International Journal of Engineering & Technology, 7 (4.15) (2018) 281-285

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Ensemble Comparative Study for Diagnosis of Breast Cancer Datasets

1Bibhuprasad Sahu, 2 Sujata Dash, 3Sachi Nandan Mohanty, 4 Saroj Kumar Rout

[1]Research Scholar , Department of CS&IT,NOU,Baripada,Odisha,India.[2]Department of Master in Computer Application NOU, Baripada, Odisha, India.,[3],[4]Department of Computer Science & Engineering, Gandhi Institute For Technology, Bhubaneswar, Odisha, India.

**Corresponding Author E-mail: prasadnikhil176@gmail.com , sujata238dash@gmail.com , dr.sachinandan@gift.edu.in , rout_sarojkumar@yahoo.co.in*

Abstract

Every disease is curable if a little amount of human effort is applied for early diagnosis. The death rate in world increases day by day as patient fail to detect it before it becomes chronic. Breast cancer is curable if detection is done at early stage before it spread across all part of body. Now-a-days computer aided diagnosis are automated assistance for the doctors to produce accurate prediction about the stage of disease. This study provided CAD system for diagnosis of breast cancer. This method uses Neural Network (NN) as a classifier model and PCA/LDA for dimension reduction method to attain higher classification rate. Multiple layers of neural network are applied to classify the breast cancer data. This system experiment done on Wisconsin breast cancer dataset (WBCD) from UCI repository. The dataset is divided into 2 parts train and test. With the result of accuracy, sensitivity, specificity, precision and recall the performance can be measured. The results obtained are this study is 97% using ANN and PCA-ANN, which is better than other state-of-art methods. As per the result analysis this system outperformed then the existing system.

Keywords: *Classification, Neural Network, features selection, PCA, LDA, NB, RF.*

1. Introduction

Now a day's affected ration of breast cancer in case of women around the world increases drastically. Early detection of cancer may reduce the death rate before it reach up to the chronic level. After lung cancer, the death rate of breast cancer is higher as compared to other diseases in women. As per the report of international agency for research on cancer (IARC), the no. of persons affected with cancer is 14.1 million, from this 8.2 million death occurred due to breast cancer [1-3]. The no. increases as doctors are failed to state the stage of the patient. For breast cancer detection, digital mammography a diagnosis model used throughout world. CAD system plays vital role for health professionals to analyze and identify the accurate stage of the patient. So that proper attention may provide to save the life. The aim of this system to develop a computer aided diagnosis model to diagnose a disease in earlier stage by using PCA-NN and LDA-NN. A neural network is a complex network consists of n hidden layers. Each layer inputted from previous layer so the error rate reduced layer by layer and produces accurate result. It is used to designed a complex hierarchy is a simple manner and it support all sort of algorithms such as supervised, unsupervised, semi supervised and reinforcement.

This proposed work of this system is to develop a computer aided diagnosis model to diagnose a disease in earlier stage by using PCA-NN and LDA-NN.

The remaining part of the paper is organized as: Section 2 discusses the related work of different classification model to predict the cancer data, section 3 represents the proposed neural network as a classification process in the system. Section 4 shows Simulation and results work of the system. Concluding remarks is providing in section 5

2. Related Work

A neural network is a branch of AI known as artificial neural network (ANN). A neural network has its own input and output channels called dendrites and axons respectively. It is a programmatic model, which can find the pattern exists is data which replicate the knowledge. ANN has million processing nodes called neuron. In ANN each neuron is called as unit. A neural network consists of layer where as each layer consists of n no. of nodes. A neural network system has a single input layer and has single (or) two hidden layers and produce output. The main objective of neural network is to produce the inputs to a signification output. The Figure 1 shows the structure of ANN

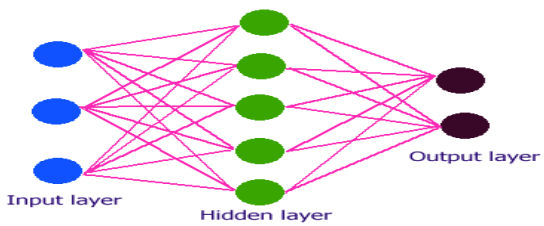

Figure 1: Neural network

Mert et al. [4] used radial basis function neural network (RBFNN) for classification of data and independent component analysis (ICA)

for feature selection. Their approach finds out one feature out of 30 features. This method accuracy obtained of 86%. Bhattacharyya et al. [5] used BPNN (Back Propagation Neural Network) and found accuracy of 99.37% of classification. Goruneseu et al. used evaluation strategy to develop intelligent medical decision model. They have used different classifiers such as NN, GA, SVM, KNN, MLP, RBF, PNN, SOM and NB. They prove SVM perform good with WBC dataset. Ahamad et al. applied there classification algorithm like RBF, MLP and PNN with WBC dataset. Result analysis proves that PNN performed better than other classification algorithm with accuracy of 97.66% [6]. Jhajharia et al. implemented model with PCA and feed forward neural network with training and test data [7]. Yin et al. used SVM with recursive feature elimination (REF) with Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The out of PCA classified with SVM and RE, it achieved 98.58% accuracy [8]. Huang et al. proposed a ensemble learning technique (bagging and boosting). Here GA is used to identify the best feature from dataset. The dataset is divided into 90-10% spits based k fold cross validation GA+RBF SVM achieve accuracy of 98% [9]. Jouni et al. used a model bound on artificial neural network with MLP and BPNN. This model identifies whether it is malignant (or) benign [10]. Bewal et al. used MLP with four different training algorithms such as quasi-Newton, gradient discussed with momentum and adaptive learning, Levenberg-Marquardt and resilient back propagation. Result analysis proves Levenberg-Marquardt + MLP achieve good accuracy rate of 94.11% with comparison to other [10]. Paulin et al. used a system with back propagation neural network (BPNN) and achieved accuracy of 99.28% with Levenberg-Marquardt algorithm. They have used median filter and min-max technique for pre-processing and normalization of data respectively [12]. Menaka et al. used SVM with genetic algorithm on breast cancer dataset and achieved a classification accuracy of 99.78% on begging and 97.67% on Malignant [13]. Hussain et al. proposed a model with genetic algorithm and fuzzy logic with breast cancer data with accuracy of 97.95% [14]. Chandra P et al .proposed a method using
Artificial Neural Network and extreme learning technique on breast cancer data and achieved accuracy of 98% [15]. Jalil et al. proposed a machine learning model using fuzzy feature, Bee colony and Neural Network on WBC data and successfully achieve an accuracy of 99.15% [16]. Asieh K et al. achieved with accuracy of 99% by implementing probabilistic neural network as classifier on breast cancer dataset [17]. Mohammed et al. presented a hybrid Genetic + K means as classifier for feature selection to classify medical datasets. The result analysis proves that the designed model achieved an accuracy of 98% [18]. Kalpana K et al. used neural network as a classifier with breast cancer dataset and the proposed model found to be a good one with accuracy of 97.6% [19].

3. Proposed Method

In this study Neural Network is used for classification process and RFE system to detect a subset of features from the dataset. The steps of the proposed method are described in Figure 2.
To reduce noise from the instance of WBC dataset pre-processed is used.
Recursive feature selection is used for iteration process.
PCA/LDA applied to the training and test dataset.
Neural Network classification technique is applied to the designed model.

3.1 Pre-Processing

In machine learning, pre-processing is used to normalize, ambiguous data from the dataset. In this study, we have the breast cancer dataset, it consist of 569 instances with 32 feature variables. This dataset support binary classification because it has two classes called benign and malignant.

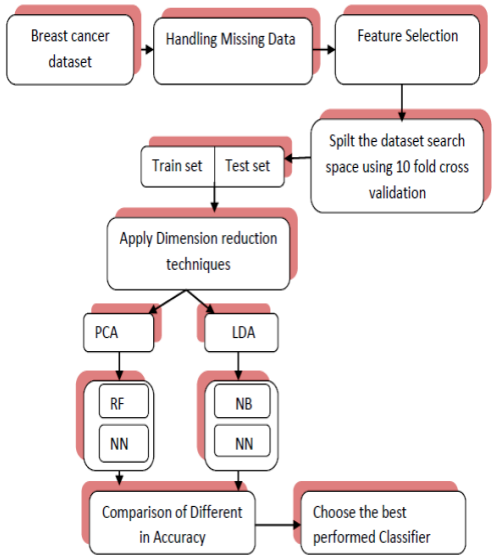

Figure 2: proposed Neural Network classification method

The class benign represents patient not having cancer. Malignant represents patients with cancerous tumors. In this dataset out of 569 instances there are 0 missing values. To improve the stability of the system we remove all missing value instances from the dataset, if any exist. Now the system predicts the best feature from the feature variables to enhance performance as well as the accuracy.

3.2 Feature Selection

Feature selection plays an inevitable role is machine learning to make the dataset free from ambiguity and reduces complexity of data. So the sine of data minimized which solve the problem like over filtering [25]. Predicting the suitable best feature improves the accuracy. Filter, wrapper and embedded feature selection approaches are used in machine learning. Many researches also using hybrid feature selection approaches for better accuracy of dataset.

3.3 Principal Component Analysis (PCA)

The main idea behind principal component analysis (PCA) is to minimize the dimension of dataset. As in the dataset the variables are correlated with each other PCA normally select the variables that are high correlated with each other. So PCA is a dimension reduction to which is used to reduce large no of variables to small set which contain the same information like the large set. In this study PCA is applied to both training and test attributes of breast cancer dataset. PCA identifies pattern from the dataset and estimates similarly and difference between individual attributes. The converted variables are called principal component variables. The output of PCA contains PCA components with decreasing variance values [20]. The covariance matrix is generated using value of Eigen vector and Eigen values [21]. Eigen values are sorted ascending order to identify most significant data and less significant data discarded to reduce dimension [22].
The variance is calculated using equation (1) to find the spread of data in breast cancer dataset [23].

$$\mathrm{Var}(x) = \sigma^2 \quad (1)$$

To find relation between two classes covariance is calculated. Covariance calculates for which zero values, there is no relation

between two dimensions covariance can be compiled using equation (2) as suggested in [24].

$$\text{Cov(x,y)}= \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\mu_{xj})(y_{ij}-\mu_{yj}) \quad (2)$$

Then Eigen values and Eigen vector for covariance matrix is calculated, then it is transformed into orthogonal rotation using equation (3)

$$\text{Det}(A-\lambda I)=0 \quad (3)$$

3.4 Linear Discriminate Analysis (LDA)

LDA is an attribute selection method to categorize the dataset with two (or) more classes [24]. LDA is a global method of fisher's linear discernment which selects linear combination of features and classification done with linear classifier. LDA is also used to regression analysis. LDA measurement of individual class is equally distributed. In LDA the maximum separability can be achieved by maximizing the ration of class variance. After splitting the original dataset into training set and test set, mean of each dataset and mean of entire dataset is found by merging the test and training set[25,26].

$$\mu_3 = (\mu_1 * p_1) + (\mu_2 * p_2) \quad (4)$$

P_1& P_2 are probability features, then inter, intra class scatter matrices is generated.

Intra class scatter is generated by considering covariance of each class. Scatter matrices are computed using equation (5)

$$s_m = \sum_j p_{j*Cov(j)} \quad (5)$$

Covariance matrix is computed as equation (6)

$$Cov_j=(x_{j-}\mu_j)(x_i-\mu_i)\wedge T \quad (6)$$

So interclass scatter matrix can be

Covariance matrix= (μ_i- mean of entire dataset)×(μ_i-mean of entire dataset) ∧T (7)

Then Eigen values and Eigen vectors matrix is generated. Eigen Vector we locate the values of that satisfy equation (8)

$$A-\lambda I \quad (8)$$

Sorting the Eigen vectors by arranging the Eigen values is a decreasing order. After transformation is over, Euclidian distance (or) root means square (RMS) distance is used to classify the data.

3.5 Classification

After the dimension reduction is such as principal component analysis (PCA) and linear discriminate analysis (LDA) is applied with the original dataset, classifiers such as Neural Network and Random Forest is applied to find out the accuracy of the given dataset.

3.5.1 Random Forest

Random Forest classifier is a ensemble algorithm, means Random Forest the combines with other one algorithm for classify the object. Random Forest classifier generates a set of decision trees from randomly selected subset of training. It then aggregates the votes from different decision trees to decide the final class of the test object. Random Forest has same hyper parameter as a decision tree (or) bagging classifier. No need to combine a decision tree with a bagging classifies, one can easily use the classifier class of random forest. In Random Forest instead of searching for biomarker white splitting a node, it searches it among a random subset of features.

4. Simulation and Results

In this study, WBC dataset consists of 569 instars with 32 features variables. This dataset supports for binary classification model as is has two class labels such as 0 for benign (Non cancer) 1 for malignant (cancerous). There are no instances from 569 instance found to be missing value. If any missing value exists then instances are removed from the dataset to reduce the error rate of the system. So, finally we considered instances for feature selection. After feature selection the dataset is applied to PCA/LDA for dimension reduction to get better accuracy. The featured dataset are spilt into training and test data and inputted for dimension reduction.
The performance of model is evaluated through confusion matrix. By the confusion matrix we can find the classified and misclassified rate of the system. Accuracy of the model represents effectiveness and performance of a system.

Accuracy can be defined as, Accuracy= (TP+TN)/TP+TN+FP+FN) (9)

Where TP= True Positives
TN=True Negatives
FP= False Positives
FN=False Negatives

Sensitivity of a model is defined as the ration between correctly identified instances versus actual positives.

$$\text{Sensitivity= TP/ (TP+FN)} \quad (10)$$

Specificity of a model is a ratio of correctly identified instances with actual negative. The accuracy of a model can be found using F-Score. To compute F-Score precision and recall also calculated,

$$\text{Specificity= TP/(TP+FN)} \quad (11)$$

$$\text{Precision =TP/ (TP+FP)} \quad (12)$$

$$\text{Recall=TP/(TP+FN)} \quad (13)$$

F- Score= $((+1)* \beta^2\text{Precision}* \text{Recall})^{/} (\beta^2*\text{Precision} +\text{Recall})$
Where B<1 then F-Score value favor for precision when B>1 it favor for recall.
In this study, we trained the network model with training set of 70-30%. We then check the performance of model with various percentage of splitting such as 80-20%, 60-40%. As expected, this model performed and provides a promising result of 97% for 70-30%, 96% for 80-20%, 94% for 60-40%. The accuracy of the designed model is presented in Table-3.

Table 2: Confusion matrix for proposed system

Classifier	Prediction		Reference
		B	M
RF	B	105	5
	M	2	58
PCA+RF	B	104	5
	M	3	58

ANN	B	107	5
	M	0	58
PCA+ANN	B	105	3
	M	2	60
NB	B	99	7
	M	8	56
LDA+NB	B	99	7
	M	8	56

Table 3: Performance for the proposed system (70%-30%)

Classifier	Kappa statistics	Accuracy	Sensitivity	Specificity	Prevalence
RF	0.91	0.95	0.92	0.98	0.37
PCA+RF	0.95	0.95	0.92	0.97	0.37
KNN	0.93	0.97	0.92	1	0.37
ANN	0.93	0.97	0.92	1	0.37
PCA+ANN	0.93	0.97	0.95	0.98	0.37
NB	0.81	0.91	0.88	0.92	0.37

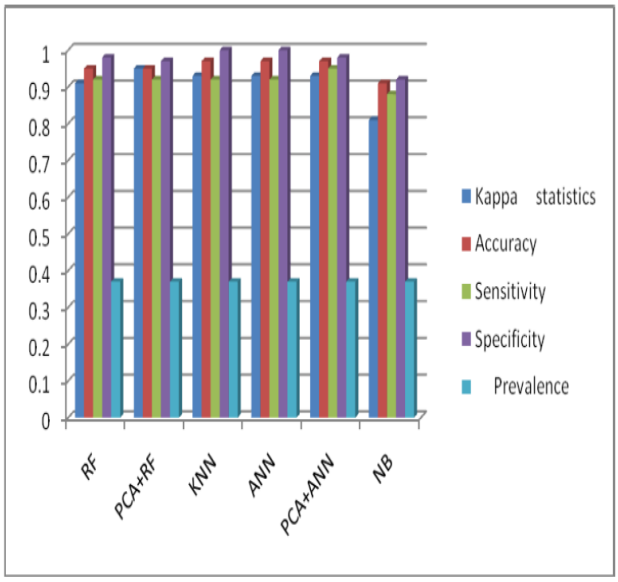

Fig 2: ROC performance of the system using Scatter Plot

This system achieved outperformed accuracy of 97 for ANN and PCA-ANN is shown in Figure 2. Partitioning confusion matrix table presents the performance in term of accuracy, sensitively, and specifying with various test-training spilt partition Figure 3 show the ROC performance of the system using Scatter Plot.

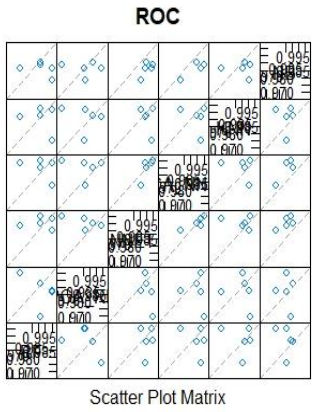

Figure 3: ROC performance of the system using Scatter Plot

5. Conclusion

In various sciences, including medicine Classification models based on artificial intelligence have had a significant impact on the predictive decision making process. From the result analysis we found that the ANN and PCA-ANN provide a result with accuracy of 97%.We have compared the result with various existing approach by different researchers. For cancer diagnosis ANN plays a major impact for early detection to save the life of human being.

References

[1] Prevention Control: Center for Diseases Control and Prevention(2014).URL https://www.cdc.gov/cancer/breast/index.htm.
[2] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control.
[3] Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA: a cancer journal for clinicians. 2015 Mar;65(2):87-108.
[4] Mert A, Kılıç N, Bilgili E, Akan A. Breast cancer detection with reduced feature set. Computational and mathematical methods in medicine. 2015;2015.
[5] Bhattacherjee A, Roy S, Paul S, Roy P, Kausar N, Dey N. Classification approach for breast cancer detection using back propagation neural network: a study. InBiomedical image analysis and mining techniques for improved health outcomes 2016 (pp. 210-221). IGI Global.
[6] Karaa WB, editor. Biomedical image analysis and mining techniques for improved health outcomes. IGI Global; 2015 Nov 3.
[7] Azar AT, El-Said SA. Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Computing and Applications. 2014 Apr 1;24(5):1163-77.
[8] Jhajharia S, Varshney HK, Verma S, Kumar R. A neural network based breast cancer prognosis model with PCA processed features. InAdvances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on 2016 Sep 21 (pp. 1896-1901). IEEE.
[9] Yin Z, Fei Z, Yang C, Chen A. A novel SVM-RFE based biomedical data processing approach: Basic and beyond. InIndustrial Electronics Society, IECON 2016-42nd Annual Conference of the IEEE 2016 Oct 23 (pp. 7143-7148). IEEE..
[10] Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM ensembles in breast cancer prediction. PloS one. 2017 Jan 6;12(1):e0161501.
[11] Jouni H, Issa M, Harb A, Jacquemod G, Leduc Y. Neural Network architecture for breast cancer detection and classification. InMultidisciplinary Conference on Engineering Technology (IMCET), IEEE International 2016 Nov 2 (pp. 37-41). IEEE
[12] Nachaliel E, Lenington S, inventors; Mirabel Medical Ltd, assignee. Breast cancer detection. United States patent US 7,409,243. 2008 Aug 5.
[13] Paulin F, Santhakumaran A. Classification of breast cancer by comparing back propagation training algorithms. International Journal on Computer Science and Engineering. 2011 Jan;3(1):327-32.

[14] Menaka K, Karpagavalli S. Breast Cancer Classification using Support Vector Machine and Genetic Programming. International Journal of Innovative Research in Computer and Communication Engineering. 2013 Sep;1(7)
[15] Lafta HA, Ayoob NK. Breast Cancer Diagnosis Using Genetic Fuzzy Rule Based System. Journal of University of Babylon. 2013;21(4):1109-20.
[16] Utomo CP, Kardiana A, Yuliwulandari R. Breast cancer diagnosis using artificial neural networks with extreme learning techniques. International Journal of Advanced Research in Artificial Intelligence. 2014 Jul;3(7):10-4.
[17] JalilAddehb ,MassoudPourmandia," Breast Cancer Diagnosis Using Fuzzy Feature and Optimized Neural Network via the Gbest-Guided Artificial Bee Colony Algorithm", Computational Research Progress in Applied Science & Engineering, Vol.1, No. 4, 152-159, 2015.
[18] Pourmandi M, Addeh J. Breast cancer diagnosis using fuzzy feature and optimized neural network via the Gbest-guided artificial bee colony algorithm. Computational Research Progress in Applied Science & Engineering. 2015;1(4):152-9.
[19] Naser MA, Hasan ZF, Hussein EA. A hybrid Genetic K-Means Algorithm forFeatures Selection to Classify Medical Datasets. journal of kerbala university. 2016(لكلية الرابع العلمي المؤتمر العلوم):139-49.
[20] Kalpana K and Anil Arora A, Breast Cancer Diagnosis using Artificial Neural Network, (IJLTET),7(2), 2016.
[21] Bro R, Smilde AK. Principal component analysis. Analytical Methods. 2014;6(9):2812-31.
[22] Kotu V and Deshpande B 2015 Predictive Analytics and Data Mining (Waltham: Morgan Kaufmann)
[23] Kavitha R and Kannan E 2016 An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining IEEE Int. Conf. on Emerging Trends in Engineering Technology and Science (ICETETS) pp 1-5
[24] Jolliffe I T 2002 Principal Component Analysis 2 nd Ed. (New York: Springer-Verlag)
[25] Johnson RA and Wichern DW 2007 Applied Multivariate Statistical Analysis 6 th Ed. (New Jersey: Pearson Prentice Hall)
[26] Sahu B. A Combo Feature Selection Method(Filter +Wrapper) for Microarray Gene Classification, International Journal of Pure and Applied Mathematics Volume 118 No. 16 2018