# A Precise Breast Cancer Detection Approach Using Ensemble of Random Forest with AdaBoost

**Conference Paper** · July 2019

DOI: 10.1109/IC4ME247184.2019.9036697



**5 authors**, including:

Tanbin Islam Rohan

**5** PUBLICATIONS **1** CITATION

SEE PROFILE

Abu Bakar Siddik

Khulna University of Engineering and Technology

**5** PUBLICATIONS **2** CITATIONS

SEE PROFILE

Monira Islam

Khulna University of Engineering and Technology

**20** PUBLICATIONS **88** CITATIONS

SEE PROFILE

Md.Salah Uddin Yusuf

Khulna University of Engineering and Technology

**37** PUBLICATIONS **104** CITATIONS

SEE PROFILE

International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 11-12 July, 2019

# A Precise Breast Cancer Detection Approach Using Ensemble of Random Forest with AdaBoost

Tanbin Islam Rohan[1], Awan-Ur-Rahman[2], Abu Bakar Siddik[3], Monira Islam[1], Md. Salah Uddin Yusuf[1]
[1]*Dept. of EEE,* [2]*Dept. of CSE,* [3]*Dept. of Biomedical Engineering*
*Khulna University of Engineering & Technology (KUET)*
Khulna-9203, Bangladesh
tanbinislam009@gmail.com, awanrahman55@gmail.com, yeashasiddik@gmail.com,
monira_kuet08@yahoo.com, ymdsalahu2@gmail.com

***Abstract*—Due to breast cancer, a number of women die every year. With an early diagnosis, breast cancer can be cured. Prognosis and early detection of cancer types have become a necessity in cancer research. Thus, a reliable and accurate system is required for the classification of benign and malignant tumor types of breast cancer. This paper explores a supervised machine learning model for classification of malignant and benign tumor types from Wisconsin Breast Cancer dataset retrieved from UCI machine learning repository. The dataset has 458 (65.50%) of benign data and 241 (34.50%) of malignant data, the total of 699 instances, 11 features and 10 attributes. Random Forest (RF) ensemble learning method is implemented with AdaBoost algorithm manifest improved metrics of performance in binary classification between tumor classes. For more accurate estimation of model prediction performance, 10-fold cross-validation is applied. The structure provided accuracy of 98.5714% along with sensitivity and specificity of 100% and 96.296% respectively in the testing phase. Matthews Correlation Coefficient is calculated 0.97 which validates of the structure being a pure binary classifier for this work. The proposed structure outperformed conventional RF classifier for classifying tumor types. Additionally, this model enhances the performance of conventional classifiers.**

***Keywords—breast cancer, ensemble technique, Random Forest, AdaBoost, 10-fold cross-validation***

## I. INTRODUCTION

Breast cancer is one of the leading cancer types in women around the world. There are two types of breast cancer tumors namely benign (non-cancerous) and malignant (cancerous). In a recent study of World Cancer Research Fund International (WCRFI) showed, breast cancer is increasing by 1.7% per year in Asia/Pacific Islander [1]. Though for the past 2 decades, cancer death rates have been constantly decreasing. Between 1991 and 2011, the risk of dying from cancer declined by 22% [2]. Nevertheless, for improving breast cancer results and survivals, a vital step would be early detection. And one of the early detection strategies is screening. Screening method includes breast self-exam, clinical breast exam which is conducted by trained health professionals and mammography. Mammography represents 2D projection images of the breast which uses low-energy X-rays to identify abnormalities in the breast. A mammogram is one of the most extensively used method for the detection of breast cancer [3]. Digital mammograms have variances in the types of tissues and low contrast [4-6]. Calcification clusters and preliminary signs of masses are important visual clues of breast cancer. Breast tumors are of different sizes, shapes and few of them have the same characteristics of normal tissue. And in the early stages of breast cancer, these tumors varied in form, and these signs are delicate. Sadly, it is hard for the specialists making the diagnosis more challenging. Specialists sometimes fail to differentiate between benign and malignant tumors. This can lead to false biopsy tests and cause problems for the individual.

Several studies have been conducted in the past years. In [7], support vector machine (SVM) model was used for recurrence estimation of breast cancer within 5 years in Korean population after breast cancer surgery. The SVM-based prediction model has beaten other prognostic models with accuracy (84.58%) sensitivity (0.89), specificity (0.73), positive predictive values (0.75), and negative predictive values (0.89). In [8], semi-supervised learning (SSL) algorithm was used as labeled patient records are not easy to collect. The model delivered a mean area under the curve of 0.81 and a mean accuracy of 76%. But, the authors appraised that more labeled data can lead to better performance. In [9], the risk of breast cancer was predicted by training and testing a 3-layer feedforward ANN with 1000 hidden-layer nodes by using 10-fold cross-validation. The ANN was well calibrated using the H-L goodness-of-fit P-value of 0.13 and the network showed superior performance. In [10], Artificial Neural Network (ANN), Decision Tree (C4.5), and Support Vector Machine (SVM) were used to establish the prognostic models. The accuracy of the techniques, DT, ANN, and SVM are 0.936, 0.947 and 0.957 respectively. In [11], a study on SEER datasets was conducted with ANN, SVM, and SSL algorithms. The authors claimed the best performance in the measure of accuracy, specificity, and sensitivity was SSL with a mean accuracy of 0.71, a mean sensitivity of 0.76, and a mean specificity of 0.65. The SVM model had the best AUC performance with a mean of 0.80. Authors in [12], proposed a (CAD) system for classifying benign and malignant tumors from image.

Finally, we find that differentiating between malignant and benign tumor is problematic which scopes to rethink about the classification mechanism. Therefore, in this paper, a boosting algorithm, AdaBoost has been deployed on the features to yield more discriminative characteristics between two classes so that the classifier can classify the two classes more accurately. After implying the AdaBoost algorithm, the

**978-1-7281-3060-6/19/$31.00 ©2019 IEEE**

Random Forest classifier was used which outperformed classification accuracy with the improved sensitivity, specificity and other performance measures. The algorithm used the ensemble technique and exhibited better performance. The proposed method showed enhanced result.

Rest of the paper is organized accordingly as section II describe the proposed methodologies for breast cancer detection. Results and discussion are explained in Section III. Finally, section IV concludes the result.

## II. Proposed Methodology

In breast cancer detection classification shows a very vital role. The classification algorithms in Machine Learning can work with linear or nonlinear problems. We generally use Logistic Regression, Naive Bayes for the linear problems and SVM (Support Vector Machine), Decision Trees, Boosted Trees, Random Forest, Neural Networks, Nearest Neighbor for the nonlinear problems.

### A. Data Collection and Provision

From the UCI machine learning repository [13], the data set called Wisconsin Breast Cancer (WBC) is retrieved. WBC data set consists of 699 instances, 10 attributes and the cases are contrasted as either benign or malignant. In the whole dataset, there are 458 (65.50%) of benign data and 241 (34.50%) of malignant data. The dataset is divided in two classes, namely 2 and 4 denoting benign class and malignant class respectively. The dataset consists of 11 features those are sequentially, 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. In this study, benign instances and malignant instances are represented with affirmative class and negative class. In the data set, 16 missing elements of features are replaced by the mean for that feature. For splitting the data set into 10 equal size portions, 10-fold cross validation is used. Ultimately, for the correct circulation, the dataset is processed. 9 portions are used for training and a single portion is used for testing. A block diagram in Fig.1 exhibits the proposed methodology.

### B. AdaBoost

AdaBoost is the short form of "Adaptive Boost". In the classification problems, the algorithm converts a bunch of weak classifiers into a strong one. On various weighted training examples, the classifier should be trained iteratively. For every iteration, the classifier provides fruitful result by minimizing the training error.

At first, the AdaBoost selects a training subset casually. By selecting the training set based on the accurate prediction of the last training, the algorithm iteratively trains the AdaBoost machine learning model. For the high probability of classification from the next iteration, it gives the higher weight to wrong classified observations. According to the accuracy of the classifier, it gives the weight to the trained classifier in every iteration so that the more accurate classifier will get high weight. The process will be continued until the whole training data fits without any error or until reached to the specified maximum number of estimators. The mathematical parameters and formulas are as follows,

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \tag{1}$$

Where $H(x)$ symbolizes weight-age of input training data, $h_t(x)$ refers to the output of weak classifier $t$ for input $x$, and $\alpha_t$ denotes the weight apportioned to the classifier. Now, $\alpha_t$ is calculated as follows:

$$\alpha_t = 0.5 * \ln(\frac{1-E}{E}) \tag{2}$$

So, the weight of classifier is calculated based on the error rate E. In the beginning, all input training example has equivalent weight-age. Afterward, a weak classifier is trained, the weight of each training example is updated with the following formula,

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t} \tag{3}$$

Where $D_t$ corresponds to the weight of the previous layer. The weights are then normalized by dividing each of them by the sum of all weights, $Z_t$. And $y_i$ is the $y$ level of training point ($x_i$, $y_i$).

### C. Random Forest

Random Forest is one of the popular supervised learning algorithms. Random forest algorithm generates decision trees on randomly nominated data examples. Then it gets a prediction from each tree for selecting the best solution by the means of voting. Through it constructs a decision tree for each sample and gets a prediction result from each decision tree then it performs a vote for each predicted result. The final prediction comes by selecting the prediction result with the most votes. Random forest uses the Gini coefficient taken from the CART (Classification and- Regression Trees) learning system to construct decision trees. The Gini coefficient calculates the dissimilarity between values of a frequency distribution. A Gini coefficient of zero states perfect likeness and a coefficient value of 1 expresses maximal inequality between values. If a dataset $T$ contains examples from $n$ classes and $P_j$ is the relative frequency. Then, the Gini coefficient is defined as:

$$Gini(T) = 1 - \sum_{j=1}^{n} (P_j)^2 \tag{4}$$

### D. Parameters of Performance Measurement

According to the machine learning approaches, the performance of the technique is measured by some parameters. A confusion matrix is formed with *True Positive* (*TP*) or correctly identified, *True negative* (*TN*) or incorrectly identified, *False Positive* (*FP*) or correctly rejected, and *False Negative* (*FN*) or incorrectly rejected parameters for actual and predicted class.

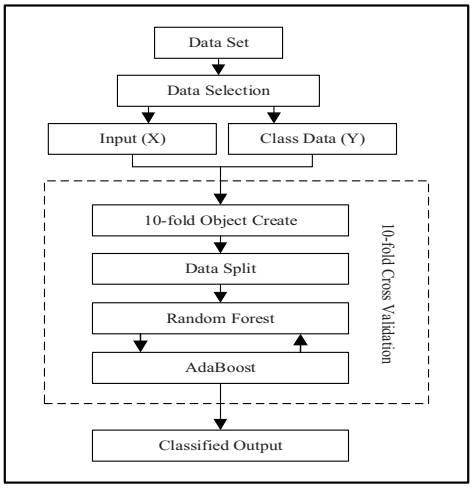


Fig. 1. Block diagram of the proposed approach.

The proposed system's performance can be calculated using (5)-(10).

$$Accuracy(Acc) = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100\% \quad (5)$$

$$Sensitivity(Sen) = \frac{TP}{(TP + FN)} \times 100\% \quad (6)$$

$$Specificity(Spec) = \frac{TN}{(TN + FP)} \times 100\% \quad (7)$$

$$FDR = \frac{FP}{(TP + FP)} \times 100\% \quad (8)$$

$$FOR = \frac{FN}{(FN + TN)} \times 100\% \quad (9)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(FN + TN)(FP + TN)(FN + TP)(FP + TP)}} \quad (10)$$

In the above equations, *FDR* represents *False Discovery Rate*, *FOR* denotes *False Omission Rate*, and *MCC* means *Matthews Correlation Coefficient*.

## III. Results and Discussion

### A. Extracted Parameters for Performance Measurement

In this work a precise breast cancer detection technique is proposed. Random Forest classifier was ensembled with. AdaBoost algorithm, and implemented on a computer configured with 16 GB RAM and Intel Core i7 processor. An open-source tool named Scikit-learn have been used which is developed in Python language for machine learning library. An open source cross-platform IDE named Spyder which is developed in Python language for scientific programming has been used as well.

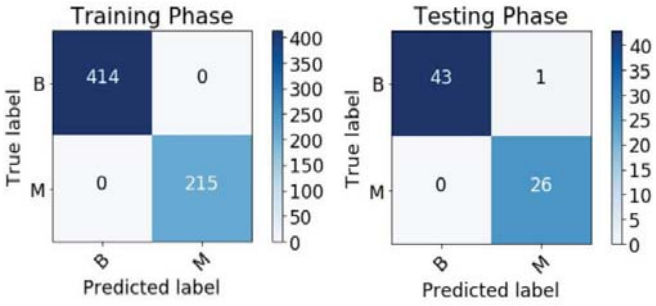


Fig. 2. Performance measurement matrix for training and testing phase in breast cancer classification.

Random Forest algorithm has been used as the base estimator of the AdaBoost classifier. 50 base estimator have been deployed for ensemble modeling. The data have been split into 629 (90%) instances for training and rest of the 70 (10%) instances for testing purpose. 10 fold cross-validation is used for the model.

Fig.2 shows a confusion matrix for both training and testing phase. In the training phase, 629 instances out of 629 are accurately identified. And in the testing phase, 69 instances among 70 instances are precisely identified. Fig. 2 illustrates that the numbers of correctly identified instances which are very high compared to the numbers of incorrectly identified instances. Table I demonstrates the performance of the classifier for training and testing phase on the basis of the percentage calculated by using the equations described above and Fig. 3 illustrates the graphical representation of the training and testing phase. From Fig. 3 it is clear that the classifier has 98.5714% accuracy in the test phase. It has 100% sensitivity (true positive rate) which indicates that the model can correctly identify those with diseases, 96.296% specificity (true negative rate) which indicates that the model can correctly identify those without diseases, 2.27% False Discovery Rate (FDR) and 0% False Omission Rate (FOR). Matthews Correlation Coefficient (MCC) is calculated to be 0.97, indicating a higher likelihood of a binary classifier.

### B. Performance Comparison of the Proposed Work

For better performance with AdaBoost algorithm, the training data needs to be of high quality and noise free. The accuracy of conventional Random Forest classifier and Random Forest with AdaBoost ensemble algorithm has been calculated. The reason behind using AdaBoost ensemble technique with a base estimator is that boosting algorithm produces strong classifiers out of conventional weak binary classifiers. Forest classifier attains accuracy just beyond random chance on a classification problem. Fig. 4 shows the comparison of accuracy between Random Forest classifier with AdaBoost and without AdaBoost ensemble algorithm.

TABLE I. Parameters for Performance Measurement of Breast Cancer Detection

| Model Parameters | Training Phase | Testing Phase |
|---|---|---|
| Accuracy (Acc) (%) | 100 | 98.5714 |
| Sensitivity (Sen) (%) | 100 | 100 |
| Specificity (Spec) (%) | 100 | 96.296 |
| False Discovery Rate (FDR) (%) | 0 | 2.27 |
| False Omission Rate (FOR) (%) | 0 | 0 |
| Matthews Correlation Coefficient (MCC) | 1 | 0.97 |

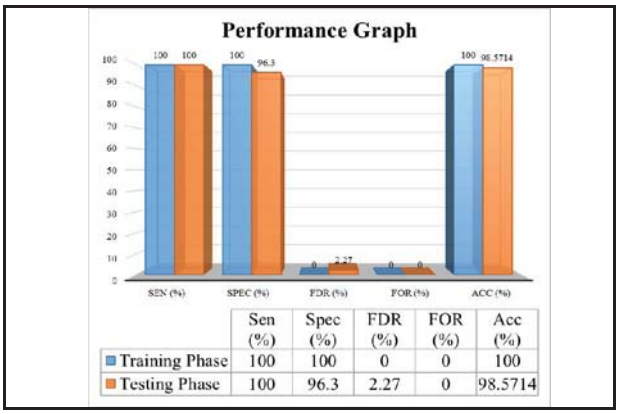


Fig. 3. Performance parameters of the proposed method for both in the training phase and testing phase.

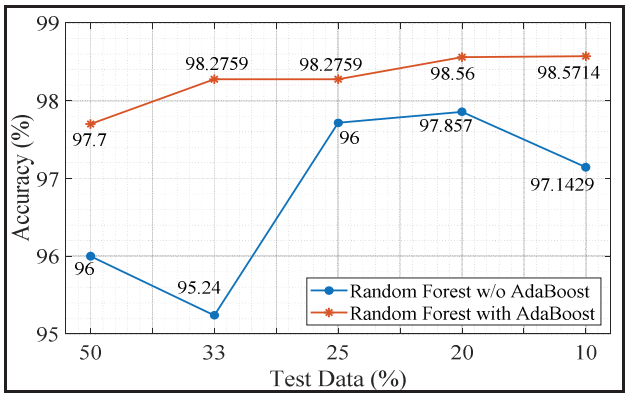


Fig. 4. Accuracy comparison between Random Forest with AdaBoost and without AdaBoost for different test data size.

TABLE II. COMPARISON OF THE PROPOSED APPROACH WITH OTHER EXISTING WORK

| Methods | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|
| LSVM [14] | 95.4286 | 96.5217 | 93.33 |
| PSVM [14] | 96 | 97.3684 | 93.4426 |
| SSVM [14] | 96.5714 | 96.5821 | 96.5517 |
| NSVM [14] | 96.5714 | 96.5812 | 96.5517 |
| LPSVM [14] | 97.1429 | 98.2456 | 95.082 |
| St-SVM [14] | 94.86 | 95.65 | 93.33 |
| Proposed Method | 98.5714 | 100 | 96.296 |

Without using AdaBoost, Random Forest shows poor testing result by varying the training and testing data percentage. By using AdaBoost along with Random Forest illustrates a better performance. It is also clear that for different test data split ratio, the accuracy of Random Forest varies and doesn't have a stable accuracy pattern. But, the model provides better accuracy measure in each case. In Table II, there is a comparison of the performance of different methods and our proposed method in the testing phase. In [14], the authors got the highest testing accuracy of 97.1429% by using LPSVM. The highest sensitivity and specificity are 98.2456 and 96.5517 by using LPSVM, NSVM, and SSVM respectively using 4-fold cross-validation technique. Our proposed model has got 98.5714% testing accuracy, 100% sensitivity, and 96.296% specificity.

## IV. CONCLUSION

Decision making in medical conditions can be sometimes a troublesome work and likely to go wrong. In these matters, machine learning classifiers yield an accurate decision. This paper presents an approach for binary classification between a malignant tumor and benign tumor in breast cancer. According to statistical data, one of the most prevalent and prevailing cancer types is breast cancer. But this cancer type is also curable if it is diagnosed early. Random forest along with AdaBoost structure had given a very promising result in classifying breast cancer. Previous studies have shown that machine learning methods are very favorable in prognoses compared to conventional statistical methods and even sometimes expert-based systems. This result strongly suggests that the proposed approach will play a significant role in the proper diagnosis of breast cancer.

## REFERENCES

[1] C. E DeSantis, J. Ma, A. G. Sauer, L. A. Newman, A. Jemal, "Breast cancer statistics, 2017, racial disparity in mortality by state," CA: a cancer journals for clinicians, vol. 67, no. 6, pp. 439-448, Nov. 2017.

[2] R. L. Siegel, K. D. Miller, A. Jemal, "Cancer statistics, 2015," CA: a cancer journals for clinicians, vol. 65, no. 1, pp. 5-29, Jan-Feb. 2015.

[3] A.T. Azar, S.A. El-Said, "Probabilistic neural network for breast cancer classification," Neural Computing and Applications, vol. 23, no. 6, pp. 1737-1751, Nov. 2013.

[4] P. C. Gøtzsche, M. Nielsen, "Screening for breast cancer with mammography," The Cochrane database of systematic reviews, vol. 19, no. 1, Jan. 2011.

[5] E. D. Pisano, C. Gatsonis, E. Hendrick et al., "Diagnostic performance of digital versus film mammography for breast-cancer screening," The New England journal of medicine, vol. 353, no. 17, pp. 1773-1783, Oct. 2005.

[6] K. Kerlikowske, D. Grady, S. M. Rubin, C. Sandrock, V. L. Ernster, "Efficacy of screening mammography. A meta-analysis," JAMA, vol. 273, no. 2, pp. 149-154, Jan. 1995.

[7] W. Kim, K. S. Kim, J. E. Lee, D. Y. Noh, S. W. Kim, Y. S. Jung, M. Y. Park, R.W. Park, "Development of novel breast cancer recurrence prediction model using support vector machine," Journal of breast cancer, vol. 15, no. 2, pp. 230-238, Jun. 2012.

[8] J. Kim, H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," Journal of the American medical informatics association, vol. 20, no. 4, pp. 613-618, Jul-Aug. 2013.

[9] T. Ayer, O. Alagoz, J. Chhatwal, J. W. Shavlik, C. E. Kahn Jr, E. S. Burnside, "Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration," Cancer, vol. 116, no.14, pp. 3310-21, 2010.

[10] A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, A. R. Razavi, L. G. Ahmad, "Using three machine learning techniques for predicting breast cancer recurrence," Journal of health & medical informatics, vol. 4, no. 2, Apr. 2013.

[11] K. Park, A. Ali, D. Kim, Y. An, M. Kim, H. Shin, "Robust predictive model for evaluating breast cancer survivability," Engineering applications of artificial intelligence, vol. 26, no. 9, pp. 2194-2205, Oct. 2013.

[12] D.A. Ragab, M. Sharkas, S. Marshal, J. Ren, "Breast cancer detection using deep convolutional neural networks and support vector machines", PeerJ 7, e6201, Jan 28, 2019.

[13] O. L. Mangasarian, R. Setiono, W.H. Wolberg, "Pattern recognition via linear programming: Theory and application to medical diagnosis," 1990. Available in: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original).

[14] A. T. Azar, S. A. El-Said, "Performance analysis of support vector machines classifiers in breast cancer mammography recognition," Neural computing and applications, Springer, vol. 24, no. 5, pp. 1163–1177, Apr. 2014.