MAT JOURNALS

Journal of Data Mining and Management
e-ISSN: 2456-9437
Volume 5 Issue 1

An Ensemble Classifier Approach for Diagnosis of Breast Cancer

Anupama Y.K.*[1], Amutha. S[2], Ramesh Babu. D. R.[2]
[1]Assistant Professor, Department of Computer Science and Engineering, Dayananda Sagar College of Engineering, VTU University, Bangalore, Karnataka, India
[2]Professor, Department of Computer Science and Engineering, Dayananda Sagar College of Engineering, VTU University, Bangalore, Karnataka India

INFO

Corresponding Author
E-mail Id:
**anupamayk@gmail.com*

DOI:

Cite as:

ABSTRACT

Accurate and early diagnosis of breast cancer increases survival rate of patients. Diagnosis of Breast cancer involves identifying tumour as either benign or malignant. In this paper, proposed methodology is an integration of ensemble classifiers AdaBoost and Random Forest named as ADARF a prediction model for diagnosis of breast cancer. The main objective is to enhance the performance and to reduce error. Experimental result shows that the proposed approach has higher accuracy of 98.8% compared to Logistic Regression (LR), K Nearest Neighbour (KNN) and Support Vector Machine (SVM) classifiers.

Keywords: *AdaBoost, ADARF, ensemble, KNN, LR, random forest, SVM*

INTRODUCTION

The occurrence of breast cancer is most prevalent among women. Rapid multiplication of cells in breast tissue causes breast tumour. Breast tumours can be categorized as cancerous (Malignant) and noncancerous (Benign) tumour. When the cancerous breast tumour cells adhesion breaks down, starts spreading to the other tissues and organs of the body. Cells of noncancerous breast tumour does not affect to the surrounding tissues. Early stage detection of Cancerous breast tumour can increase 90% survival rate of patients. These patients can live for minimum of 5 years. Whereas, Cancerous breast tumour detected at later stages have lower survival rate of 15% [1]. Nowadays, the decision tree is extensively used in the field of medical. Most of the studies use decision trees to extract the patterns from clinical data sets. AdaBoost is a well-known method due to its low error rate and good performance. As the successor of the boosting algorithm, AdaBoost is used to integrate weak classifiers to get a model with higher prediction results [2]. Proposed ensemble classifier ADARF is an integration of AdaBoost and Random Forest for predicting breast cancer outcome on the collected Wisconsin diagnostic breast cancer dataset (WDBC).

MATERIALS AND METHODS

This section explores the breast cancer dataset used for implementation and the way of proposed methodology implemented.

Dataset

WDBC data set obtained from the repository of UCI [3]. WDBC has benign and malignant data instances. Out of 569, 212 of the cases are malignant and 357 are benign cases. The attributes of the dataset consist of:

ID number: Identification number of patients.

Diagnosis: Depending on 10 real valued features that are evaluated from each single cell nucleus used for diagnosis of breast cancer as shown in Table 1:

Table 1: *Breast cancer dataset attributes.*

Sl. No.	Features
1.	Radius
2.	Texture
3.	Perimeter
4.	Area
5.	Smoothness
6.	Compactness
7.	Concavity
8.	Concave points
9.	Symmetry
10.	Fractal dimension

PROPOSED METHODOLOGY

Proposed methodology carried out as shown in Fig. 1. Data pre-processing is performed on WDBC to check for missing data in the dataset. PCA (Principal Component Analysis) algorithm applied on the cleaned (pre-processed) dataset to perform dimensionality reduction. The PCA algorithm enables us to get the principal components from a set of possibly correlated variables by performing orthogonal transformation. These principal components are uncorrelated eigenvectors, each representing some proportion of variance in the data.

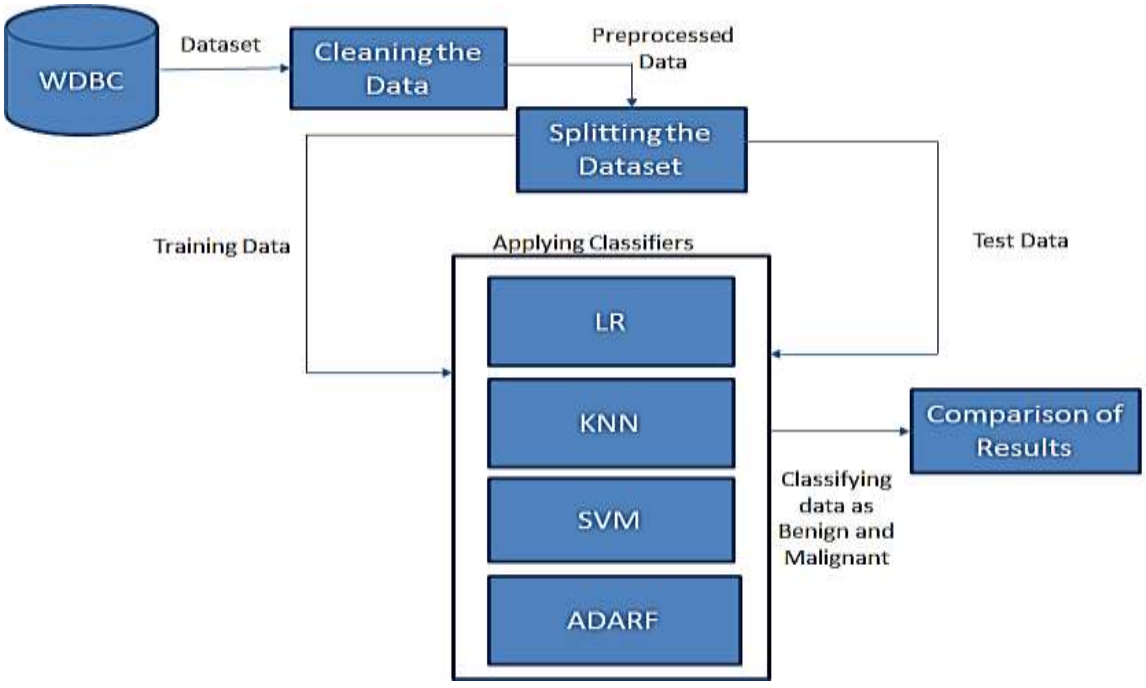

Figure 1: *Architecture of proposed methodology.*

Let X=$\{x_i\}$ $_{i=1}^{m}$ be the training data where x_i represents a tuple with dimension D. PCA extracts the most relevant attribute from x_m and compress the dimensionality by retaining only the most relevant attribute. PCA is an orthogonal projection of the original D-dimensional data onto a new 2D-dimensional space and minimizes the variance of the projected data.

Dataset partitioned into training dataset and testing dataset. AdaBoost is a boosting ensemble model that can be used to boost the weak learner. Random Forest considered as the weak learner. Since Random Forest used as the weak learner for AdaBoost, first the Random Forest is trained. The error rates associated with weights are computed. Weights of the wrongly classified ones are updated. Finally, AdaBoost prediction model diagnose the breast tumour as either cancerous or noncancerous. This technique has improved the accuracy of classifier.

LITERATURE REVIEW

Majority of women in current generation are suffering from cancerous breast tumour. Scientists and Researchers have conducted several experiments regarding the breast cancer diagnosis. Doctors examine tumour to identify whether it is cancerous or noncancerous tumour. Exploring on different techniques applied to breast cancer dataset showed that there are several studies on the early detection and prevention of cancerous breast tumour using machine learning techniques.

Md. Milon et al. applied SVM and KNN classifiers to Wisconsin Diagnostic Breast Cancer dataset (WDBC) for 629 instances [4]. The proposed method SVM performed better compared to the other variants of SVM and KNN algorithm with an accuracy of 98.57%. Reem Alyami et al. implemented Support Vector Machines and Artificial Neural Network on WDBC for 629 instances [5]. In this study SVM has outperformed ANN with accuracy of 97.14%.

A novel method proposed to identify the breast cancer by Moh'd Rasoul Al-hadidi et al. has higher accuracy [6]. Method has two major portions. First portion uses the image processing techniques to obtain relevant feature patterns from the mammography images. Feature patterns obtained are taken as input for Back Propagation Neural Network (BPNN) model and LR model. Afterwards accuracy results of both models are compared. It is observed in results that the number of feature patterns used in LR model was more than with the BPNN. Better regression value 93% is obtained using BPNN with 240 attribute features.

U. Karthik Kumar et al. explored the performance generalization of J48, Naïve Bayes, and SVM classifiers [7]. These classifiers analysed to improve the system of decision making for the survivability of breast cancer patients. In this study a novel classifier for voting is developed. New voting classifier is a combination of three classifiers for the prediction of breast tumour.

A combination of SVM and Ensemble classifier proposed by Haifeng Wang et al. [8]. In this paper, Weighted Area under the Receiver Operating Characteristic Curve is used to check the performance of the classifier technique. This model achieved a higher accuracy around 97% for WDBC than the common ensemble methods adaptive boosting and bagging, performing better than the individual models on small datasets.

Decision tree algorithm used by Lavanya et al. to increase the classification accuracy of breast Cancer dataset [9, 10]. 10-fold cross validation method applied on the training dataset. Then feature selection methods applied to remove features that have no relevance in the process of classification. Bagging applied to the training dataset to improve the accuracy of decision tree classifier.

EXPERIMENTAL RESULTS

Confusion Matrix, Performance Metrics and Receiving Operating Characteristic (ROC) curve used to compare the results obtained from LR, KNN, SVM and ADARF classifiers.

Confusion Matrix (CF)

The confusion matrix is a tool commonly used to represent performance of classifiers in classification tasks visually [11]. The efficiency of classifiers is computed with a number of correctly classified and misclassified instances from each value of attributes being classified in CF [12]. CF computes the parameters true positive (TPO), false positive (FPO), true negative (TNE) and false negative (FNE) as shown in Table 2.

Table 2: *Confusion matrix.*

	Predicted	
Actual	Benign	Malignant
Benign	Tpo	Fpo
Malignant	Fne	Tne

Performance Evaluation Metrics

Performance metrics used for evaluation of prediction models are Accuracy, sensitivity, and specificity.

Accuracy (Acc) is the percentage of correctly classified outcomes as positive and negative among the test set data that has been evaluated.

$$Acc = \frac{TPO + TNE}{TPO + FPO + TNE + FNE}$$

Sensitivity (Sens) is the percentage of data being correctly identified as the ones with the cancer (true positive rate).

$$Sens = \frac{TPO}{TPO + FNE} X100$$

Specificity (Spec) is the percentage of correct detection of the instances, those without the disease (true negative rate).

$$Spec = \frac{TNE}{(TNE + FPO)} X100$$

LR, KNN SVM and ADARF models performance evaluated using accuracy metrics as shown in Fig. 2.

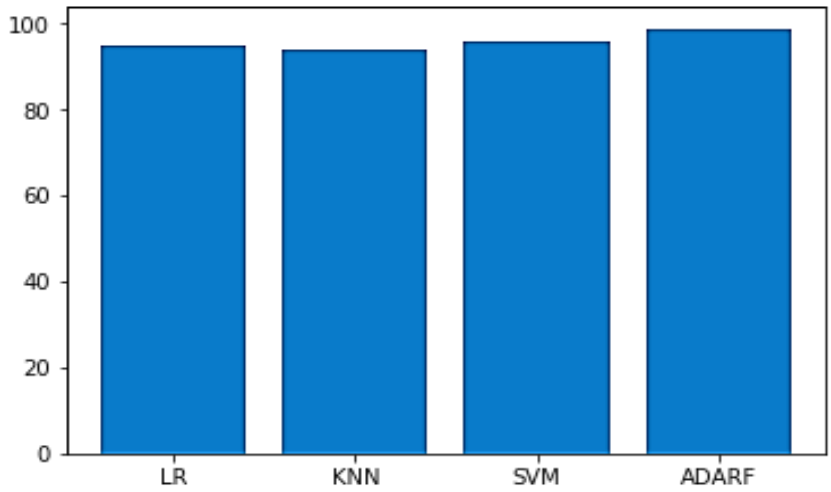

Figure 2: *Accuracy comparison of classifiers.*

Table 3: *Performance metrics of classifiers.*

	Accuracy	Accuracy	Specificity
Logistic Regression	94.7%	94.6%	94.9%
K-Nearest Neighbour	94.1%	95.5%	96.6%
SVM	95.5%	95.5%	96.6%
ADARF	98.8%	98.2%	98.3%

Comparison values of accuracy, sensitivity and specificity metrics are as shown in Table 3.

Receiving Operating Characteristic (Roc) Curve

Visual tool ROC curve is used for comparing the classifiers. In the plot diagonal line shows random guessing. The curve closer to the diagonal line is less accurate model. Curve is plotted for true positive rate (Sens) and false positive rate (1-Spec) to know the efficiency of the models. Classifications are shown in Fig. 3.

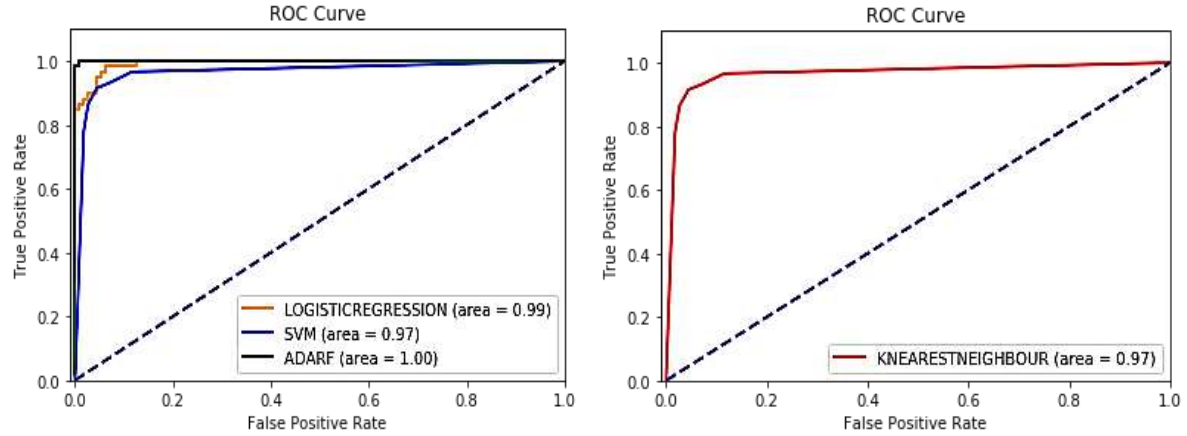

Figure 3: *ROC curve of classifiers.*

DISCUSSION

The result in Table 3, Fig. 2 and 3 shows that the proposed methodology, combination of Random Forest and AdaBoost ADARF has performed better in terms of Acc., Sens. and Spec. comparing to the existing classifiers LR, KNN and SVM. This indicates that proposed approach has a higher probability of correctly differentiating between malignant and benign tumour.

CONCLUSION

In this study classification techniques LR, KNN, SVM and ADARF are used for diagnosing the outcome of breast cancer. Comparative study shows that the proposed methodology ADARF has higher accuracy of 98.8% compared to LR, KNN and SVM classifiers.

The future work involves the study of the expansion of the number of classifiers that can be used in the ensemble and to improve its efficiency in terms of accuracy.

REFERENCES

1. Cancer Research UK, [Online] Available from: https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important.
2. Pei-Chann Chang, Chen-Hao Liu, Chin-Yuan Fan, Jun-Lin Lin, Chih-Ming Lai (16–19 September, 2009), "An ensemble of neural networks for stock trading decision making", *International Conference on Intelligent Computing, Ulsan,* South Korea.
3. M. Lichman, UCI Machine Learning Repository (2013), [Online] Available

from: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin.

4. Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, Md. Kamrul Hasan (21–23 December, 2017), "Prediction of breast cancer using support vector machine and K-nearest neighbours", *IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka,* Bangladesh.
5. Reem Alyami, Jinnan Alhajjaj, Batool Alnajrani, Ilham Elaalami, Abdullah Alqahtani, Nahier Aldhafferi, Sunday O. Olatunji (21–23 February, 2017), "Investigating the effect of Correlation based feature selection on breast cancer diagnosis using artificial neural network and support vector machines", *International Conference on Informatics, Health & Technology (ICIHT), Riyadh,* Saudi Arabia.
6. Moh'd Rasoul Al-hadidi, Abdulsalam Alarabeyyat, Mohannad Alhanahnah (31st August–2nd September, 2016), "Breast cancer detection using K-nearest neighbour machine learning algorithm", *International Conference on Developments in eSystems Engineering (DeSE), Liverpool,* UK.
7. U. Karthik Kumar, M.B. Sai Nikhil, K. Sumangali (2–4 Aug. 2017), "Prediction of breast cancer using voting classifier technique", *IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Chennai,* India.
8. Haifeng Wang, Bichen Zheng, Sang Won Yoona, Hoo Sang Ko (2018), "A support vector machine-based ensemble algorithm for breast cancer diagnosis", *Euro. Jo. of Opl. Res.*, Volume 267, Issue 2, pp. 687–699, DOI: 10.1016/j.ejor.2017.12.001.
9. Lavanya Doddipalli, K. Usha Rani (2012), "Ensemble decision tree classifier for breast cancer data", *Int. Jo. of Inf. Tech. Con. & Ser.*, Volume 2, Issue 1, pp. 17–24, DOI: 10.5121/ijitcs.2012.2103.
10. Tan A. C, Gilbert D (2003), "Ensemble machine learning on gene expression data for cancer classification", *App. Bioinf.*, Volume 2, Issue 3, pp. 75–83.
11. J. Han, M. Kamber (2006), "Data mining: concepts and techniques", *Elsevier Science,* 2nd edition, San Francisco; Morgan Kaufmann, pp. 1–703.
12. P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, A. Zanasi (1998), "Discovering data mining from concept to implementation", *Upper Saddle River,* N. J. Prentice Hall.