Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors

Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, and Md. Kamrul Hasan
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
miloncsekuet@gmail.com, pranto00250@gmail.com, r.haque.249.rh@gmail.com and mhgolap11@gmail.com

***Abstract*— Breast Cancer is one of the most exquisite and internecine disease among all of the diseases in medical science. It is one of the crucial reasons of death among the females all over the world. We present a novel modality for the prediction of breast cancer and introduces with the Support Vector Machine and K-Nearest Neighbors which are the supervised machine learning techniques for breast cancer detection by training its attributes. The proposed system uses 10-fold cross validation to get an accurate outcome. The breast cancer termed as Wisconsin breast cancer diagnosis data set is taken from UCI machine learning repository. The performance of the proposed system is appraised considering accuracy, sensitivity, specificity, false discovery rate, false omission rate and Matthews correlation coefficient. The approach provides better result both for training and testing. Furthermore, the techniques achieved the accuracy of 98.57% and 97.14% by Support Vector Machine and K-Nearest Neighbors individually along with the specificity of 95.65% and 92.31% in testing phase.**

Keywords- Breast cancer; Prediction; Support Vector Machine; K-Nearest Neighbors; Performance Measure Indices.

I. INTRODUCTION

A recent statistics from World Cancer Research Fund International (WCRFI) [1] shows that breast cancer is the most widely recognized cancer in ladies around the world. There are almost 1.7 million new cases which are detected in 2012 that signifies around 12 percent of altogether new cancer cases and 25 percent of all cancers in ladies. Breast cancer outperformed the position of fifth for the reason of death in ladies. In numerous nations with advanced technology in medical science, the 5-year survival rate of initial phase breast cancer is 80–90%, dropping to 24% for breast cancer analyzed at progressive phase.

Breast cancer cells are found in the tissues of the breast. In very recent years there are numerous modalities have been developed for the prediction of breast cancer. In biopsy testing [2], the biopsy is occupied from the tissues of the breast. The test provides higher accurate result but the procedure to take the biopsy from breast is very painful and pathetic. So, the most of the patients are not intrigued with this test. A mammogram [3] is the most widely used technique for the detection of breast cancer which provides the 2D projection images of the breast. There are two kinds of mammography techniques that are widely used. These are the screen-film mammography (SFM) and digital mammography (DM). SFM is utilized in asymptomatic ladies breast i.e. problem free breast where it receipts two sights of both breasts. The time duration of screening mammograms (conventional mammography) is around 20 minute. It cannot detect benign cancer properly. Digital Mammography overcomes the problem of screening mammograms. It is related with computer system i.e. the data of digital mammography are kept in a computer. In DM, the images are taken and image processing methods are applied to improving the quality of the image. DM performs better in case of misdiagnosed cancer samples. Magnetic Reasoning Imaging (MRI) [4] is another most common technique for the diagnosis of breast cancer. Although MRI is a very complex test, sometimes it miscues some cancer whereas mammograms may detect. MRI is used for the ladies who have attacked in breast cancer to define the real size of the breast and find another disease in the breast. It provides an excellent result for 3D images and displays the dynamic functionality. MRI is done with the help of contrast-enhanced imaging.

Although several techniques have been introduced, none of the techniques are able to provide an accurate and reliable outcome. All of the modalities are involved with doctors or physicians or other medical staffs. So, a system which can operate without any medical equipment's and medical staffs may lead to an appropriate solution. We introduce an innovative modality to classify the input attributes according to the presence or absence of benign or malignant types of breast cancer. We have used two supervised machine learning techniques termed as Support Vector Machine and K-Nearest Neighbors which are related with learning computations that investigate data utilized for regression analysis and classification to identify the breast cancer.

The remaining part of the paper is planned as follows: Section II demonstrated the related works that have been done in this field. The theoretical explanation of the supervised machine learning techniques is investigated in Section III. The proposed methodology including the performance measure indices is illustrated in Section IV. The implementation and results analysis are represented in Section V and the conclusion of the paper is drawn in Section VI.

978-1-5386-2175-2/17/$31.00 ©2017 IEEE

II. Related Works

There are numerous modern techniques have been evolved with the evolution of technology for the prediction of breast cancer. The work related to this field is outlined shortly as follows.

Azar et al. [3] proposed a novel technique for the detection of breast cancer. The approach used three classification algorithms named as radial basis function (RBF), probabilistic neural networks (PNN) and multi-layer perceptron (MLP).The technique trained the attributes of breast cancer dataset and testing methodology also applied. The performance of the system is calculated in terms of some machine learning performance measure indices like accuracy, specificity, sensitivity etc. MLP achieved accuracy of 97.80% and 97.66% for training and testing respectively. The authors in [5] demonstrated a system for the identification of breast cancer which is applied for the two different Wisconsin Breast Cancer datasets using GA feature selection and Rotation Forest (RF). The Genetic Algorithm (GA) eliminates the unnecessary attributes of the data and provides the appropriate data which can speed up the system. Various machine learning techniques have been applied for classification purpose. The higher accuracy 99.48% is achieved by Rotation Forest model with GA-based attributes selection.

Ahmad et al. [6] proposed a technique named as GA-MOO-NN for the detection of breast cancer. The Genetic Algorithm is used for selecting the optimal attributes from the overall attributes. The approach divides the dataset into three parts named as training (50%), testing (25%) and validation (25%). The number of connections, hidden nodes size, number of selected features are also considered as a performance measure. The algorithms with the combination of objectives attained the accuracy of 98.85% and 98.10% individually in best and average cases. The paper also provides a comparison with the existing techniques that have been already introduced in this field. Islam et al. [7] developed a classifier for the diagnosis of breast cancer using the symbolic regression of Multigene Genetic Programming. The model uses 10 fold cross validation technique. The scheme provides a mathematical expression of the attributes of data set. The result presented in the paper is not clear whether it is training or testing. The accuracy obtained by the model is 99.28% along with 0.1303 RMSE.

The authors in [8] illustrated an innovative technique for the prediction of breast cancer by classifying the attributes of breast cancer dataset using a hybrid neuro genetic framework comprising of Genetic Algorithm and Training Feed Forward Back Propagation. The framework is trained according to leave one out fashion thus causes over fitting. The overall accuracy obtained by the framework is 97%. A comparison study is also shown in this paper. The most widely used machine learning techniques named as Random Forest (RF), Bayesian Networks (BN) and Support Vector Machine (SVM) are introduced in [9] for the detection of breast cancer. The system used 10 fold cross validation for the avoidance of over fitting. The authors implemented the system in WEKA and calculated the performance measure like as accuracy, recall, precision etc. The outcomes illustrated an accuracy rate of 97% along with the precision of 97.2% in Bayesian Networks.

III. Theoretical Considerations

Classification plays a major role in any pattern recognition problem. The classification model can be able to work with the linear or nonlinear problem. The Logistic Regression, Support Vector Machine (SVM) etc. are utilized for a linear problem. The K-Nearest Neighbors (K-NN), Kernel SVM, Random Forest etc. are used for the nonlinear problem.

A. Support Vector Machine

Support Vector Machine is a discriminative classifier that can be defined by a separating hyperplane. It is the generalization of maximal margin classifier which comes with the definition of hyperplane. In an n-dimensional space, the hyperplane is of (n-1) dimension with flat subspace that need not pass through the origin. The hyperplane is not visualized in higher dimension but the notion of an (n-1) dimensional flat subspace still applies [10]. If there doesn't exist any linearly separable hyperplane for any dataset, linear classifier can't be formed in that case. Kernel trick have to be applied to maximum-margin hyperplanes to develop nonlinear classifier. According this, nonlinear kernel function will be applied to the hyperplanes in replacement of dot product. Cubic, quadratic or higher-order polynomial function, Gaussian Radial basis function or Sigmoid function are forms of nonlinear kernel function. In p-dimensions, a hyperplane is described as follows.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0 \qquad (1)$$

where $\beta_0, \beta_1, \beta_2 \ldots \beta_p$ are the hypothetical values and X_p are the data points in sample space of p dimension.

B. K-Nearest Neighbors

K-Nearest Neighbors (K-NN) algorithm is another supervised machine learning technique used for classification and regression. K-NN doesn't make any assumptions on the fundamental data distribution. It performs great in pattern recognition and predictive analysis. For any new data point, firstly K-NN gather data points that are close to it. Any attributes that can vary on a large scale may have effective impact on the distance between data points [10]. The algorithm then sort those closest data points in terms of distance from the arrival data point. This distance can be measure in various way but Euclidian distance is the suggested one by experts. Next step is to take a specific number of data points whose distance are lesser among all and then categorize those data point. In K-NN, the number of closest data points are usually chosen as an odd number if the number of classes is 2. The category with highest number of data point will be the category of the new data point.

IV. The Proposed Methodology

A. Data Collection and Preparation

The breast cancer named as Wisconsin Breast Cancer (WBC) data set is retrieved from UCI machine learning repository dataset [11]. This dataset comprises of 699 instances, where the cases are labeled as either benign or malignant and 458 (65.50%) of the cases are benign and 241 (34.50%) are

malignant. The dataset is partitioned into two classes 2 and 4, where 2 denotes the benign class and 4 denotes the malignant class. The dataset has 11 features that are Clump Thickness(x_1), Uniformity of Cell Size(x_2), Uniformity of Cell Shape(x_3), Marginal Adhesion(x_4), Single Epithelial Cell Size(x_5), Bare Nuclei(x_6), Bland Chromatin(x_7), Normal Nuclei(x_8), Mitoses(x_9) except sample code number and class. The benign instances are represented as positive class and the malignant instances are represented as negative class in our study. There are 16 missing values of features in the data set. The missing features are substituted by the mean for that feature. Finally, the dataset is randomized to guarantee the correct circulation of data. The k-fold cross-validation is used where the given data set is split into k equal size chunks. A single chunk is used for testing and k-1 chunks is used for training.

B. The Performance Measure Indices

The performance of machine learning techniques is measured in terms of some performance measure indices. A confusion matrix for actual and predicted class is formed comprising of TP, FP, TN, and FN to evaluate the parameter. The significance of the terms is given below.

TP = True Positive (Correctly Identified)

TN = True Negative (Incorrectly Identified)

FP = False Positive (Correctly Rejected)

FN = False Negative (Incorrectly Rejected)

The performance of the proposed system is measured by the following formulas:

$$Accuracy\ (Acc) = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2)$$

$$Sensitivity\ (Sen) = \frac{TP}{(TP+FN)} \quad (3)$$

$$Specificity\ (Spec) = \frac{TN}{(TN+FP)} \quad (4)$$

$$False\ Discovery\ Rate\ (FDR) = \frac{FP}{(FP+TP)} \quad (5)$$

$$False\ Omission\ Rate\ (FOR) = \frac{FN}{(FN+TN)} \quad (6)$$

$$Matthews\ Correlation\ Coefficient\ (MCC) = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (7)$$

V. Implementation and Results Analysis

We have proposed a model using Kernel Support Vector Machine and K-Nearest Neighbors which is implemented in a high configuration computer. The computer configuration was Intel Core i7 with 8GB RAM. We have used Scikit-learn which is an open-source software developed in Python for machine learning library. An Integrated development environment named as Spyder is used to run the program.

We have utilized the 10-fold technique i.e. the data set was split into 10 chunks. The 10 fold technique is utilized to approve the methodical model. 9 folds are utilized for training and the rest one for testing in 10 fold cross validation. We have formed a confusion matrix from the classifier. We have utilized 629 (90%) instances of total data for training both in Support Vector Machine and K-Nearest Neighbors individually. The remaining 70 (10%) instances used for testing both in SVM and K-NN individually. The graphical representation of the confusion matrix for each modality is illustrated in Fig. 1. Fig. 1 depicts that the correctly identified value is comparatively higher in SVM and K-NN for training and testing. Among the 629 instances 409 and 404 is correctly identified in SVM and K-NN respectively. The performance measures indices are calculated both for training and testing using the above-described equation. The calculated values are depicted in Table I and the graphical view of the performance measure indices is illustrated in Fig. 2.

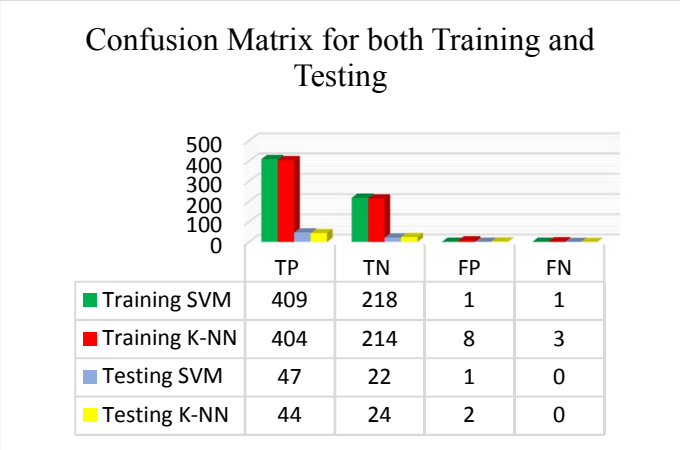

Fig. 1. Confusion matrix both in training and testing phases for the prediction of breast cancer.

TABLE I. PERFORMANCE MEASURE INDICES

Parameters	Training Phase		Testing Phase	
	SVM	K-NN	SVM	K-NN
Accuracy (%)	99.68	98.25	98.57	97.14
Sensitivity (%)	99.76	99.26	100	100
Specificity (%)	99.54	96.40	95.65	92.31
Geometric Mean of Sensitivity and Specificity (%)	99.65	97.83	97.83	96.16
False Discovery Rate (%)	0.24	1.94	2.08	4.35
False Omission Rate (%)	0.46	1.38	0	0
Matthews Correlation Coefficient	0.99	0.96	0.97	0.94

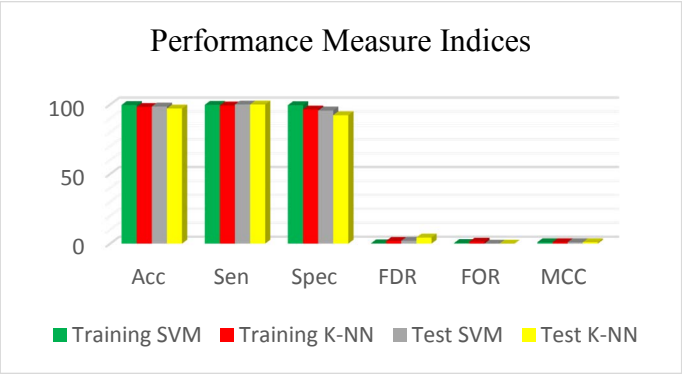

Fig. 2. Performance measure indices both in training and testing phases for the prediction of breast cancer.

The result presented in Table I and Fig. 2 shows that SVM has the best performance over K-NN in terms of specificity, sensitivity, and accuracy. The accuracy, sensitivity, and specificity are obtained by SVM are 99.68%, 99.76%, and 99.54% respectively in training phase. The Matthews Correlation Coefficient value towards 1 indicates a higher possibility of being a pure binary classifier. We found the Matthews Correlation Coefficient value of 0.99 for SVM and 0.96 for K-NN in training phase. This includes that both the classifier are a correct binary classifier. The false omission rate is near about zero for each modality in training and testing phase. The false discovery rate is comparatively high in K-NN rather than SVM. The SVM performs comparatively better than K-NN.

A comparison study is drawn in Table II for testing phase only. In [12], the authors measured the performance of standard SVM (St-SVM) and compared with the variants of SVM. The highest accuracy, sensitivity and specificity obtained by LP-SVM, L-SVM, SSVM and NSVM are 97.1429%, 98.2456% and 96.5517% respectively in testing phase. The highest accuracy achieved by our model 98.57% and 97.14% in SVM and K-NN individually. The St-SVM shows lower performance in testing phase. The authors in [12] used 4 fold technique where we utilized 10 fold technique.

TABLE II. COMPARISON WITH EXISTING METHODS IN TESTING PHASE

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)
LPSVM[12]	97.1429	98.2456	95.082
LSVM [12]	95.4286	96.5217	93.33
SSVM[12]	96.5714	96.5812	96.5517
PSVM[12]	96	97.3684	93.4426
NSVM[12]	96.5714	96.5812	96.5517
St-SVM [12]	94.86	95.65	93.33
Proposed model using SVM	**98.57**	**100**	**95.65**
Proposed model using K-NN	**97.14**	**100**	**92.31**

VI. CONCLUSION

Breast cancer prediction is very significant in the area of Medicare and Biomedical. In this paper we focused on building a classifier which aims at predicting the most severe cancer known as breast cancer. Breast cancer is a remarkably risky disease that causes a lot of death for numerous ladies all over the world. So, early detection of this cancer can save a lot of valuable life. We proposed a model that predict the breast cancer based on Support Vector Machine and K-Nearest Neighbors. The SVM has been implemented by the Python to be the most effective in classifying the diagnostic data set into the two classes in view of the seriousness of the cancer. We end up with an accuracy of 99.68% in SVM in training phase. The proposed model will be very helpful for the medical staffs as well as general people. The classifier obtained by supervised machine learning techniques will be very supportive in the field of medical disorders and proper diagnosing.

REFERENCES

[1] Breast cancer statistics. [Online]. Available: http://www.wcrf.org/int/cancer-facts-figures/data-specific-ancers/breast-cancer-statistics, accessed on: Aug. 25, 2017.

[2] Arbab Masood Ahmad, Gul Muhammad, Khan, S.Ali Mahmud, Julian F. Miller, “Breast Cancer Detection Using Cartesian Genetic Programming evolved Artificial Neural Networks,” Philadelphia, Pennsylvania, USA, GECCO’12, July 7–11, 2012.

[3] Ahmad Taher Azar, Shaimaa Ahmed El-Said, “Probabilistic neural network for breast cancer classification,” Neural Computing and Applications, Springer, vol. 23, pp.1737-1751, 2013.

[4] Warner E, Messersmith H, Causer P et al, “Systematic review: using magnetic resonance imaging to screen women at high risk for breast cancer,” Annals of Internal Medicine,148(9):671–679,06 May, 2008.

[5] Emina Alic˘kovic´, Abdulhamit Subasi, “Breast cancer diagnosis using GA feature selection and Rotation Forest” Neural Computing and Applications, Springer, Volume 28, issue 4, pp 753–763, April 2017.

[6] Fadzil Ahmad, Nor Ashidi Mat Isa, Zakaria Hussain, Siti Noraini Sulaiman, “A genetic algorithm-based multi-objective optimization of an artificial neural network classifier for breast cancer diagnosis”, Neural Computing and Applications, Springer, Volume 23, Issue 5, pp 1427–1435 ,October 2013,.

[7] M. K. Hasan, M. M. Islam and M. M. A. Hashem, "Mathematical model development to detect breast cancer using multigene genetic programming," 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, pp. 574-579, 2016.

[8] 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. 144-149, 2017.

[9] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, pp. 1-4, 2016.

[10] Gareth James and Daniela Witten and Trevor Hastie and Robert Tibshirani, An Introduction to Statistical Learning, 1st ed. , 2013.

[11] Breast Cancer Wisconsin (Original) Data Set, [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data, accessed on: Aug. 25, 2017.

[12] Ahmad Taher Azar, Shaimaa Ahmed ,El-Said ,“Performance analysis of support vector machines classifiers in breast cancer mammography recognition” Neural Computing and Applications, Springer, Volume 24, Issue 5, pp 1163–1177, April 2014.