classifier" or classification model on a collection of test data for which the true values are known. The level of effectiveness of the classification model is calculated with the number of incorrect and correct classification in each possible value of the variable being classified in the confusion matrix. In this paper a testing dataset is taken where there are 203 malignant data out of which 201 have been predicted correctly and only 2 are predicted wrongly. Also, there are 66 benign data of which 5 are predicted wrongly and 61 are predicted correctly. This analysis is shown in Table 2.

**Table 2. Confusion matrix of the dataset**

| | | Predicted | |
|---|---|---|---|
| | | **Malignant** | **Benign** |
| **Actual** | **Malignant** | 201 | 2 |
| | **Benign** | 5 | 61 |

**Table 3. Execution time of the classification algorithms**

| Sl. No. | Name of Algorithm | Execution Time (in ms) |
|---|---|---|
| 1. | Naïve Bayes | 0.102023 |
| 2. | Support Vector Machine | 0.311502 |
| 3. | Ensemble | 32.404418 |

In Table 3 the execution time has been obtained for each of the three classification algorithms. The execution times for Naïve Bayes, SVM and Ensemble algorithms are 0.102123, 0.311502 and 32.404418 milliseconds respectively.

The comparative study of the classification algorithms in Table 4 reveals that Naive Bayes and SVM algorithms give almost equal accuracy when five most dominant attributes (Worst Concave Points, Worst Perimeter, Mean Concave Points, Worst Radius and Mean Perimeter) are considered and it is obtained by using Pearson Correlation Coefficient technique, without binning. But, if 19 features are considered then SVM gives better result. Introducing the concept of binning it is observed that accuracy percentage has increased and it becomes constant for three methods when 5 most dominant features are considered. Here, again if the number of dominant feature is increased to 19, SVM gives better result. So this observation shows that for both binning and without binning, SVM gives better result with respect to the other two algorithms. But here the major drawback is the consideration of a large number of dominant features. Also, the execution time has been taken for the three classification methods i.e., Naive Bayes, Ensemble and SVM. Although they are giving same accuracy percentage after binning yet the execution time for Naïve Bayes is the lowest compared to other two classifiers. So it is conclude that Naïve Bayes is the best classification technique having least time complexity and it gives better classification accuracy with only five dominant features after introduction of the binning concept.

**Table 4. Classification accuracies of different classifiers (Naïve Bayes, SVM and Ensemble) using dominant features obtained by Pearson Correlation Coefficient technique**

| Sl. No. | Name of Algorithm (Number of Features Used) | Classification Accuracy (in %) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Without Binning | | With Binning | | | |
| | | Considering Individual Feature | Considering Combination of Features | Considering Individual Feature and Taking Upper Bound of Each Bin | Considering Individual Feature and Taking Lower Bound of Each | Considering Combination of Features Stepwise and Taking Upper | Considering Combination of Features Stepwise and Taking Lower |
| 1. | Naïve Bayes (using 5 features) | 93.3086 | 95.5390 | 96.2825 | 96.2825 | 97.3978 | 97.3978 |
| 2. | Support Vector Machine (using 5 features) | 94.4238 | 95.9108 | 96.2825 | 96.2825 | 97.3978 | 97.3978 |
| 3. | Support Vector Machine (using 19 features) | 94.4238 | 98.8848 | 96.2825 | 96.2825 | 98.5130 | 98.5130 |
| 4. | Ensemble (using 5 features) | 92.1933 | 93.6803 | 96.2825 | 96.2825 | 97.3978 | 97.3978 |

## 5. CONCLUSION

Comparing to all other cancers, breast cancer is one of the major causes of death in women. So, the early detection of breast cancer is needed in reducing life losses. This early breast cancer cell detection can be predicted with the help of modern machine learning techniques. In this paper data cleaning, feature selection, feature extraction, data discretization and classification techniques have been applied for predicting breast cancer as accurately as possible. This project reveals that Naïve Bayes classifier gives the maximum accuracy of 97.3978% with only five dominant features and time complexity of this algorithm is 0.102023 millisecond which is least compared to other two classifiers.

This work can further be enhanced by modifying Support Vector Machine which gives maximum accuracy with nineteen dominant features. It is a challenging task in machine learning and data mining areas to construct a specific and computationally efficient classifiers for medical applications. With the help of machine learning methods it is really difficult to diagnose the different medical conditions of a breast cancer patient and prediction of conditions are also more critical in nature. For big datasets how these classification algorithms behave, that is one of the future scopes of this project. Moreover the identification of particular stage of breast cancer can be done in near future.

## 6. REFERENCES

[1] Shoon, Lei Win and Htike@Muhammad Yusof, Zaw Zaw and Yusof,Faridah and Ibrahim Ali, Noorbatcha (2014)“Cancer recognition from DNA microarray gene expression data using averaged one-dependence estimators.” International Journal on Cybernetics & Informatics (IJCI) , 3 (2). pp. 1-10. ISSN 2277-548X (O) 2320-8430 (P).

[2] http://www.nationalbreastcancer.org/breast-cancer-facts.

[3] www.aicr.org/assets/docs/pdf/brochures/reduce-your-risk-of-breast.pdf.

[4] https://dialnet.unirioja.es/descarga/articulo/4036558.pdf

[5] Daniele Soria, Jonathan M. Garibaldi, Elia Biganzoli, Ian O. Ellis,"A comparison of three different methods for classification of breast cancer data."Machine Learning and Applications, 2008. ICMLA'08.Seventh International Conference on. IEEE, 2008.

[6] https://en.wikipedia.org/wiki/Principal_component_analysis.

[7] https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php.

[8] https://books.google.co.in/books?isbn=0230637604.

[9] pradeeploganathan.com/support-vector-machines-svm/

[10] Jiawei Han, Jian Pei, Micheline Kamber ”Data Mining Concepts and Techniques”, Third Edition, Elsevier Inc, 2012, ISBN:978-0-12-381479-1.

[11] Sau Loong Ang, Hong Choon Ong and Heng ChinLow, "Classification Using the General BayesianNetwork." Pertanika Journal of Science & Technology24.1 (2016).

[12] K. Sivakami, "Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model." International Journal of Scientific Engineering and Applied Science (IJSEAS) -Volume-1, Issue-5, August 2015.

[13] Shweta Kharya, Shika Agrawal and Sunita Soni, "Naïve Bayes Classifiers: Probabilistic Detection Model for Breast Cancer."International Journal of Computer Applications 92.10 (2014).

[14] G. Ravi Kumar, Dr G. A. Ramachandra and K. Nagamani. "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques." International Journal of Innovations in Engineering and Technology (IJIET) 2.4 (2013): 139.

[15] C. D. Katsis, I. Gkogkou, C.A. Papadopulos, Y.Goletsis, P. V. Boufounou, G. Stylios "Using artificial immune recognition systems in order to detect early breast cancer." International Journal of Intelligent Systems and Applications 5.2 (2013): 34.

[16] Gouda I. Salama, M. B. Abdelhalim and Magdy Abd-elghany Zeid, "Breast cancer diagnosis on three different datasets using multi-classifiers." Breast Cancer (WDBC) 32.569 (2012): 2.

[17] Kim W, Kim KS, Lee JE, Noh DY, Kim SW, Jung YS, Park MY, Park RW, "Development of novel breast cancer recurence prediction model using support vector machine." Journal of breast cancer 15.2 (2012): 230-238.

[18] Mehmet Fatih Akay,“Support vector machines combined with feature selection for breast cancer diagnosis.” Expert Systems with Applications 36 (2009) 3240–3247.

[19] Diana Dumitru,"Prediction of recurrent events in breast cancer using the Naive Bayesian classification."Annals of the University of Craiova-Mathematics and Computer Science Series 36.2 (2009): 92-96.

[20] https://en.wikipedia.org/wiki/C4.5_algorithm.

[21] https://archive.ics.uci.edu/ml/machine-learning databases /breast cancer-wisconsin/wdbc.data.