An ensemble classification based approach for breast cancer prediction

Vijayalakshmi G V Mahesh¹, Mohan Kumar M²

¹Department of Electronics and Communication Engineering, BMS Institute of Technology and management, Bangalore, India
²Department of Electronics and Communication Engineering, Yenepoya Institute of Technology, Mangalore, India

E-mail: vijayalakshmi@bmsit.in, mohankumar@yit.edu.in

Abstract. Breast cancer is the second major reason for deaths in women. Early detection of the breast cancer and receiving the appropriate treatment can reduce the death rates as survival becomes hard in the higher stages of the tumor growth. Application of machine learning in healthcare play a key role in aiding the clinical experts to detect the disease at an early stage and perform precise assessment. This paper proposes a pattern recognition methodology that uses breast cancer biomarkers as the attributes and ensemble classification approach for accurately detecting the presence of cancer. The proposed method was evaluated for the samples of the breast cancer Coimbra dataset by fusing the decisions of naïve Baye's, radial basis function neural network and linear discriminant analysis classifiers based on majority voting rule. The experimental results demonstrated the enhanced performance of the system with fusion of classifiers as compared to the single classifiers.

1. Introduction

An uncontrolled way of cell division leads to cancer and these cells often invade into other healthy tissues. Breast cancer is the most common and is the leading cause of death in women. It can be invasive or noninvasive: invasive cancer spreads from the ducts or glands to other parts of the breast while noninvasive cancer does not spread from the original tissue. Based on this, the following types of breast cancer are identified (i) Ductal carcinoma in situ (DCIS): It is non invasive where the cancer cells are confined only to the ducts of the breast, (ii) Lobular carcinoma in situ (LCIS): A non invasive condition where the cancer cells are limited to the milk producing glands of the tissue, (iii) Invasive ductal carcinoma (IDC): This is invasive type of cancer. The cancer cells develop in the milk ducts of the breast and then invade to the nearby tissues and organs. IDC is the most common type of breast cancer and (iv) Invasive lobular carcinoma (ILC): This type of cancer begins in the lobules of the breast and spreads to the surrounding tissue.

Based on the size of the tumor and the extent of its spread into nearby tissues and organs the doctors recognize the different stages of cancer as stage 1, stage 2, stage 3 and stage 4 breast cancer. Stage 1 is a primary tumor with its size around 2cms while stage 4 is advanced where the tumor can have any size and the cancer cells have spread to nearby tissues and organs. Early detection of the cancer i.e.,
in the initial stages improves possibility of survival. If the cancer reaches stage 4, the chances of survival comes down. Thus early detection of cancer is very crucial and regular checks and screening can aid in detecting the symptoms earlier. Clinical experts diagnose the cancer after routine screening. The use of biomarkers and imaging tests helps the clinicians to identify the breast cancer. The most widely used image screening methods are Mamogram, ultrasound and Magnetic resonance imaging (MRI). Further, the researchers are trying to find new breast imaging methods to improve the detection and treatment of the cancer and have come up with Scintimammography (molecular breast imaging), Positron emission mammography (PEM), Electrical impedance imaging (EIT) and Elastography.

Though imaging methods play an important role in decision making followed by treatment. They require expertise to deliver significant and accurate information. For example, extraction of feature descriptors from images using mathematical techniques to detect and differentiate the tumor and non-tumorous region. Also the reliability of the detection method needs to be assessed based on (i) Accuracy, (ii) Precision, (iii) Correlation to the disease and (iv) Usability. Biomarkers[1], on the other hand are gaining lot of importance in accurately detecting and diagnosing the cancer as a result of advancements in bioinformatics tools. Recognition of significant biochemical changes in cancer has facilitated towards advances in detection and its treatment. They can be relied upon for early detection and accurate evaluation. Breast and prostate cancers are diagnosed comparatively earlier as compared to other cancers. Additionally, the cancer biomarkers are significant in detecting the disease for asymptomatic persons at the early stage. Ideally a portion of a protein or a protein in the blood or urine samples is considered as a marker for tumor detection. Thus this paper proposes a pattern recognition system that use biomarkers as features or descriptors combined with machine learning algorithms to effectively identify the presence or absence of cancer. Accordingly, this is a classification task where the classifier predicts the class label of the sample. The performance of the model is later quantitatively evaluated using the metrics derived from the confusion matrix.

The rest of the paper is organized as follows, Section 2 presents a comprehension of related work. Section 3 describes the methodology of the works. The experimental results and discussion are provided in section 4. Finally, section 5 concludes the paper.

2. Related work

Biomarkers in cancer are the characteristics that are measured and they describe the pathological state of an living being. Cancer biomarkers are categorized into predictive, diagnostic and prognostic based on how they are handled. The first cancer biomarker was recognized from the blood sample of a patient with colon cancer [2]. Later different biomarkers were developed to identify other cancers. The standards and measurements of breast cancer biomarkers are provided in [3] and a overview is found in[2]. Further, the advancements in technology and data analytics is seen to solve various health related issues to improve the quality of health care and provide effective treatment. Recent years have witnessed a biggest transformation of Artificial intelligence in health care domain powered by machine learning algorithms that can assist the clinicians/experts in diagnosing the disease and build up effectual solutions. Thus the acquired dataset is fed to the algorithms that can discover the patterns and anomalies to make right decisions about the disease diagnosis.

Many researchers have put forward their ideas in using cancer biomarkers with machine learning algorithms to improve the precision of detection and analysis. Li and Chen[4] investigated the adaption of cancer attributes in detecting the disease which can reduce the probability of death rate. They used Wisconsin Breast cancer and Breast Cancer Coimbra datasets with Decision tree (DT), Random forest (RF), Support Vector Machine (SVM), Neural Network (NN) and Logistics Regression (LR) Classifiers and compared their performances. The experimental study showed that the Random forest algorithm provided better results in terms of prediction accuracy, F-score and Area Under the Curve (AUC) as compared to other classifiers. Farah Sardouk et al.[5] studied the predictors of Breast Cancer Coimbra dataset to estimate the presence or absence of the cancer. They created six machine models using Adaboost, Regression, Random Forest, Jrip, RBFNN and J48 classifiers in WEKA and MATLAB environments. The models were assessed using True positive rate, False positive rate, precision, recall, F1-score MCC and AUC. The results proved that the machine
learning algorithms were efficient in predicting the outcome. Miguel Patricio et al., [6] at university of Coimbra collected clinical attributes from 166 individuals based on anthropometric data. The measured data or features were then fed to logistic regression, Support vector machine and random forest machine learning algorithms to evaluate the probability of using the anthropometric data as biomarkers for cancer prediction. which were given to the developed a model that can be used as a biomarker. The evaluation of the classifier models for the metrics sensitivity, AUC and specificity indicated the better performance of SVM for the following predictors: Glucose, Resistin, BMI and age. Thus they can be efficient biomarkers of breast cancer.

Jin Yue et al.,[7] carried out statistical analysis, spectral analysis and least squares fitting analysis on the Breast Cancer Coimbra dataset to check the interclass variation. The analysis was able to clearly indicate the significant distinction between the two classes: cancerous and non cancerous. Yolanda D. Austria et al.,[8] investigated 11 classification algorithms and their variants : LR, k Nearest Neighbor(kNN), SVM, DT, RF, Gradient boosting and naive Baye's for predicting the breast cancer on the Coimbra dataset. The performance of the models were evaluated using accuracy and prediction time. The evaluation displayed an higher accuracy of 74.14% with Gradient boosting classifier and non linear SVM was fast in predicting the label without any delay. The work also identified the anthropometric data body mass index(BMI) as the principal predictor. Ratula Ray et al.,[9] explored the use of DT, kNN, RF and Guassian naive Baye's machine learning algorithms in aiding the pathologists to make precise decision about the presence of the disease considering both image and numeric(biomarkers) databases. Authors used breast histopathology images from Kaggle and breast Coimbra dataset for the evaluation of classifier models. The analysis indicated the better performance of random forest classifier which provided precision, recall and F1 score of 71%, 71% and 70% respectively for the numeric dataset. Similarly it presented precision, recall and F1 score of 91%, 92% and 92% respectively. The study also implies the importance of resistin and BMI biomarkers that influence the decision process.

From the evaluation of the contributions it is found that majority of the methods used various machine learning algorithms independently to improve the success rate of cancer detection. This paper presents a model that makes use of a ensemble classification where the classifiers are fused at the decision level to enhance the performance of the system.

3. Methodology
The methodology of the proposed system to predict the presence of breast cancer is shown in Figure1. The frame work includes (i) Feature extraction (ii) Classifier/Machine learning model creation (iii) Classification.

The extraction of suitable features or descriptors are essential in deciding the accuracy of the system. The anthropometric data measured from the input samples or the biomarkers form the features. These features are given to different classifiers for models creation which can predict the class label of the given sample. To enhance the presentation of the system, the decisions of the individual classifiers are combined based on decision level fusion strategy. The created models are evaluated to validate their performance. The details of the process is described in subsequent sub sections.
3.1 Dataset and Feature extraction

The work considered Breast Cancer Coimbra dataset[6] collected from 116 persons. From each participant nine clinical features were measured as displayed in Table 1.

| Feature | Description          | Unit  |
|---------|----------------------|-------|
| A       | Age                  | Years |
| BMI     | Body Mass Index      | Kg/m2 |
| G       | Glucose              | Mg/dL |
| I       | Insulin              | µU/mL |
| HOMA    | Homeostasis Model    | --    |
|         | Assessment index     |       |
| L       | Leptin               | Ng/mL |
| Ad      | Adiponectin          | µg/mL |
| R       | Resistin             | Ng/mL |

These clinical measures are the anthropometric data collected during procedural blood analysis. Out of 116 input samples, 64 samples are with breast cancer and the remaining 52 are healthy samples. An observation of the features of dataset as illustrated in Figure 2 identifies the distinction between the cancerous and non-cancerous samples that clearly indicate that they are potential biomarkers for the detection of the pathology.
Figure 2. Features of Healthy and Cancerous samples

The features are combined to form a feature vector as shown,

\[ FV = [A, BMI, G, I, HOMA, L, Ad, R, MCP-1] \]  \hspace{1cm} (1)
which is provided to the classifiers for training and testing. To validate the proposed system, the dataset with its features was divided into two non-overlapping training and testing sets with a ratio of 75:25. The training set is labeled using the class labels \( y = \{1\, \text{'1'},\, 0\} \), where '1' indicates healthy and '0' indicates Cancerous. The labeled training set \( \{FV, y\} \) is passed to the classifier for creating the model. The model created is then tested using the samples from the testing set to predict its outcome \( C \in y \).

3.2 Machine Learning Algorithms

The classifier predicts the outcome of an unseen sample \('Q'\) as a function of labeled training set and the learning parameters of the algorithm.

\[
C = \text{f}(FV_0, \{[FV, y] : \text{training parameters}\})
\]

where \( FV \) is the feature vector of the training samples and \( FV_0 \) is the feature vector of the query/unseen sample.

The proposed system used three classifiers: Naive Baye's(NB), Radial Basis Function Neural network(RBFNN) and Linear Discriminant Analysis(LDA).

3.2.1 Naive Baye's classifier

It is a simple probabilistic classifier which is based on Baye's theorem[10]. It requires fewer number of training samples for prediction. Baye's rule can be formulated as

\[
P(y/X) = \frac{P(X/y).P(y)}{P(X)} \quad (2)
\]

where \( X \) signify the features and \( y \) is the label representing the category of the input sample. In the context of the work presented here, the above equation can be rewritten by expanding the feature set as,

\[
P(C/f_1,f_2,...........,f_0) = \frac{P(f_1/C).P(f_2/C).P(f_3/C)............P(f_9/C).P(C)}{P(f_1).P(f_2)............P(f_0)} \quad (3)
\]

For all the elements of the dataset, the denominator is constant, thus we get

\[
P(C/f_1,f_2,...........,f_0) \propto P(C) \prod_{i=1}^{9} P(f_i/C) \quad (4)
\]

Here \( C \) can be one of the class label belonging to \( y \). Thus the outcome \( C \) is found based on maximum probability, i.e.,

\[
C = \text{argmax}_c P(C) \prod_{i=1}^{9} P(f_i/C) \quad (5)
\]

The classifier finds its applications in biometrics(face recognition), healthcare, weather prediction, security, emotion recognition, sentiment analysis etc.

3.2.2 Radial Basis Function Neural network(RBFNN)

RBFNN[11] is one of the variants of feed forward neural network that uses radial basis functions as the activation function in the hidden layer. The network converges faster to reach the performance goal. The NN architecture has three layers: Input layer, Hidden layer and Output layer. The number of neurons in the input layer correspond to the dimension of the feature vector and the neurons in output layer correspond to the number of class labels. Each layer is associated with its activation function. The activation functions in the hidden layer are implemented as Gaussian functions.

The labeled training set \( \{FV, y\} \) is provided to the network for training. During training the parameters such as the network weights, RBF activation function centres and their distributions are tuned reach the performance goal: Mean squared error(MSE)
For a given input \( FV \), the output at a node of output layer is computed as,

\[
C_i = \sum_{j=0}^{L_2} w_{ij} \varphi_j (\|FV - M_j\|) \tag{6}
\]

Where, \( FV \) is the input to the network/Feature vector
\( w_{ij} \) is the weight from hidden neuron \( j \) to output neuron \( i \)
\( \|.\|_2 \) / Euclidean norm
\( \varphi_j \) is the RBF in the hidden neuron \( j \) with center \( M_j \).

### 3.2.3 Linear Discriminant Analysis (LDA)

LDA classifier proposed by Fisher [12] is based on the concept of reducing the within (intra) class variation and maximizing the between (inter) class variation. This classifier finds class separation by minimizing the spread within class and maximizing the distance between each class. Thus it is based on the statistical measures mean and variance. For the given dataset and class labels, LDA works by projecting it on to hyperplane to find the scalar, of all the possible lines, the line that gives maximum class separability is chosen.

The best projection vector is selected by maximizing the Fisher's criteria given by,

\[
J(W) = \frac{W^T S_W W}{W^T S_B W} \tag{7}
\]

here \( J(W) \) indicates the measure of difference between means of the classes

\( S_W \) represents the within class scatter matrix
\( S_B \) represents between class scatter matrix
\( W \) is the projection matrix.

Maximizing \( J(W) \) yields
which is Fisher's Linear Discriminant

\[
W = \text{argmax}(J(W)) = S_W^{-1}(\mu_1 - \mu_2), \tag{8}
\]

where, \((\mu_1 - \mu_2)\) is the distance between the projected means for two classes.

### 4. Results and Discussion

The experiment was conducted to identify the presence of the breast cancer for the given sample based on fusion of classifiers. The proposed method is validated with hold out method where the dataset with 116 samples was separated into training and testing sets with a ratio of 3:1. Thus the training set has 87 samples and testing set 29 samples.

As a first step, all the training samples were passed to feature extractor that provided the feature vector \( FV \) with 9 dimensions. Each feature of the \( FV \) has different ranges, so the data is normalized using Z-score normalization that gives normalized \( FV \) with zero mean and standard deviation of one as shown

\[
FV_n = \frac{FV - \mu}{\sigma} \tag{9}
\]

Later the \( FV_n \) was labeled to form training data set \( Z = \{FV_n, y\}_{n=1}^{87}\). The training data set was accepted by the Naive Baye's(NB), Radial Basis Function Neural network(RBFNN) and Linear Discriminant Analysis(LDA) classifiers for creating the models. During training, the learnable parameters of the classifiers were modified based on the performance function. The RBFNN was optimized for a spread constant of 3.5 and for NB classifier the prior probabilities was set to [0.2 0.2]. The trained models were tested with the testing data set \( Z_{Q} = \{FV_n\}_{Q=1}^{29} \), where \( Q \) represents the Query or unseen sample.

The results of the individual classifiers were observed and analyzed using the parameters Receiver Operating Characteristics(ROC), Classification accuracy(CA), Precision, Recall(Sensitivity), Specificity and F1 score which are computed using the elements of confusion matrix[13] as shown,
Classification accuracy (CA) = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (10)

Precision = \frac{(TP)}{(TP+FP)} \quad (11)

Recall (Sensitivity) = \frac{(TP)}{(TP+FN)} \quad (12)

Specificity = \frac{(TN+FP)}{(TN+FN)} \quad (13)

F1 score = \frac{2 * (Precision \times Recall)}{(Precision + Recall)} \quad (14)

where

TP - Number of true positives
FN - Number of false negatives
FP - Number of false positives
TN - Number of true negatives

The ROC obtained for all the classifiers is displayed in Figure 3 and other metrics are shown in Table 2.

The analysis of ROC and the performance measures indicate the poor presentation of the classifiers when used individually. The highest CA, precision and specificity was obtained from RBFNN that provided 69%, 70% and 81.25% respectively.
Table 2. Classifier results

| Classifier | NB  | RBFNN | LDA  | Ensemble classification |
|------------|-----|-------|------|-------------------------|
| CA(%)      | 62.06 | 69    | 58.17 | 75.86                   |
| Precision(%) | 56.25 | 70    | 53.84 | 75                      |
| Recall(%)  | 69.23 | 53.84 | 53.84 | 69.23                   |
| Specificity(%) | 56.25 | 81.25 | 62.5  | 81.25                   |
| F1-score   | 0.6206 | 0.6086 | 0.5383 | 0.7199                 |

To improve the performance of the system, the decision of the individual classifiers can be fused [14] which is applied in the proposed framework. A review of the literature finds different fusion strategies based on max, min, sum, product, mean, median and majority vote rules. This work utilizes binary classifiers with class label vector $y = \{1', 0'\}$. Let $C_1$, $C_2$ and $C_3$ be the outputs of NB, RBFNN and LDA classifiers such that $C_i \in y$. The outputs are fused as shown,

$$C = F(C_1, C_2, C_3)$$

(15)

Here $F$ signifies the fusion rule.

Now based on the fusion, we arrive at the final decision by opting for majority voting rule which requires odd number of single classifiers. This is the simple and a sensitive approach that assigns the sample with a most frequent class labels. Finally to demonstrate the significance of ensemble classification, the final decision was analyzed using the performance metrics and the results are tabulated in Table 1. The results show an 75% precision with ensemble classification that represents a low false positive rate. The analysis of F1-score which is the harmonic mean of precision and recall is advantageous in our work as the dataset considered has uneven class distribution (52 healthy and 64 cancerous samples). The range of F1-score is [0 1]. Thus a value of 0.7199 is better as compared to individual classifier results.

It is observed from the results that, the ensemble classification provided the results greater than the average of the single classifiers indicating a enhanced performance of the system subjected to CA, precision, recall, specificity and F1-score.

5. Conclusion

This paper proposes a pattern recognition system with ensemble of classifiers for classifying a sample into two classes viz presence of breast cancer and healthy. The work utilized anthropometric data/biomarkers as the features for data representation. These features or descriptors were provided to the NB, RBFNN and LDA machine learning algorithms for classifier model creation and evaluation. The experimental study on breast cancer Coimbra dataset with the proposed framework revealed the efficacy of the biomarkers as better features. Further the analysis of results obtained from fusing the decisions of single classifiers indicated improved performance measures in terms of classification accuracy, precision, recall, specificity and F1-score. This justified the deployment of ensemble classifiers in better discrimination of classes leading to enhanced performance of the system.

6. References

[1] Kim SH, Hoffmann U, Borggrefe M, Akin I, Behnes M. Advantages and limitations of current biomarker research: from experimental research to clinical application. Current Pharmaceutical Biotechnology. 2017 May 1;18(6):445-55.
[2] Chatterjee, Sabarni K., and Bruce R. Zetter. "Cancer biomarkers: knowing the present and predicting the future." (2005): 37-50.
[3] Cole KD, He HJ, Wang L. Breast cancer biomarker measurements and standards. PROTEOMICS–Clinical Applications. 2013 Jan;7(1-2):17-29.
[4] Li Y, Chen Z. Performance evaluation of machine learning methods for breast cancer prediction. Appl Comput Math. 2018 Oct 18;7(4):212-6.
[5] Sardouk F, Duru AD, Bayat O. Classification of breast cancer using data mining. American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS). 2019 Jan 4;51(1):38-46.

[6] Patrício M, Pereira J, Crisóstomo J, Matafome P, Gomes M, Seiça R, Caramelo F. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC cancer. 2018 Dec 1;18(1):29

[7] Yue J, Zhao N, Liu L. Prediction and Monitoring Method for Breast Cancer: A Case Study for Data from the University Hospital Centre of Coimbra. Cancer Management and Research. 2020;12:1887.

[8] Austria YD, Jay-ar PL, Maria Jr LB, Goh JE, Goh ML, Vicente HN. Comparison of Machine Learning Algorithms in Breast Cancer Prediction using the Coimbra Dataset. cancer. 2019;7:10.

[9] Ray R, Abdullah AA, Mallick DK, Dash SR. Classification of Benign and Malignant Breast Cancer using Supervised Machine Learning Algorithms Based on Image and Numeric Datasets. InJournal of Physics: Conference Series 2019 Nov (Vol. 1372, No. 1, p. 012062). IOP Publishing.

[10] Rish I. An empirical study of the naive Bayes classifier. InIJCAI 2001 workshop on empirical methods in artificial intelligence 2001 Aug 4 (Vol. 3, No. 22, pp. 41-46).

[11] Mahesh VG, Raj AN. Invariant face recognition using Zernike moments combined with feed forward neural network. International Journal of Biometrics. 2015;7(3):286-307.

[12] Bouzalmat A, Kharroubi J, Zarghili A. Comparative study of PCA, ICA, LDA using SVM classifier. Journal of Emerging Technologies in Web Intelligence. 2014 Feb 1;6(1):64-8.

[13] Fawcett T. ROC graphs: Notes and practical considerations for researchers. Machine learning. 2004 Mar 16;31(1):1-38.

[14] Mi A, Wang L, Qi J. A multiple classifier fusion algorithm using weighted decision templates. Scientific Programming. 2016 Jan 1;2016.