Hybrid Approach to predict Breast Cancer using Machine Learning Techniques

Megha Rathi
Department of computer science & Engineering
Jaypee Institute of Information Technology
Noida, India
megha.rathi@jiit.ac.in

Vikas Pareek
Department of computer Science & Information Technology
Banasthali University
Banasthali (Rajasthan), India
er_pareekvikas@yahoo.co.in

Abstract— The research is about the prediction of breast cancer using machine learning techniques. Prediction of cancer type is one of the crucial aspect of statistical analysis. In the paper we presented a hybrid approach for the development of tool which can predict breast cancer with the help of Machine Learning techniques. The main reason behind developing this tool is that the number of bioinformatics tool for prediction of target class is very scarce and rare. In this paper we analyze the performance of machine learning algorithm for predicting the target class based on some attributes. We have implemented MRMR feature selection algorithm along with four different classifiers to find out the best classifier for the breast cancer domain. Classifiers used in this study are SVM(Support Vector Machine), FT(Function Tree), End Meta, and Naïve Bayes for comparison on different parameters that are accuracy, root mean square error, mean absolute error, Kappa statistics, Sensitivity and specificity. The tool also helps in determining the chances of getting cancer in the future also. We experimented with different datasets with same approach and found that our research produces better results for every statistical measure. Also we presented the comparative study which shows that our approach is better than other existing approaches. Data sets are taken from UCI ML Repository [1] and tested on both multi class and binary data sets namely WDBC, WBC and Breast tissue dataset. This tool could serve as a boon because it could help oncologists to determine the type of breast cancer also predicting the chance of getting breast cancer within no time.

Keywords- Breast Cancer, Machine Learning, Prediction, Data Pre-processing, Feature Selection, Hybrid Approach.

I. INTRODUCTION

Breast Cancer is one of the most life threatening disease of the world. During the past few decades, breast cancer has emerged to be one of the top cancer killers amongst women worldwide [2]. In the year 2009 it has been investigated that, many new cases of breast cancer were detected and thousands of lives were claimed to be taken by breast cancer worldwide [2] .According to [3] in the year 2010 one and a half million new cases was diagnosed. Despite the high incidence rates of occurrence of breast cancer in women, survival rate of breast cancer patient is still 5 years after the patient is diagnosed with breast cancer and this is because of treatment and diagnosis [3]. Thus this shows that early prediction of Breast Cancer disease and its treatment at an early stage can reduce the mortality rate globally. Thus it is very essential to diagnose this life threatening disease at an early stage. One of the major clinical problems of this disease is the prediction of type of breast cancer. Thus the primary time goes in knowing about the type of breast cancer during which many of the states of breast cancer gets passed. Thus this tool could helps to predict type of breast cancer of patients, hence saving valuable time which could help in controlling mortality rate to a large extent. Prediction of type of breast Cancer more accurately and precisely would help oncologists in diagnosis and in the treatment of breast cancer. Controlling the mortality rate is the main reason or motivation behind the development of this tool which could predict the cancer type also predict the risk of having cancer in future. This would further help in reducing the colossal cost of treatment for the patients and would thus help oncologists to make more accurate decisions in diagnosis and treatment of the patient's disease. In the research [4] it has been found that predictive models are very helpful when experienced oncologists are not available. Data mining techniques are applied for the creation these predictive models which is very supportive tool for physicians in decision making. In this study we have proposed a Bioinformatics tool which could help in prediction of the type of breast cancer of a patient by using different machine learning algorithms that are SVM, End Meta, FT, and Naïve Bayes. Also we compared these algorithms and came to a conclusion about the best classifier on the parameters of accuracy, precision and

Kappa statistics, Sensitivity, and specificity. For this purpose we experiment our Machine Learning algorithm on Breast Cancer Dataset taken from UCI ML Repository. Inputted data has some attributes which are helpful in determining the Cancer type as Benign or Malignant. The entire framework is developed in Java Netbeans interface.As the amount of data inreases rapidly so there is need of some techniques to anaylze the data and produce important results which is helpful in medical diagnosis. Data mining is very helful in analyzing, estimating medical data also proves to be crucial in decision making in medical diagnosis.

In this study an approach for breast cancer prediction and prediction for chanes of having breast cancer in future is proposed.We have implemented a hybrid approach which makes use of MRMR feature selection algorithm with four different classifiers to check which one is producing better result for the breast cancer data set. This study also presents the performance of all the implemented four classifiers with MRMR feature selection algorithm. Total three data sets namely Wisconsin Breast cancer data set, Wisconsin Diagnostic Breast Cancer, and Breast Tissue data set are processed with our tool after algorithm assessment SVM shows the best relevant output for breast cancer prediction. The main difficulties of this work are the limited amount of Breast Cancer Data and missing data of some attributes.

The rest of the paper is organized as follows. Section 2 presents the Related Work on breast cancer classification using various approaches. In section 3 we elaborate the methodologies used in the paper which in turn further contains the description all data sets with the entire algorithmic concept used in the study. Section 4 presents the proposed framework. Section 5 shows all experimental study and results. Finally Section 7 presents the conclusion of the paper.

II. RELATED WORK

In the paper [5] a tool is developed which assists doctors in decision making for suggesting treatment methods for the patients suffering from breast cancer. In this study data mining techniques are applied for the treatment of breast cancer. Tool helps for the determination of treatment methods for the post surgical operation for breast cancer patient. The data set is obtained from Ankara Oncology Hospital consists of 462 records and Classification algorithms are applied one by one on this data set to find out the best classifier. Results obtained in this study are that; for hormonotherapy output IBI with accuracy 94.6237%, for tamoxifen output Multilayer Perceptron with accuracy 92%, for chemotherapy output and accuracy is 97.77% , and for radiotherapy output Multilayer Perceptron with accuracy 95.23% are defined as best algorithm. In the work [6] author presented a hybrid breast cancer detection system using neural network and feature selection based on Sequential backward and Sequential forward selection. In this study feature selection techniques SFS and SBS is developed using tenfold cross validation are combined with Principal Component Analysis in order to obtain two new hybrid feature selection techniques as SFSP and SBSP. According to the results found in this study it has been found that SBSP method for the selection of feature is better than SFSP feature selection method. Feature space is reduced from 9 to 4 after using SFSP method .Neural Network is used as a classifier in this study and classification accuracy achieved is 98.57%(Neural Network + SBSP) and 97.57%(Neural Network+ SFSP).In the study [7] an algorithm called Class AMP helps in the prediction of propensity of protein sequence for identifying the antibacterial, antifungal, or antiviral activity. The algorithm is developed using Random Forest and Support Vector Machine which is well known data mining techniques. In paper [7] three different approaches were used for the purpose of AMP's classification. First Approach helps in the classification of AMP which is active against one or more classes, while second and third approach aims to find AMP which is active against a particular class of microbes. The positive data set for all the Methods are same, However negative data set are different. The negative dataset in Methods I is a sequence of nonantimicrobial. Method II negative data set consists of sequences which were active against the other two classes. In method III the multiclass classification algorithm was employed. In this study no separate negative training data set was used for all the three methods support vector machine and random forest were used for the prediction purpose. In this study it is also found that model built using Method II performed better than model built using Method I and Method III and hence been deployed in Class AMP. Data Mining techniques using CART method are applied in the study [8] for the development of Remote Health Monitoring system of heart failure which upgrade the efficiency of detecting the severe heart problem of the patient. In this study developed application detects the severe heart problem and classify into severe and mild heart failure. In this study author implemented Classification and regression tree in a telemedicine platform to detect heart failure and its severity. Results obtained in this study in terms of accuracy and precision are 96.39% and 100% for the detection of heart failure and 79.31% and 82.35% for classifying the heart failure into severe and mild heart failure. According to the paper [9] performance of neural network classifier is improved using floating centroid method and particle swarm optimization with inertia weight as optimizer. In conventional neural network classifier position of centroid and classes are set manually also the count of centroid remains constant with respect to the number of classes. Results obtained in this study are also very promising and achieve accuracy equivalent to 96.47% which is higher than other conventional neural network classifiers. In the paper [10] Francesco Folino and Clara Pizzuti presented an approach for the prediction of disease that combines various data mining techniques like clustering, Markov models and association analysis. A cluster of medical records is

generated and then Markov model for each cluster is created for the detection of disease. The proposed model is known as CORE^{+} that combines clustering, association analysis with Markov Model. . The developed model uses the past patient medical history for generating models able to determine the risk of individuals to develop future disease. Patient medical records are clustered and markov model is generated for each cluster for the prediction of disease a patient could likely to be affected. If probability of generated markov models is not high it starts sequential analysis for the selected items by considering high confidence rules produced by recurring disease pattern. The dataset used consists of 1105 patient records with 330 distinct diseases. According to the result of this study prediction accuracy improved with the combination of all models (clustering, association analysis, and Markov Model). In the study [11] improved classifier Least Square Support Vector Machine is developed for the detection of breast cancer and achieves accuracy 98.53% using ten cross fold validation. Xu et al. [12] proposed linear orthogonal transform algorithm for the diagnosis of breast cancer and accuracy achieved in this study is 98.53%. In the study [13] Yeh proposed a hybrid approach for breast cancer pattern mining and achieves accuracy of 98.71%. Author combined two techniques discrete particle swarm optimization and statistical method for diagnosis. In the study [13] author presented an approach for the classification and recognition of breast cancer using least square support vector machine algorithm and accuracy achieve through this method is 98.53%. Yeh WC, Chang WW, and Chung YY in the paper [14] developed a rule based classifier using simplified swarm optimization. In this study data set used is the thyroid gland dataset obtained from the UCI ML repository. Accuracy of classifier improved after adding close interval encoding to present the rule structure, and orthogonal array test to prune rules. In 2011 Xu Y, Qi Z, and Wang J presented a study [15] in which author applied machine learning techniques for breast cancer diagnosis. In this study author propose a technique known as kernel orthogonal transform for breast cancer diagnosis. Results obtained from this study showed that accuracy of classifier is improved and classifier classifies with more accuracy than the entire previously used machine learning techniques.

III. METHODOLOGY

The proposed framework consists of four main modules namely: Data Collection, Data Pre-processing, Feature Selection, and Classification. The details of all of them are discussed below.

A. Data Collection

First step is the collection of data. We experimented with three different data seta collected from UCI ML Repository. First data set is Wisconsin Diagnostic Breast Cancer dataset, Second is Wisconsin Breast Cancer Data set and last one is Breast tissue dataset. The details of all data sets are provided below.

- *Wisconsin Diagnostic Breast Cancer Dataset:* Data Set is taken from the UCI ML Repository [1]. We have experimented with three different data sets. For the paper first data set used is Wisconsin Diagnostic Breast Cancer data set which was obtained from University of Wisconsin Hospitals and is used by many researchers who is doing research on breast cancer. The WDBC data set contains a fine needle of data mass where features are extracted from digitized image. All the features represent the characteristics of cell nuclei present in the image. Data Set is linearly separable using all 30 input features. The dataset contains total 569 instances with 32 attributes (ID, Diagnosis, 30 real valued features). Table I presents the dataset attribute information. Table I presented the details of attribute information where first attribute is the unique id per patient and second represent the class label either malignant or benign. The attribute range in between 3-32 is ten real valued features for each cell nucleus. Radius is the mean of distances from centre to point on the perimeter. Texture is the standard deviation of gray scale values. Smoothness is the local variation in radius length. Compactness is computed as: (perimeter power 2/area-1.0). Concavity is the severity of concave portions of the contour and Fractal dimension is the (coastline approximation)-1. 30 features are computed using the mean, standard error and worst of these features. For instance field 3 is mean radius, field 13 is Radius SE, and field 23 is worst radius.

TABLE I WISCONSIN DIAGNOSTIC BREAST CANCER DATASET ATTRIBUTES

1). ID_number
2).Class_Label(M=malignanat ,B=benign)
3-32). Ten features are computed for each nucleus
a) radius
b) texture
c) perimeter
d) area
e) smoothness
f) compactness
g) concavity

h) concave points
i) symmetry
j) fractal dimension

- *Wisconsin Breast Cancer Dataset:* This dataset also taken from UCI ML Repository and has total 699 instances. Wisconsin Breast Cancer data set is collected from Wisconsin Hospitals, Madison. Samples in the database collected periodically by Dr. Wolberg. The data set sample consists of visually accessed nuclear features of FNA (fine needle aspirates) taken from the breasts. Each sample assigned a 9 dimensional vector (attributes 3 to 9 below). Each attribute value lies in the range 1-10 where 1 denotes the normal state and 10 represents the most abnormal state. Inputted data has some attributes which are helpful in determining the Cancer type as Benign or Malignant. Wisconsin Breast Cancer data sets is used by many researchers and prove to be very useful in visualizing many results for cancer diagnosis .Attributes are depicted in the Table II.

TABLE II WISCONSIN BREAST CANCER DATASET

Clump_Thickness	[1,10] integer
Cell_Size_Uniformity	[1,10] integer
Cell_Shape_Uniformity	[1,10] integer
Marginal_Adhesion	[1,10] integer
Single_Epi_Cell_size	[1,10] integer
Bare_Nuclei	[1,10] integer
Bland_Chromatin	[1,10] integer
Normal_Nucleoli	[1,10] integer
Mitoses	[1,10] integer
Class	{benign, malignant}

- *Breast Tissue Dataset:* This dataset contains impedance measurements of freshly excised breast tissue which were made at the following frequencies: 15.625, 31.25, 62.5, 125, 250, 500, 1000 KHz. These measurements lies in the real and imaginary plane constitute the impedance spectrum from where the breast tissues are calculated. The dataset can be used for predicting the classification of either the original 6 classes or of 4 classes by merging together the fibro-adenoma, mastopathy and glandular classes whose discrimination is not important. This dataset is about the electrical impedance measurements of freshly excised tissue samples from the breast. The data set consists of total 106 instances with 10 attributes (9 features+1 class attribute).Table III present the attribute description of breast tissue data set and Table IV present the six classes of breast tissue dataset.

TABLE III BREAST TISSUE DATASET

Attribute	Description
IO	Impedivity at zero frequency
PA500	Phase angle at 500 KHZ
HFS	High frequency slope of phase angle
DA	Impedance distance between spectral ends
AREA	Area under Spectrum
A/DA	Area normalized by DA
MAX IP	Maximum of the spectrum
DR	Distance between IO and real part of the maximum frequency point
P	Length of spectral curve

TABLE IV CLASS DESCRIPTION OF BREAST TISSUE DATASET

Class Labels
Car Carcinoma
Fad Fibro-adenoma
Mas Mastopathy
Gla Glandular
Con Connective
Adi Adipose

B. Data Preprocessing

Data Pre-processing is the second step in which we process the data set so that high quality data is available which is error free. Data should be non ambiguous, correct, and complete because classification accuracy depends on the quality of data. Data Pre-processing is applied to remove inconsistencies from the data set, also to fill missing values. Data set obtained from different sources contains redundant and irrelevant data. In order to remove such kind of inconsistencies from the data set we apply data cleaning techniques. Data cleaning techniques helps in dealing with anomalies of existing data. Data cleaning mainly deals with error checking, error detection, and Data Validation. In the study we analyze all three datasets and it has been found that first data set i.e. Wisconsin Diagnostic Breast Cancer dataset contains no error, second data set i.e. Breast Cancer data set contains some missing data and some irrelevant data so we clean the data by filling missing values and removing the irrelevant values by relevant one. Missing values are filled by using the attribute mean for all samples belonging to the same class. Last data set i.e. Breast Tissue data set is also error free so no data cleaning techniques applied on the data set.

C. Feature Extraction

In third module we select the relevant features out of the given data set so that dimensionality of data set is reduced. Feature Selection is the process of selecting the subset of features or attributes that is inputted to the system. Accuracy of classifier depends upon the features of data set which contribute to the prediction of breast cancer. In this study features are selected using MRMR (Maximum Relevance and Minimum Redundancy) [16] algorithm. Minimum redundancy is a feature selection algorithm which is used for the identification of some very important characteristics of phenotype and genes and reduces their relevancy [16]. This algorithms works by selecting those features which were mutually far away and having "high" correlation to the classification variable

D. Classification Techniques for Cancer Prediction

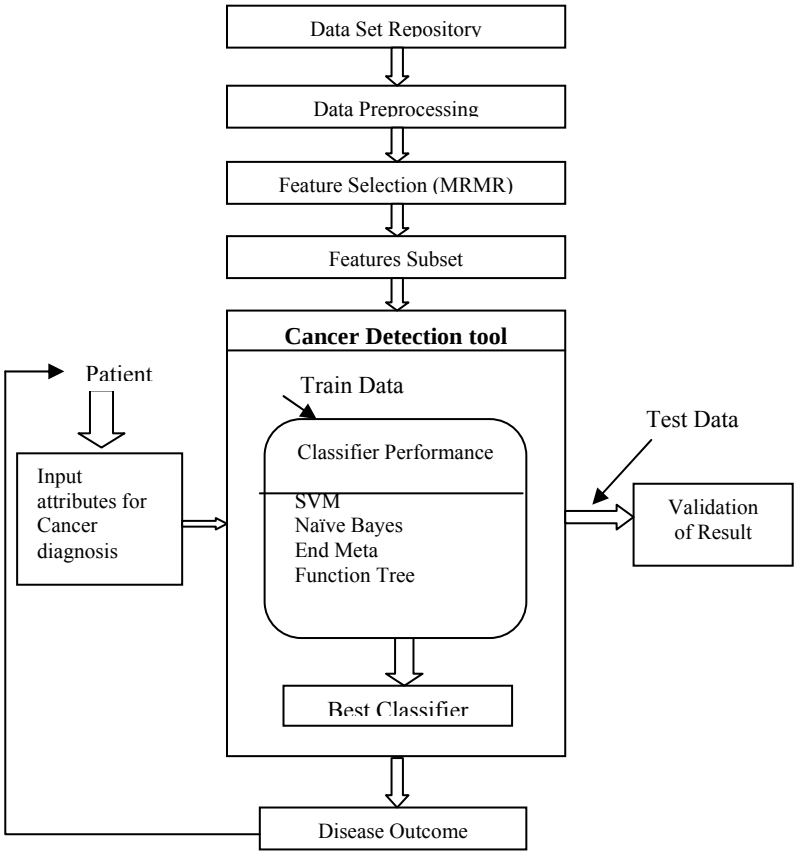
In last module we apply classification algorithm one by one on the reduced data set to find out the best classifier. We have implemented four algorithms End Meta, Naïve Bayes, SVM (Support Vector Machine), and FT (Function Tree) to find out which one is giving better result for breast cancer detection. After extracting the relevant features from the data set we apply classifier to check the performance of classifier. Also classifiers are used to predict the chance of having cancer in future also. A Support Vector Machine (SVM) performs classification by constructing an *N*-dimensional hyper plane that optimally separates the data into two categories [17]. The Naïve Bayes Technique follows the Bayesian approach which is very simple and used for fast classification. This technique considers mutually independent features and is used in many concerned areas to achieve significant results in machine learning [18]. End Meta works upon search algorithm and evaluate next to the base classifier. It helps in reflecting the transparency in feature selection and classifier receives only the reduced features [19]. Function Tree presents the dependencies between main functions of the system. The entire problem is split into two or more subset problems which enhance tree visualization [20]. A Function Tree contains nodes and nodes represent which function calls another function.

IV. PROPOSED FRAMEWORK

In this study we presented a framework for the prediction of breast cancer at initial stage. In this study we implemented a tool known as “Breast Cancer Prediction Tool” with the help of Java Net Beans Interface which detect disease at initial stage and treatment start at initial stage of disease which decrease mortality rate due to breast cancer. In this framework we collect the data set and apply data cleaning techniques in order to improve classifier performance. Then features are selected using feature selection algorithm. Then we split the data in two parts one is test data and other is train data. Classifiers are trained using train data then classification

algorithms are applied one by one to find out which one is producing better result in terms of accuracy for the given data set. The proposed framework prove to be very useful in healthcare domain as it detects disease at initial stage and prediction accuracy also enhances because of the use of features selection algorithm MRMR. If a tool is available which detect disease just after inputting some values one can easily predict the outcome of disease without doctor intervention. Working of the tool depicted below:

- Features are selected using MRMR algorithm.
- Classifiers are trained using training data set.
- Classifiers are also trained to predict the risk of having cancer in future also.
- Input the data in the tool for prediction of breast cancer.
- Choose any algorithm out of the given four algorithms: SVM, End Meta, Naïve Bayes, and Function Tree.
- Predict Class attribute.
- Predict accuracy and performance of all prediction algorithms on the inputted test data.



]

Fig.1 Overall architecture of Proposed system

Fig.1 presents the architecture of the proposed system. In this study we have implemented four different classifiers using Java NetBeans interface for the prediction of breast cancer. Accuracy of the classifiers depends on the quality of data inputted to the system so it is essential to select relevant and important features to enhance classifiers performance in terms of accuracy. We apply MRMR algorithm for the selection of subset of features and it is found that accuracy of classifiers improved after selecting the subset of features. In this study we also find out the best classifier for the given breast cancer domain out of the implemented four classifiers. This study shows that machine learning is very useful in healthcare domain especially for the prediction of disease classifiers with MRMR feature selection algorithm. We have implemented four machine learning algorithms SVM, End Meta, Function Tree, and Naïve Bayes for comparison on different parameters like accuracy, root mean squared error, mean absolute error, and Kappa Statistics and also to check performance of all these classifiers for the prediction of breast cancer of a patient on inputting the attribute values to the system.

V. EXPERIMENTAL STUDY AND RESULTS

A. Snapshot of Selected Features

We experimented with three different sets Figure 2, 3 and 4 present the attribute selected for the three data sets.

Feature Selection Method: MRMR
Total number of instances:569 with 32 attribute(one class attribute: Benign or Malignant)
Selected Attribute: Attribute Id: 3,4,6,10,11,13,18,19,21,23,26,27,30
Name of the features selected: radius se, radius worst, texture_se, perimeter_worst, area_mean, area_worst, compactness se, compactness worst, concavity se,

Fig. 2 Breast Cancer diagnostic dataset after Feature Extraction

Feature Selection Method: MRMR
Total Number of instances: 699 with 10 attribute(one class attribute)
Selected Attribute: Attribute Id:1,2,3,4,5,6,7,8,9
Attribute Name: Clump Thickness, Cell Size_Uniformity, Cell_shape_Uniformity, Marginal Adhesion, Single_Epi_Cell_size, Bare_Nuclei, Normal_Nuclei, Mitoses.
, concave points mean, symmetry mean, symmetry se,

Fig. 3 Breast Cancer Dataset after feature extraction

Feature Selection Method: MRMR
Total Number of instances: 699 with 10 attribute(one class attribute)
Selected Attribute: Attribute Id:1,2,3,4,5,6,7,8,9
Attribute Name: Clump_Thickness, Cell_Size_Uniformity, Cell_shape_Uniformity, Marginal Adhesion, Single_Epi_Cell_size, Bare_Nuclei, Normal_Nuclei, Mitoses.
, concave points mean, symmetry mean, symmetry_se, fractal_dimensio_se, fractal_dimension_worst,

Fig. 4 Breast Tissue dataset after Feature Selection

B. Estimation and Validation

In this study we presented an approach for the prediction of breast cancer using machine learning techniques. For the same we developed a software tool using NetBeans IDE which assists doctors for breast cancer detection. Developed software tool uses four machine learning algorithm namely SVM, Naïve Bayes, End Meta and Function Tree for comparison on various statistical factors is implemented in Java NetBeans interface. We will use 10 fold cross validation training data to calculate the performance of machine learning algorithm. Results of features extracted are described in table V,VI, and VII.

Listed are some different statistical parameters which we consider for statistical analysis and also to compare the performance of all classifiers on three different data sets [21].

- Accuracy
- Kappa Statistics (KS)
- Mean Absolute Error (MAE)
- Root Mean Square error (RMSE)
- Sensitivity
- Specificity

Table V represent the matrix for performance evaluation:

TABLE V PERFORMANCE EVALUATION METRICS

	Predicted Class		
Actual Class		Class=Yes	Class=No
	Class=Yes	a: TP	b: FN
	Class=No	c: FP	d: TN

In the above table where TP denotes the number of true positive, TN denotes the number of true negative and FP denotes the number of False Positive, and FN represents the number of False Negative.

Accuracy is the correct prediction percentage and is computed using the formula [22]:

Accuracy=TP+TN/TP+TN+FP+FN(1)

To Compute Kappa Statistics first we need to calculate the observed level of agreement [23]

Pr(α)= (a+d)/(a+b+c+d) (2)

This value needs to be compared to the value that you would expect if the two raters were totally independent

Pr (e)=(a+c)/(a+b+c+d)*(a+b)/(a+b+c+d)+(b+d)(a+b+c+d) * (c+d)/(a+b+c+d) (3)

The value of Kappa is defined as

K = Pr(α)-Pr(e)/1-Pr(e) (4)

Mean absolute error (MAE) is a quantity used to measure how close is predictions are to the eventual outcomes [24]. The mean absolute error is given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n} \quad (5)$$

As the name suggests, the mean absolute error is an average of the absolute errors $e_i = f_i - y_i$, where f_i is the pred iction and y_i the true value.

The RMSE of an estimator with respect to the estimated parameter θ' is defined as the square root of the mean square error [24]:

$$RMSE = \sqrt{MSE(\theta')} = \sqrt{(E(\theta' - \theta))^2} \quad (6)$$

For an unbiased estimator, the RMSE is the square root of the variance, known as the standard error.

Sensitivity and specificity is computed as [25]:

Sensitivity= TP/TP+FN (7)

Specificity= TN/TN+FP (8)

C. Classification and other statistical Results

About Dataset: We experimented with three different breast cancer data sets. First data set has 569 total instances, second has total 699 instances and third has total 106 instances. All the experiments are implemented in Java NetBeans interface .We will use 10 cross fold validation training data to calculate the performance of all classifiers. Table VI present the result of our tool with first data set i.e. Wisconsin Diagnostic Breast Cancer Data Set.

TABLE VI RESULTS WITH FIRST DATASET

Algorithm	**Accuracy (%)**	**Kappa Statistics**	**RMSE**	**Mean absolute error**	**Sensitivity (%)**	**Specificity (%)**
SVM	99	0.9495	0.1500	0.035	96.21	98.93
Naïve Bayes	98.17	0.9441	0.1510	0.063	96.10	98.34
FT	97	0.9321	0.1953	0.030	95.65	97.7
End Meta	96.5	0.9111	0.1900	0.0465	94.38	96.6

Table VII present the result of our tool with second data set i.e. Wisconsin Breast Cancer Data Set.

TABLE VII RESULTS WITH SECOND DATASET

Algorithm	Accuracy (%)	Kappa Statistics	RMSE	Mean absolute error	Sensitivity (%)	Specificity (%)
SVM	98.97	0.9591	0.1499	0.0131	97.12	99
Naïve Bayes	98.2	0.9571	0.1510	0.0146	96.97	98.78
FT	97.8	0.9338	0.1733	0.03	95.87	98.10
End Meta	96.0	0.9095	0.1985	0.0405	94.87	98.81

Table VIII present the result of our tool with third data set i.e. Breast Tissue Data Set.

TABLE VIII RESULTS WITH THIRD DATASET

Algorithm	Accuracy	Kappa Statistics	RMSE	Mean absolute error
SVM	99	0.9611	0.1450	0.0111
Naïve Bayes	98.8	0.9572	0.1421	0.0124
FT	97	0.9330	0.1953	0.031
End Meta	96.65	0.9330	0.1954	0.030

Table IX present the sensitivity and specificity results for breast tissue data set for all the six classes. In the table shown below we presented the sensitivity and specificity for six classes namely: Car, Fad, Mas, Gla, Con, and Adi. I represent the sensitivity value while II represent the specificity value for the breast tissue data set.

TABLE IX SENSITIVITY AND SPECIFICITY OF BREAST TISSUE DATASET

Algorithm	Classes											
	Car		Fad		Mas		Gla		Con		Adi	
	I	II	I	II	I	II	I	II	I	II	I	II
SVM	100	100	99	100	100	100	99	100	99	100	100	100
Naïve Bayes	99	100	100	100	93	95.7	96	98.1	95	97.23	98	99
FT	96.7	97.1	94.3	96.5	94.4	96.9	93.9	95.9	93.9	93.0	99	100
End Meta	99	100	93.3	95.67	93.3	95.6	93.7	95.13	93.9	94.67	94	97.21

From the results shown above it has been found that our hybrid approach increases the efficiency of all classifiers in terms of all statistical parameters. For the selected input features the results are very promising as SVM and naïve Bayes achieve accuracy of 99% which is quiet high. We also compare our results with other existing hybrid approaches and with WEKA also and it has been found that our results are much better than other. Table X present the result on WEKA on 10 cross fold validation on three different data sets

TABLE X WEKA RESULTS ON BREAST CANCER DATASET

Algorithm	Accuracy (Data Set I)	Accuracy (Data set II)	Accuracy (Data Set III)
SVM	97.5	96.9957	96
Naïve Bayes	97.23	95.9943	97.13
FT	96	96.995	94.34
End Meta	94	94.56	92.34

Table XI present the comparison of our hybrid classifier with other hybrid approaches on the domain of breast cancer and it is seen that our approach works well for the diagnosis of breast cancer when compared with other existing approaches. Our approach achieves almost 99% accuracy for the classification of breast cancer. From the results shown above it has been found that SVM produces better output in terms of correctly identified instances when combined with MRMR algorithm on entire three data sets. SVM classifier achieves accuracy of 99% on an average when combined with MRMR algorithm while classifier performance without using feature selection algorithm is 97.5% only. After SVM then Naïve Bayes achieves second highest accuracy when merged with the given feature selection algorithm. Accuracy achieved by Naïve Bayes classifier is 98.5% on an average. FT achieves accuracy 97.5% on an average. Amongst all implemented algorithm End Meta shows worst accuracy 96% when combined with MRMR feature selection algorithm. The performance of our hybrid approach is

shown in graphical format in Fig.5 while Fig.6 presented the performance of WEKA classifiers on the three data sets.

TABLE XI COMPARATIVE ANALYSIS WITH OTHER EXISTING HYBRID APPROACH

Comparison of our and other approaches	**Algorithm I**	**Algorithm II**	**Accuracy**
Our Approach for Breast Cancer Diagnosis	SVM	MRMR	99%
Hybrid Breast Cancer Detection system [8]	Neural Network	Sequential Forward Selection and Sequential Backward Selection	SFSP+NN (97.5%) and SBSP+ NN (98.5%)
Breast Cancer Classification [9]	Neural Network	Particle Swarm Optimization	96.47%
Breast Cancer Diagnosis [13]	Least Square SVM	SVM	98.5%
Breast Cancer Diagnosis [14]	Particle Swarm Optimization	Statistical Method	98.7 %
Breast Cancer Diagnosis [26]	Feature Selection Artificial Immune System	C4.5 Decision Tree	98.5%
Intelligent Hybrid Method for Breast Cancer diagnosis [27]	Fuzzy Clustering	SVM	97.34%
Novel Algorithm for Breast Cancer Detection [28]	Constrained search sequential floating forward search(CSSFFS)	SVM	98%
Breast Cancer Detection in peripheral Blood [29]	Recursive Feature elimination and cross validation	SVM	98.4%

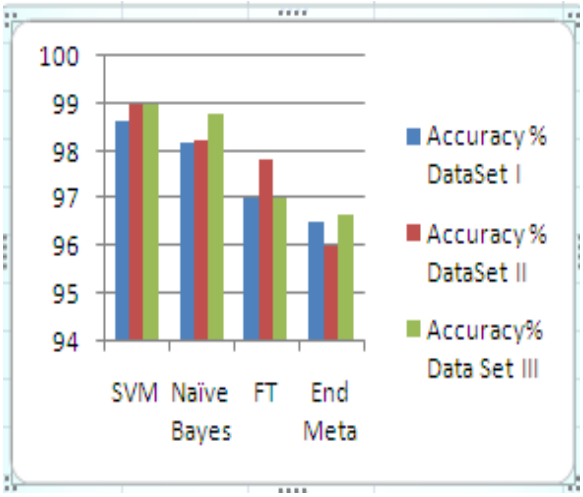


Fig. 5 Performance Analysis of our Approach

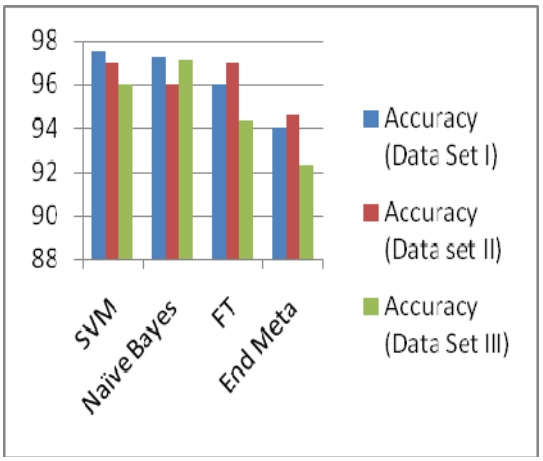


Fig. 6 Performance of WEKA Classifiers

VI. CONCLUSION

The tool is developed specially for oncologist for predicting cancer type with in no time and thus helps in decision making for the treatment method.This paper implemented using machine learning techniques can be helpful in diagnosing the cancer type and to assist oncologist for decision support. For this purpose we propose a hybrid approach to help oncologist in diagnosing the breast cancer and risk of having breast cancer using machine learning techniques. In our framework we embed MRMR feature selection with four classifiers namely SVM, Naïve Byes, Function Tree, and End Meta for comparison on different parameters like accuracy , root mean squared error, mean absolute error, and Kappa Statistics, sensitivity and specificity to find out the performance of all classifier for the prediction of breast cancer with the MRMR algorithm and it has been found that SVM performance is better than other classifier when combined with feature selection algorithm for the breast cancer domain. Also the results are compared with WEKA [30] and it is found that our approach achieve higher accuracy for all the four implemented classifier. Also we compare the results with other existing hybrid techniques for breast cancer detection and our results shows that we achieve higher rate of accuracy than the other existing techniques. In conclusion this study shows that machine learning techniques can be a useful tool for medical diagnosis and applications particularly at treatment decision statement. This tool helps oncologist or patient to decide in a short time whether the person is suffering from cancer or not .Just input some attributes in the tool and result is there to detect breast cancer.

REFERENCES

[1] “Data set Repository”(http://archive.ics.uci.edu /ml)
[2] “Breast cancer awareness” (http:// www. notouchbreastscan.com /awareness globaldisease.html)
[3] "World Cancer Report". International Agency for Research on Cancer. 2008. Retrieved 2011-02-26.
[4] “Breast cancer statistics” (http:// www. world widebreastcancer.com / learn/ breast-cancer-statistics-worldwide).
[5] American cancer society. Breast cancer facts and figures 2005-2006.
[6] Abdulkadir Cakir, Burcin Demirel, "A Software Tool for determination of Breast Cancer Treatment methods using Data Mining Approach".Springer,2010.
[7] Shaini Joseph, Shreyas Karnik, Pravin Nilwae, V.K. Jayaram, and susan Idicula-Thomas, "ClassAMP: A Prediction Tool for classification of Antimicrobial Peptides". IEEE/ACM.
[8] Mustafa serter Uzer, Onur Inan, Nihat Yilmaz," A hybrid breast cancer detection system via neural network and feature selection based on SBS, SFS, and PCA. Springer Journal of Neural Computing and application, 2012.
[9] Lei Zhang, Lin wang, Xujiewen Wang, Keke Liu, and Ajith Abraham, "Research of Neural Network Classifier Based on FCM and PSO for Breast Cancer Classification”. Springer, 2012.
[10] Leandro Pecchia, Paolo Meilillo, and Marcello Bracale," Remote Health Monitoring of Heart Failure with Data Mining via CART Method on HRV Feature". IEEE Transaction on Biomedical Engineering, Vol.58, 2011.
[11] Francesco Folino and Clara Pizzuti,"Combining Markov Models and Association Analysis for Disease Prediction". pp. 39-52, ITBAM-2011.
[12] “Emerging Technologies for Patient Specific Healthcare”, IEEE Transactions on Information Technology in Biomedicine, Vol.16, No.2, 2012.
[13] Polat K, Gunes S ,"Breast Cancer Diagnosis using least square support vector machine". Digit signal Process 17(4):694-701, 2007.
[14] Yeh WC, Chang WW, Chung YY," A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method". Expert System application, 2009.
[15] Xu Y, Qi Z, Wang J," Breast Cancer diagnosis based on kernel orthogonal transforms". Neural computing application,2011.
[16] Hanchuan Peng, Fuhui Long, and Chris Ding, “"Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy ,"IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp.1226-1238, 2005.
[17] V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995
[18] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques,2nd Edition. San Fransisco:Morgan Kaufmann;2005.
[19] Ricardo Vilalta, Christophe Giraud-Carrier, Pavel Brazdil, Carlos Soares,”Using Meta Learning to support Data Mining”, International Journal of Computer Science & Applications, Vol. I, No. 1, pp. 31 – 45,2004.

[20] Rokach, Lior. "Data Mining with Decision Trees: Theory and Applications." 69 (2008): Web. 3 Feb. 2013.[14].
[21] J.Han and M.Kambes, “Data Mining Concepts & Techniques”. ,CA: Elsev Elsevier: Morgan Kaufmann Publisher,2006.
[22] Vladimir N. Vapnik. Statistical Learning theory. New York: Wiley, 1998.
[23] L Gaelle, N Nicolas,” Data Preprocessing and Kappa Coefficient”, LNAI 3641, pp.176-184,2005.
[24] Kumar, Yugal; Sahoo, G.,”Study of Parametric Performance Evaluation of Machine Learning and Statistical Classifiers” International Journal of Information Technology & Computer Science , Vol. 5 Issue 6, p57-64. 8p, May 2013.
[25] D.G. Altman, J.M. Bland,” Diagnostic tests Sensitivity and Specificity”, BMJ 308(6943):1552, 1994.
[26] Polat K, Sahan S, Halife, and Gunes S,” A New Classification Method for Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System (FS-AIRS)”, Lecture Notes in Computer Science Volume 3611, Springer 2005.
[27] Jalil Addeh and Ata Ebrahimzadeh, “Breast Cancer Recognition using a Novel Intelligent Hybrid Method”, J Med Signals Sens. 2012 Apr-Jun; 2(2): 95–102.
[28] S. Aruna, S.P. Rajagopalan,” A Novel SVM Based CSSFFS Feature Selection Algorithm for Detecting Breast Cancer”, International Journal of computer Applications, 2011.
[29] Zhang F, Kaufman H L, Deng Y, and Drabier R, “Recursive SVM Biomarker selection for early detection of breast cancer in peripheral blood”, International Conference on Bioinformatics and Computational Biology, Vol.6, 2011.
[30] WEKA at http://www.cs.waikato.ac.nz/~ml/weka.