Expert Systems With Applications 85 (2017) 134–145

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Optimal breast cancer classification using Gauss–Newton representation based algorithm

Lingraj Dora[a], Sanjay Agrawal[b], Rutuparna Panda[b,*], Ajith Abraham[c]

[a] Department of Electrical and Electronics Engineering, VSSUT, Burla 768018, India
[b] Department of Electronics and Telecommunication Engineering, VSSUT, Burla 768018, India
[c] Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, Washington-98071-2259, USA

ARTICLE INFO

*Article history:*
Received 27 December 2016
Revised 26 April 2017
Accepted 15 May 2017
Available online 15 May 2017

*Keywords:*
Breast cancer classification
Sparse representation
Gauss-Newton representation based algorithm
Euclidean distance measure

ABSTRACT

Breast cancer is a decisive disease worldwide. It is one of the most widely spread cancer among women. As per the survey, one out of eight women in the world are at risk of breast cancer at some point of time in her life. One of the methods to reduce breast cancer mortality rate is timely detection and effective treatment. That is why, more accurate classification of a breast cancer tumor has become a challenging problem in the medical field. Many classification techniques are proposed in the literature. Today, expert systems and machine learning techniques are being extensively used in the breast cancer classification problem. They provide high classification accuracy and effective diagnostic capabilities. In this paper, we have proposed a novel Gauss-Newton representation based algorithm (GNRBA) for breast cancer classification. It uses the sparse representation with training sample selection. Until now, sparse representation has been successfully applied in pattern recognition only. The proposed method introduces a novel Gauss-Newton based approach to find the optimal weights for the training samples for classification. In addition, it evaluates the sparsity in a computationally efficient way as compared to the conventional $l_1$-norm method. The effectiveness of the GNRBA is examined on the Wisconsin Breast Cancer Database (WBCD) and the Wisconsin Diagnosis Breast Cancer (WDBC) database from the UCI Machine Learning repository. Various performance measures like classification accuracy, sensitivity, specificity, confusion matrices, a statistical test and the area under the receiver operating characteristic (AUC) are reported to show the superiority of the proposed method as compared to classical models. The experimental results show that the proposed GNRBA could be a good alternative for breast cancer classification for clinical experts.



## 1. Introduction

According to American Cancer Society's report, 2017 (https://www.cancer.org), breast cancer is the second major cause of death among women. There is a possibility that 1 in 37 (about 2.7%) woman may die from breast cancer. The chance of developing breast cancer in women increases significantly, as they get older (McPherson, Steel, & Dixon, 2000). Yet there is no established theory regarding the causes and treatment of the disease (Christoyianni, Dermatas, & Kokkinakis, 2000; Rodrigues, Chang, & Suri, 2006). Breast cancer arises due to uncontrolled growth of cells. A normal cell in its life cycle grows in size, divide into new cells and die at the proper time. However, cancerous cells behave differently from normal cells. Any changes or mutation in DNA can affect normal cells to become cancerous. Some genes control the behavior of normal cells, such as cell growth, division into new cells and repair or perish at the proper time. One such gene is called Proto-oncogenes that control the cell growth. When it mutates (changes) or there are too many replicas of it, it becomes a "bad" gene. This bad gene is called oncogenes. Further, another gene called tumor suppressor genes slow down cell division rate (a process known as apoptosis). When these genes do not function properly, uncontrolled growth of cells occurs. This can lead to cancer. Sometimes, certain inherited DNA mutations (changes) also increase the risk of developing breast cancer. The mutations of oncogenes and/or tumor suppressor genes are also influenced by some other factors such as radiation or cancer-causing chemicals. Yet the causes of mutations that lead to breast cancer are still unknown (Hanahan & Weinberg, 2011).

Cancerous cells usually accumulate to form a lump called a tumor or a mass that can be seen with X-Ray or felt by hands. However, not all tumors are cancerous. Non-cancerous tumors are called benign. Tumors that are cancerous are called malignant. Ma-

* Corresponding author.
*E-mail addresses:* lingraj02uce157ster@gmail.com (L. Dora), agrawals_72@yahoo.com (S. Agrawal), r_ppanda@yahoo.co.in (R. Panda), ajith.abraham@ieee.org (A. Abraham).

http://dx.doi.org/10.1016/j.eswa.2017.05.035
0957-4174/

lignant (cancerous) tumor can spread into surrounding tissues or distant metastasis. Malignant tumors can spread when the cancer cells get into the blood or lymph system and are transferred to the nearby parts of the body. Mostly, breast cancer starts in the ducts that carry milk to the nipple (ductal cancers). Some begin in the glands that make breast milk (lobular cancers). There are also some other types of cancer that start in other breast tissues like sarcomas and lymphomas, but they are very rare. The malignant tumor found in the breast tissue is identified as breast cancer (Muto, Bussey, & Morson, 1975). Benign breast tumor is also the result of abnormal growths inside the breast. However, they grow deliberately, and do not spread outside the breast. They are not life threatening. However, some benign tumors can increase the risk of getting breast cancer. The most common symptom of breast cancer is a new painless lump or hard mass developed inside the breast. Sometimes breast cancer can be soft and painful. For this reason, it is suggested that any breast lump or growth needs to be checked by the doctor to identify whether it is benign or malignant. It also reduces the impact of getting future cancer risk.

A lot of research around the world is going on to detect the causes of breast cancer and to develop a preventive measure. Early detection of breast cancer and finding state-of-the-art cancer treatment could be a preventive measure to reduce its mortality rate (Christoyianni et al., 2000; Rodrigues et al., 2006). Diagnosis of breast cancer at an early stage is associated with improved outcome because timely treatment can be given preventing the loco regional spread or distant metastasis. During its early stage, a tumor is small and has not spread, can be localized for effective treatment. Regular screening is the most reliable way of early breast cancer detection. This will be a reliable method of predicting the prognosis of a woman with this disease. Normally, breast cancer diagnosis is clinical and biological in nature, performed by doctors. The development of more effective diagnostic techniques and improved treatment planning has attracted significant attention to breast cancer cases. The most common techniques available for breast cancer diagnosis are biopsy, mammography, ultrasound (US) imaging, magnetic resonance imaging (MRI) scans and experimental breast imaging.

Among these techniques, mammography is the first modality for screening or detection of breast cancer. In this technique, the doctor to look for the symptoms of breast cancer uses an X-ray image of the breast (mammogram). Usually, symptoms like small white spots called calcifications, lump or tumor and other suspicious area indicate signs of breast cancer. However, the radiologist's decision may vary while interpreting the mammography (Elmore, Wells, Lee, Howard, & Feinstein, 1994). Moreover, mammography suffers from limitations such as false-negative results, false-positive results, etc. Ultrasound is often used to examine some changes in the breast that can be felt but not visualized on a mammogram. It may also be used for woman with dense breast tissues. Further, it is also used to guide a biopsy needle into the suspicious area that can be taken out and tested for cancer. However, breast ultrasound imaging test solely depends on the skill and experience of the doctor interpreting the images. Breast biopsy is a technique usually followed after getting suspicious results from a screening test. Cells from suspicious area is taken out by a surgeon and tested in the laboratory to detect the presence of cancer cells. There are different types of biopsy such as fine needle aspiration cytology (FNAC) biopsy, core needle biopsy, surgical biopsy and lymph node biopsy. FNAC is a standard work-up in patients with suspected breast cancer. A doctor takes the tissue sample, irrespective of the prescribed biopsy, and a pathologist for final diagnosis examines it. The judgement given by the multidisciplinary oncology board consisting of radiologist, oncologist and pathologist is the final decision in breast cancer diagnosis. However, machine learning techniques could assist them to improve their diagnostic capability. It will further reduce the errors committed due to fatigue or inexperience. In addition, they can expedite examining medical data in a shorter time with details (Şahan, Polat, Kodaz, & Güneş, 2007).

A wide variety of methods based on machine learning, expert systems and soft computing have been proposed in the literature, to solve the breast cancer diagnosis problem (Bhardwaj & Tiwari, 2015; Malmir, Farokhi, & Sabbaghi-Nadooshan, 2013; Marcano-Cedeño, Quintanilla-Domínguez, & Andina, 2011). Earlier, data mining and machine learning techniques were integrated into a computer aided diagnosis (CAD) system for classification of breast cancer. The successful execution of this approach transformed the breast cancer diagnosis into a 2-class (benign or malignant) classification problem (Wolberg, Street, & Mangasarian, 1995). The objective of breast cancer classification is to identify an unknown tumor sample (test sample) as benign or malignant by using labeled training samples. Almost all the methods consider WBCD taken from the UCI machine learning repository (Bache & Lichman, 2013) to experiment.

Quinlan (1996) proposed a 10 – fold cross validation in conjunction with C4.5 tree method. The author was silent about the effect of changing the training samples on the classification accuracy. Hamilton, Shan, and Cercone (1996) used a rule induction algorithm based on the approximate classification method to achieve an accuracy of 96%, which is low. Abonyi and Szeifert (2003) applied the supervised fuzzy clustering technique and obtained an accuracy of 95.57%. This is also a lower classification rate and thus, not fit for clinical usages. Şahan et al. (2007) employed a hybrid machine learning method in breast cancer diagnosis. The method involves two stages. Firstly, fuzzy-artificial immune system was applied to reduce the dimension of the dataset. Secondly, k-NN performs the classification in the reduced space. In this manner, the processing time for classification is reduced. The method achieved a high classification accuracy of 99.14% via 10-fold cross validation. However, the authors did not investigate the effect of partitioning the training and the test samples.

Peng, Wu, and Jiang (2010) suggested a technique which integrates wrapper-based and filter-based feature extraction methods. Such feature extraction methods are efficient enough to select a subset of features from original feature space, without loss of significant information. The method achieved an accuracy of 99.5%. The method suffers from computational complexity. Recently, Chen (2014) extended the idea of feature selection to cluster analysis techniques to improve their operation. The author proposed a hybrid intelligent technique for breast cancer diagnosis. Chen, Yang, Liu, and Liu (2011) used rough set (RS) theory for feature selection. The authors implemented RS to select the significant optimal feature followed by support vector machine (SVM) for classification. They applied a subset that contains five optimal features. These authors worked in a high dimensional space using SVM.

Übeyli (2007) applied five different classifiers using neural networks for breast cancer diagnosis. The Levenberg-Marquardt algorithm is used to train the dataset for different classifiers. Zheng, Yoon, and Lam (2014) proposed a hybrid method (K-means + SVM) and achieved a classification accuracy of 97.38% using 10-fold cross validation. The above authors used highly computational complex methods for classification.

Örkcü and Bal (2011) implemented back propagation neural network (BPNN), binary coded genetic algorithm (GA) and real coded GA for breast cancer diagnosis. Marcano-Cedeño et al. (2011) explained a novel artificial metaplasticity multilayer perceptron (MLP) for breast cancer diagnosis. Malmir et al. (2013) trained an MLP using an imperialistic competitive algorithm (ICA) and particle swarm optimization (PSO) for 40 iterations to achieve a classification accuracy of 97.75%

and 97.63% respectively. Koyuncu and Ceylan (2013) implemented rotation forest artificial neural network (ANN) using 9-classifiers and achieved a classification accuracy of 98.05%. Xue, Zhang, and Browne (2014) proposed a novel technique for initializing and updating in PSO for feature selection to achieve an accuracy of 94.74%. It is observed that most of the authors used neural network for classification, which is inherently complex.

Bhardwaj and Tiwari (2015) proposed a genetically optimized neural network (GONN) algorithm to carry out the breast cancer classification. The method used GA to develop an optimized neural network architecture. The method implements new crossover and mutation operators for GA. However, the computational complexity is more due to the use of GA and ANN. Further, the authors investigated the diagnosis problem on WBCD only. Moreover, to deal with real life problem, efficiency of methods on the breast cancer represented by imaging techniques like mammography, US etc. need to be evaluated. The imaging techniques define the diagnostic features based on analysis of the images and are usually not very well correlated with a class. In this paper, our focus is on evaluation of GNRBA on two different breast cancer datasets (WBCD and WDBC). The performance of GNRBA on different imaging modalities is beyond the scope of the paper and can be considered as our future work. Nilashi, Ibrahim, Ahmadi, and Shahmoradi (2017) proposed a knowledge-based system, which uses expectation maximization to cluster the data into similar groups. The method uses classification and regression trees to generate a set of fuzzy rules. Classification of breast cancer is done based on the fuzzy rules. Further, to deal with multi-collinearity problem, principal component analysis (PCA) is used. The method shows a good classification accuracy on WBCD and mammography mass datasets. However, the method is based on non-incremental data mining technique and evaluated on small datasets. Sometimes, classification algorithms may be influenced by certain properties of the diagnostic images such as mammography. Early detection of breast cancer may depend on factors like differences in instrument setting or breast positioning by the operator. Taking these limitations into account, Magna et al. (2016) proposed an adaptive artificial immune based system to investigate the properties of classification methods, applied to mammography images. The classification methods are trained to measure bilateral asymmetry between paired regions of the right and left breasts. Still, such a system requires proper investigation to identify focal and global asymmetric features. Additionally, attention is required to extract features to localize the source of asymmetry. Wang, Hu, Li, Liu, and Zhu (2016) proposed an automatic quantitative image analysis method for classification of breast cell nuclei as benign or malignant. The author uses breast cell histopathology (BCH) images. Segmentation of cell nuclei is done by using wavelet decomposition and multi-scale region growing. A double-strategy splitting model is used to partition the overlapping cells. A hybrid technique, which includes SVM and chain-like genetic algorithm to extract optimal features (shape-based and texture-based features) for classification of cell nuclei, is presented. The method shows a good classification accuracy of 96.19%, using 68 BCH images. However, to prove the robustness of the method, evaluation on a large dataset is preferred.

Nowadays, computer-aided diagnosis (CAD) systems are gaining attention on the breast cancer classification problem for automatic diagnosis. Abdel-Zaher and Eldeib (2016) proposed a CAD system based on deep-belief neural network. The network uses back propagation with Levenberg-Marquardt learning. The method uses WBCD for breast cancer classification. Nevertheless, such algorithm has inherent high computational complexity. Therefore, it is difficult to implement them on any commercially available hardware, to assist doctors in early detection of breast cancer. Moon, Chen, Chang, Shin, Lo, and Chang (2016) proposed an adaptive CAD system for classification of breast tumor. The CAD system utilizes tumor sizes for classification. US images of breast tumor detected during screening are used for classification. The CAD system uses quantitative morphological and texture features for classification. However, speckle noise and other artifacts that are inherent in US image may degrade the performance of such a system. To overcome such problems, Abdel-Nasser, Melendez, Moreno, Omer, and Puig (2017) proposed a super-resolution based CAD system. The system uses multiple US images of breast instead of a single image. The CAD system uses four different steps like super-resolution computation, segmentation, feature extraction and classification to detect tumors and classify them as benign or malignant. The idea can be extended to other medical imaging modalities. Nowadays, a new ultrasound elastography imaging technique called shear wave elastography (SWE) is increasingly used for classification of breast lesion. SWE provides opportunity to evaluate elasticity parameters of breast, an important information about cancerous tissue. Acharya et al. (2017) proposed a method for automatic characterization of malignant breast lesion. The method uses SWE to evaluate discrete wavelet coefficients in three different levels. Features like run length statistics and Hu's moments are extracted from the coefficients. Significant features are then extracted by using sequential forward selection methods and ranked using the ReliefF ranking technique. Different classifiers for classification of benign and malignant lesions use the ranked features. The method achieved a classification accuracy of 93.59%.

Recently, sparse-based representation methods are being successfully implemented in pattern recognition for face recognition and verification (Mei, Ling, & Jacobs, 2011; Yuan, Liu, & Yan, 2012). It requires that a test face image be sparsely represented by the contributions of each class. A group of similar training face images represents a class. For each class, the contribution ability is represented by the weighted sum of the training images. The class contribution ability reflects the contribution of each class on a test image. Then the test face image is classified in a class having maximum class contribution ability. Here, sparse indicates that weighting coefficients of some of the training face images in a class are zero with respect to the matched class. The extent of sparsity of the weighting coefficients can be evaluated by using $l_1$ - norm of the weighting coefficient vector (Mahmoudi & Sapiro, 2012; Zhang, Yang, & Feng, 2011). A smaller norm indicates a stronger sparsity. In reality, conventional sparse based representation methods tend to minimize the $l_1$-norm of the coefficient vector.

To the best of our knowledge, sparse-based representation methods are not applied to breast cancer classification problem. This motivates us to apply it to breast cancer classification. Here, we propose a novel Gauss-Newton representation based (GNRBA) method for the breast cancer classification problem. It is wise to reiterate the fact that labeled tumor features represents training samples and an unknown tumor sample represents test sample. The proposed method is based on sparse representation of a test sample by the linear weighted sum of all the training samples. First, a subset of significant training samples (tumor features) is selected from the total training samples by using the Euclidean distance measure. All the significant training samples in the subset belongs to either of the two classes i.e. benign or malignant. Then, for each class, the proposed method evaluates the class contribution ability. The maximum class contribution ability criteria are used to classify a test sample. The block diagram of the proposed method is shown in Fig. 1.

The main contributions of the proposed method are: (1) Selection of a subset of significant training samples which are compatible with the proposed GNRBA to obtain accurate classification in a reduced dimension space. (2) Evaluation of sparsity of the weighting coefficients with reduced computational complexity. (3) Providing an optimal solution to evaluate the sparsity of weighting coefficients, as compared to the conventional $l_1$-norm method. To show

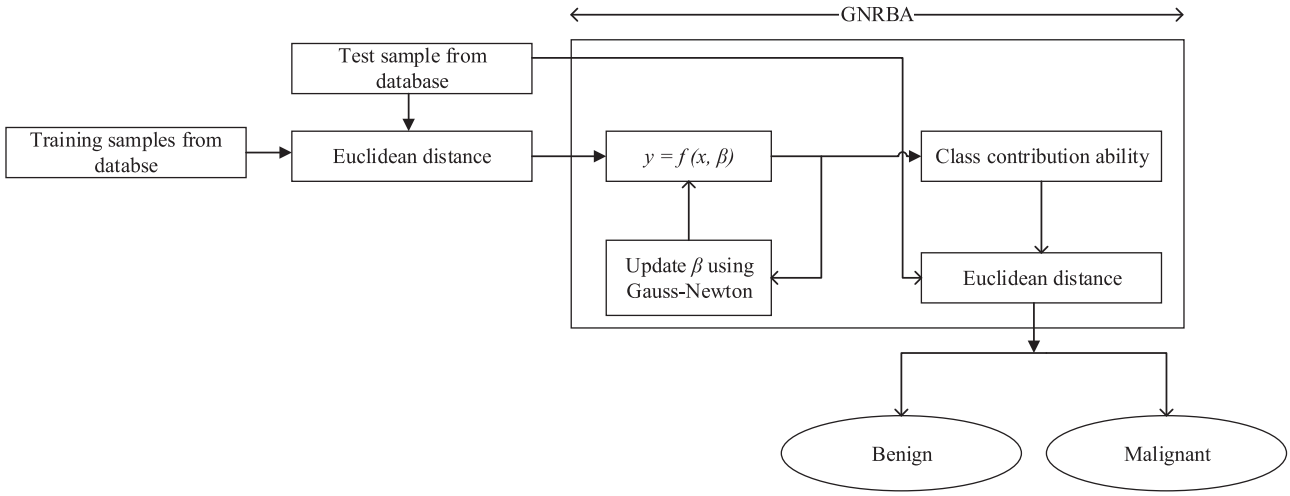


**Fig. 1.** Block diagram of the proposed GNRBA.

**Table 1**
Description of WBCD.

| Attribute number | Attribute Description | Values | Mean | Standard deviation |
|---|---|---|---|---|
| 1 | Clump thickness | 1–10 | 4.44 | 2.83 |
| 2 | Uniformity of cell size | 1–10 | 3.15 | 3.07 |
| 3 | Uniformity of cell shape | 1–10 | 3.22 | 2.99 |
| 4 | Marginal adhesion | 1–10 | 2.83 | 2.86 |
| 5 | Single epithelial cell size | 1–10 | 2.23 | 2.22 |
| 6 | Bare nuclei | 1–10 | 3.54 | 3.64 |
| 7 | Bland chromatin | 1–10 | 3.45 | 2.45 |
| 8 | Normal nucleoli | 1–10 | 2.87 | 3.05 |
| 9 | Mitoses | 1–10 | 1.60 | 1.73 |

the accuracy of the proposed method, we have compared with other methods applied to WBCD. From the study, it is observed that most of the above discussed methods use WBCD to experiment (Abdel-Zaher & Eldeib, 2016; Bhardwaj & Tiwari, 2015; Chen et al., 2011; Koyuncu & Ceylan, 2013; Malmir et al., 2013; Nilashi et al., 2017). To the best of our knowledge, very few researchers have used WDBC to experiment. In this paper, we have considered both WBCD and WDBC to experiment to show the robustness of our proposed method. The results show that the proposed method performs well as compared to the other methods. Further, it can be a better alternative to the well-established expert systems and machine learning methods. Nevertheless, we can verify the method using some private data as well but then such data won't be available for public and verification would be difficult. The rest of the paper is organized as follows. Section 2 describes the databases used to experiment. Section 3 presents the proposed method. Experimental results and discussion are presented in Section 4. Finally, Section 5 is the conclusion.

## 2. Breast cancer database

In this paper, we have carried out the experiment on WBCD and WDBC database (Bache & Lichman, 2013; Bennett & Mangasarian, 1992; Mangasarian & Wolberg, 1990; Mangasarian, Setiono, & Wolberg, 1990; Wolberg & Mangasarian, 1990). The UCI machine learning repository contains two types of breast cancer datasets taken from human breast tissue, for breast cancer classification problem.

*WBCD*: This dataset contains tumor features of Fine Needle Aspirates (FNA) of human breast tissue. It contains tumor features of 699 subjects. The dataset has 9 attributes along with a class label (benign or malignant) and a subject ID, corresponding to each subject. The name and value of each of the 9 attributes are listed in Table 1. Each attribute is an integer value [1–10], where the value of 10 represents the critical state. Out of the 699 subjects, 16 subjects have some attribute values missing. Thus, they are not included in the experiment. We consider only the rest 683 subjects. Out of the 683 subjects, 444 are in the benign class and 239 in the malignant class. The primary objective of the breast cancer classification problem is to correctly classify the tumor feature of an unknown subject as benign or malignant.

*WDBC*: This dataset contains tumor features obtained from a digital image of breast FNA. In this dataset, 32 tumor features of 569 subjects are presented. The 32 features represent (a) 30 actual tumor features, (b) a subject ID number and (c) a class label, which denotes each subject has benign or malignant tumor. For each subject, 10 attributes of cell nuclei (visible in a digital image of breast FNA) are obtained, such as radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, concave points and fractal dimension. Then different measurement like mean, standard error and maximum of these 10 attributes are calculated which results in 30 features, as depicted in Table 2. These measurements are considered as tumor features in the dataset. As the measurements are in different scales, we need to normalize the dataset before training.

## 3. Proposed method

In this paper, we have proposed a novel GNRBA for solving breast cancer classification problem. The proposed work is based on sparse representation of a test sample by the linear weighted sum of all the training samples. The input training samples are taken from the database. The classification task is performed in two steps. In the first step, a subset of the most significant training samples is selected. In the second step, the subset is utilized to classify the unknown test sample.

**Table 2**
Description of WDBC database.

| Attribute number | Attributes | Attributes range | | |
|---|---|---|---|---|
| | | Mean | Standard error | Maximum |
| 1 | Radius | 6.98–28.11 | 0.112 – 2.873 | 7.93 – 36.04 |
| 2 | Texture | 9.71–39.28 | 0.36–4.89 | 12.02–49.54 |
| 3 | Perimeter | 43.79–188.50 | 0.76–21.98 | 50.41–251.20 |
| 4 | Area | 143.50–2501.00 | 6.80–542.20 | 185.20–4254.00 |
| 5 | Smoothness | 0.053–0.163 | 0.002–0.031 | 0.071–0.223 |
| 6 | Compactness | 0.019–0.345 | 0.002–0.135 | 0.027–1.058 |
| 7 | Concavity | 0.000–0.427 | 0.000–0.396 | 0.000–1.252 |
| 8 | Concave points | 0.000–0.201 | 0.000–0.053 | 0.000–0.291 |
| 9 | Symmetry | 0.106–0.304 | 0.008–0.079 | 0.157–0.664 |
| 10 | Fractal dimension | 0.050–0.097 | 0.001–0.030 | 0.055–0.208 |

### 3.1. Subset selection

Let $X=[x_1, x_2,\cdots, x_t]^T$ denotes a set of $t$ training samples taken from the database, where $T$ represents the matrix transpose. Each element $x_i$ for $i=1, 2, \cdots, t$ is a training sample (tumor feature of a subject) that belongs to either of the two classes i.e. benign or malignant, as discussed in Section 2. In sparse representation, the class contribution ability is guided by the training samples of that class. The training samples that minimize their class contribution ability are treated as the unsuitable samples. On the contrary, the samples that maximize their ability are considered as suitable training samples. Earlier research has shown that the training samples that are nearest to the test sample are considered useful for accurate classification. Therefore, we need to select a subset of the training samples nearest to the test sample in order to maximize their class contribution ability. In the literature, many approaches are available to find the distance measure between the training samples and the test sample. Then this measure is used to select the nearest and the farthest training samples from the test sample (Xu, Zhang, Yang, & Yang, 2011). In this paper, we have taken the Euclidean distance $d_i$ to calculate the distance between a test sample $y$ and a training sample $x_i$ as given below.

$$d_i = \|y - x_i\|_2 \tag{1}$$

From Eq. (1), a small $d_i$ value indicates the training sample $x_i$ is nearest to the test sample $y$. A large $d_i$ indicates that the sample $x_i$ is an unsuitable training sample. Thus, we use Eq. (1) to select a subset containing an $N$ number of suitable training samples. The remaining training samples are discarded. Let $X_S=[x_1,x_2,\cdots, x_N]^T$ be the subset of the $N$ selected suitable training samples. The class labels (benign or malignant) of the corresponding training samples in the subset are saved in $C=[c_1,c_2,\cdots, c_N]^T$.

### 3.2. GNRBA

In this section, we explain the proposed GNRBA to classify the unknown test sample. The subset ($X_S$) of the training samples is used as the input to the method. The proposed GNRBA represents a test sample $y$ as a linear weighted summation of the training samples, given as follows:

$$y = x_1\beta_1 + x_2\beta_2 + \cdots + x_N\beta_N \tag{2}$$

where, $y$ is an unknown test sample; $\beta=[\beta_1,\beta_2,\cdots, \beta_N]$ is a vector consisting of $N$ weighting coefficients assigned to each training sample $x_i$ for $i=1, 2, \cdots, N$. Each sample $x_i$ from the subset belongs to either of the two class i.e. benign or malignant. Let the benign class contain a $C_B$ number of samples and malignant class contains $C_M$ samples such that $C_B+C_M=N$. Then Eq. (2) can also be represented as:

$$y = x_b\beta_b + x_m\beta_m \tag{3}$$

Where $x_b$ for $b=1, 2, \cdots, C_B$ is a training sample that belongs to benign class having a weighting coefficient $\beta_b$. Note that $x_m$ is a training sample that belongs to malignant class having a weighting coefficient $\beta_m$ for $m= 1, 2, \cdots, C_M$. Now Eq. (3) represents a test sample as a weighted summation of training samples from each class. In addition, it also represents the class contribution ability of each class, which is guided by all the training samples of that class.

Initially, $\beta$ is taken randomly. Different random values of $\beta$ will inconsistently represent the contribution of training samples towards their class. So, it is required to generate the optimal $\beta$ such that it consistently represents the contribution of training samples. In other words, if $\beta$ is the best solution, contribution of the training samples towards their class can be calculated according to the elements of $\beta$. In the literature, it is suggested that, a method to find optimal $\beta$ is to optimize the sum of square error (SSE), which is given as follows (Gill, Murray, & Wright, 1981)

$$SSE = \min_{\beta} \|y - X_S\beta\|_2^2 + \lambda\|\beta\|_2^2 \tag{4}$$

where $\lambda$ is a regularization parameter used to obtain a stable solution by avoiding singularity.

We have used the Gauss-Newton formula (Gill et al., 1981) to minimize the objective function in Eq. (4) in order to generate the optimal $\beta$. The update equation is given as follows,

$$\beta_{next} = \beta_{now} + \Delta\beta \tag{5}$$

where $\Delta\beta = (X_S^T X_S + \lambda I)^{-1} X_S^T \Delta Y$, $I$ is the identity matrix and $\Delta Y$ represents the difference between the desired output value and the actual output value when the input is $X$ as defined in (Gill et al., 1981).

The proposed method is easy to implement with less computational complexity as compared to the conventional sparse representation method. The regularization parameter tends to zero, when approaching optimum point. However, in this paper, we have used the regularization parameter, $\lambda=0.01$ as suggested in (Jang, Sun, & Mizutani, 1997; Zhang et al., 2011) for this update. Specifically, we have to decide a stopping criterion to obtain best $\beta$, which is an optimum solution. The termination condition used to obtain the best $\beta$ can be the maximum number of iterations or until the objective function converges. We have considered the former as the stopping criteria. Once the optimal $\beta$ is obtained using Eq. (5), class contribution ability of the $N$ selected training samples in the subset ($X_S$) can be used to represent the test sample.

However, the selected training samples in the subset might be from different classes. Therefore, we consider the sum of the weighted training samples from the same class (i.e. class contribution ability) to represent the test sample. Then the sum is used to compute the distance between the test sample and the class contribution of each class. If the selected training samples in the subset are from $k^{th}$ class, then their class contribution ability is

**Table 3**
Parameters setting for the proposed GNRBA.

| Parameters | Value | |
|---|---|---|
| | WBCD | WDBC |
| Size of the subset, $N$ | 74 (50–50 partition) | 160 (50–50 partition) |
| | 92 (60–40 partition) | 54 (60–40 partition) |
| | 2 (70–30 partition) | 43 (70–30 partition) |
| | 2 (10-fold cross validation) | 43 (10-fold cross validation) |
| Regularization parameter, $\lambda$ | 0.01 | 0.01 |
| Termination criteria | 50 iterations | 50 iterations |

calculated as:

$$X_k = \sum_{i=1}^{N_k} \beta_i X_{Si} \tag{6}$$

where, $N_k$ is the number of training samples from $k^{th}$ class; $\beta_i$ is the $i^{th}$ coefficient value of the $i^{th}$ training sample $X_{Si}$ and $X_k$ indicates the class contribution ability of $k^{th}$ class. Then the sum in Eq. (6) is utilized to classify the test sample $y$ by using

$$D_k = \|y - X_k\|_2 \tag{7}$$

In Eq. (7), a smaller distance $D_k$ between test sample $y$ and $X_k$ indicates a greater contribution of the $k^{th}$ class to represent the test sample. Hence, the test sample is classified into the class $k$. The steps of the proposed GNRBA are presented as follows

*Algorithm* of the proposed GNRBA for classification of breast cancer.

**Input**: $X$, training samples i.e. tumor features from WBCD/WDBC database.
**Output**: classification of the testing samples or unknown tumor feature.
**for** $j = 1$ *to* $T$ (number of testing samples)
Calculate $d_i$ between $X$ and $y_j$ ($i^{th}$ test sample) using Eq. (1)
$X_S = \min_{d_i}\{X_i\}$ for $i = 1, 2, \cdots, N$ (selection of subset)
  **if** class of $X_S$ = benign
    label = 1
  **else**
    label = 2
  **end if**
  **for** iteration = *1 to termination criteria*
    find optimal $\beta$ using (5)
  **end for**
  **for** $c = 1$ *to* 2 (number of classes)
    **for** $k = 1$ *to* $N$
      **if** $c$ = *label*
        find $X_k$ using (6)
      **end if**
    **end for**
  **end for**
  **for** $k = 1$ *to* 2
    find $D_k$ using (7)
  **end for**
  **if** $min\{D_k\} = 1$
    class = benign
  **else**
    class = malignant
  **end if**
**end for**

## 4. Results and discussions

The proposed GNRBA is implemented in MATLAB and on a MAC computer using Intel core i5. Two different datasets (WBCD and WDBC) as discussed in Section 2, are used to demonstrate the performance of the proposed method. Experiments are carried out on the two datasets with the parameters setting, as described in Table 3. To justify the robustness of the proposed approach, the noisiness of the evaluation dataset is performed. Attribute noise has been generated and introduced in the training and test samples of both the datasets.

In expert systems and machine learning techniques, a usual practice is to split the entire dataset into two distinct sets. To determine the performance of the proposed method as compared with other methods, we have partitioned the entire dataset in four different ways. A standard 50–50 training – testing partition is used, where half of the samples from the dataset are used to train the classifier and the remaining samples are used for testing. To show the influence of training data on the proposed method, we have also partitioned the dataset (60–40), (70–30) training and testing ratios, respectively. The process of partitioning the database (WBCD and WDBC) into training and testing sets is represented in Tables 4 and 5, respectively.

In addition, a 10-fold cross validation technique (Hastie, Tibshirani, & Friedman, 2009) is also used to measure the robustness of the proposed method. 10-fold cross validation method partitions the entire data set into ten blocks of equal size. While implementing the 10-fold cross validation in the proposed GNRBA, we have used 90% of the dataset for training and the remaining 10% for testing. The total number of test samples, cover the whole dataset (i.e. 10 repetitions of the 10% of testing data, exchanged in all runs). For each dataset, we evaluate 10-fold cross validation 10 times. The average accuracy of all the runs is taken as the final classification accuracy.

The WBCD database contains a total of 683 samples. For 10 – fold cross validation, these samples are divided into 10 blocks. Each block is used for testing in one run. The process is repeated 10 times, by exchanging the blocks in each run. As the total sample is not a multiple of 10, hence, 68 samples are used for testing in 9 runs and 71 samples are used for testing in the last run. In this way, the testing samples cover the whole dataset.

The WDBC database contains a total of 569 samples. It is wise to reiterate that, here the total sample is not a multiple of 10. For 10 – fold cross validation, 57 samples are used for testing in 9 runs and 56 samples are used for testing in the final run. In this way, the testing samples cover the whole dataset.

Validation of the classification method is a necessary step to evaluate its performance and limitations. Moreover, in the literature, it is also suggested to validate a method before applying for clinical use. In this paper, we have presented different validation measures like classification accuracy (CA), sensitivity, specificity, confusion matrix, receiver operating characteristic (ROC) curves and area under ROC curves (AUC), to evaluate the performance of the proposed GNRBA (Fawcett, 2006; Sokolova & Lapalme, 2009).

We have carried out experiments on WBCD and WDBC to show the dependence of the GNRBA on $N$ (size of the subset). The experimental results are presented in Fig. 2. The results show the variation of CA with respect to $N$, for different training – testing partition.

From Fig. 2, it is observed that when the value of $N$ is small, the GNRBA tends to obtain highest classification accuracy. The reason behind this is that, when $N$ is small, many of the unsuitable training samples that put a negative effect on the classification of the test sample, are eliminated. Normally, the value of $N$ is kept in the range($0.05 \times t$, $0.02 \times t$), where $t$ is the total number of train-

**Table 4**
Training and testing set partition of WBCD (Total 683 samples).

| Training-testing partition ratio | Number of samples in training | | | Number of samples in testing | | |
|---|---|---|---|---|---|---|
| | Total training samples | Benign | Malignant | Total testing samples | Benign | Malignant |
| 50–50 | 341 | 222 | 119 | 342 | 222 | 120 |
| 60–40 | 410 | 266 | 144 | 273 | 178 | 95 |
| 70–30 | 478 | 311 | 167 | 205 | 133 | 72 |

**Table 5**
Training and testing set partition of WDBC (Total 569 samples).

| Training-testing partition ratio | Number of samples in training | | | Number of samples in testing | | |
|---|---|---|---|---|---|---|
| | Total training samples | Benign | Malignant | Total testing samples | Benign | Malignant |
| 50–50 | 284 | 139 | 145 | 285 | 218 | 67 |
| 60–40 | 340 | 183 | 157 | 229 | 174 | 55 |
| 70–30 | 397 | 224 | 173 | 172 | 133 | 39 |

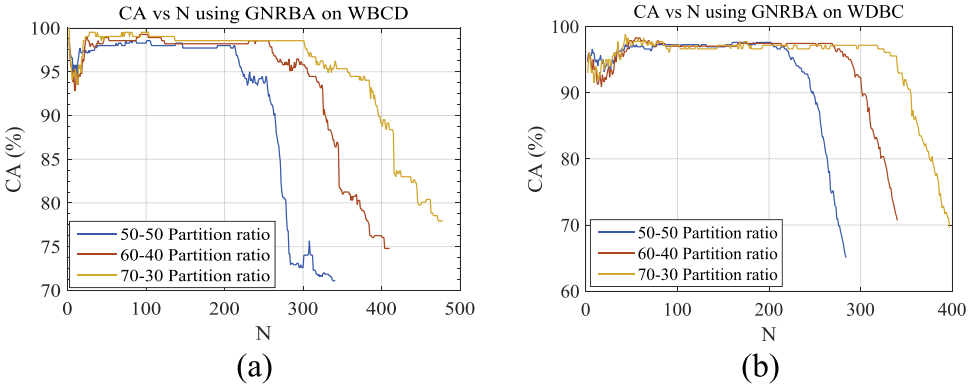


**Fig. 2.** Variation of CA with respect to $N$ using GNRBA. (a) WBCD (b) WDBC.

**Table 6**
Comparison of classification accuracy on WBCD.

| Training-testing partition ratio | BPNN | Koza's model | GONN | GNRBA |
|---|---|---|---|---|
| 50–50 | 89.81% | 89.63% | 97.73% | **98.56%** |
| 60–40 | 92.28% | 92.84% | 99.11% | **99.27%** |
| 70–30 | 93.45% | 94.14% | 99.21% | **100%** |
| 10–fold cross validation | 89.11% | 93.47% | 99.26% | **99.23%** |

ing samples. From Fig. 2(a), it is observed that for $N = 74$, 92 and 2, our proposed method results in the highest CA of 98.56% for 50 – 50 partition, 99.27% for 60 – 40 partition, 100% for 70 – 30 partition respectively. From Fig. 2(b), it is observed that $N = 160$, 54 and 43 produces a maximum CA of 97.54% for 50 – 50 partition, 98.25% for 60 – 40 partition, 98.86% for 70 – 30 partition respectively. It is to be noted that all these values of $N$ are within the specified range. Fig. 2 shows that the selection of $N$ works appropriately on both the datasets, which indicates that inclusion of too many training samples affects the classification accuracy. In real time applications such as clinical diagnosis, first the value of $N$ is set to the range($0.05 \times t$, $0.02 \times t$). Then, the cross-validation method is applied to select the optimal value of $N$.

The classification accuracy of GNRBA is compared with the Koza's model (Koza & Rice, 1991), BPNN model (Hagan, Demuth, Beale, & De Jesús, 1996) and GONN model (Bhardwaj & Tiwari, 2015) using WBCD for different training – testing partition, as depicted in Table 6. The architecture of Koza's model and GONN model is set to 50 genetic programming runs. They contain one input layer, one hidden layer and one output layer as per the architecture of BPNN. The Koza's and GONN models used only 4 inputs in the input layer. However, the BPNN model used all the 9 features with a bias input in the input layer and 8 neurons in the hidden layer. The activation function used in all the models (Koza's, BPNN and GONN) is the standard sigmodal function (Bhardwaj & Tiwari, 2015). The value of $N$ in the proposed GNRBA is set to 74 for 50–50 partition ratio, 92 for 60 – 40 partition ratio, 2 for 70–30 partition ratio and 2 for 10 – fold cross validation. The values of $N$ are obtained from Fig. 2.

From Table 6, it is observed that the proposed GNRBA outperforms all the three models (Koza's, BPNN and GONN) in terms of CA. The maximum, mean and standard value of CA for 10 – fold cross validation is 100%, 99.23% and 0.6708, respectively. The CA for 70–30 partition ratio is 100%, which is much higher as compared to Koza's and BPNN models. This improvement in CA is mainly due to two reasons: (1) the selection of a subset instead of using all the training samples provides an opportunity to eliminate unsuitable training samples which could result in misclassification and (2) optimized weights are used to find class contribution ability on the test sample for correct classification, as discussed in Section 3. However, the mean value of 10 – fold cross validation for GNRBA is marginally lower as compared to GONN model. The reason behind this may be the improper selection of the subset in 10 – fold cross validation. The results show that the proposed GNRBA can be a very supportive tool for experts decision.

The comparison of sensitivity and specificity for BPNN, Koza's, GONN and GNRBA is presented in Table 7 using WBCD for different training – testing partition. High values of sensitivity and specificity represents the correct classification ability of a model. For all the methods, the maximum values of sensitivity and specificity are not reported in the literature. From Table 7, it is observed that the GNRBA outperforms all other methods in terms of sensitivity and specificity. However, sensitivity for 50–50 partition and specificity for 10 – fold cross validation of the GNRBA is comparable to GONN

**Table 7**
Comparison of sensitivity and specificity on WBCD.

| | | | Training-testing partition ratio | | | |
|---|---|---|---|---|---|---|
| | | | 50–50 | 60–40 | 70–30 | 10-fold cross validation |
| BPNN | Sensitivity | Mean | 93.9 | 96.30 | 96.56 | 94.46 |
| | | Std dev | 0.213 | 0.282 | 0.378 | 1.144 |
| | Specificity | Mean | 82.05 | 84.77 | 87.21 | 77.21 |
| | | Std dev | 0.426 | 0.532 | 1.380 | 2.110 |
| Koza's model | Sensitivity | Mean | 93.79 | 96.31 | 96.26 | 96.5 |
| | | Std dev | 1.094 | 0.858 | 1.079 | 1.141 |
| | Specificity | Mean | 82.98 | 85.84 | 87.67 | 88.06 |
| | | Std dev | 0.416 | 1.605 | 1.349 | 0.655 |
| GONN | Sensitivity | Mean | 98.85 | 99.17 | 99.51 | 98.77 |
| | | Std dev | 0.228 | 0.508 | 0.372 | 1.167 |
| | Specificity | Mean | 95.77 | 98.45 | 99.21 | 100 |
| | | Std dev | 0.842 | 0.532 | 0.700 | 0 |
| GNRBA | Sensitivity | Mean | – | – | – | **99.44** |
| | | Std dev | – | – | – | **1.7392** |
| | | Max | **98.50** | **99.44** | **100** | **100** |
| | Specificity | Mean | – | – | – | **99.98** |
| | | Std dev | – | – | – | **0.2211** |
| | | Max | **99.17** | **98.95** | **100** | **100** |

**Table 8**
Performance measures of GNRBA on WDBC.

| Training-testing partition ratio | | CA | Sensitivity | Specificity |
|---|---|---|---|---|
| 50–50 | | 97.54% | 98.62 | 94.30 |
| 60–40 | | 98.25% | 97.70 | 100 |
| 70–30 | | 98.86% | 98.50 | 100 |
| 10–fold cross validation | Mean | 98.46% | 97.89 | 99.78 |
| | Std dev | 0.9437 | 0.8080 | 0.6640 |
| | Max | 100% | 100 | 100 |

model. The reason may be the inclusion of some unsuitable training samples in the subset.

The performance measures of the GNRBA on WDBC is shown in Table 8. The value of $N$ is set to 160 for 50–50 partition, 54 for 60–40 partition, 43 for 70–30 partition and 43 for 10 – fold cross validation, for calculation of different performance measures.

Previously reported hybrid methods like ant colony optimization plus SVM (ACO-SVM) (Prasad, Biswas, & Jain, 2010), genetic optimization plus SVM (GA-SVM) (Prasad et al., 2010), PSO-SVM (Prasad et al., 2010) and k-means plus SVM (K-SVM) (Zheng et al., 2014) gives classification accuracies of 95.96%, 97.19%, 97.37% and 97.38% respectively using WDBC. It is observed that the GNRBA is superior in comparison to all other methods for all training – testing partition ratios as well as for 10 – fold cross validation method in terms of CA only. To the best of our knowledge, the other performance measures like sensitivity and specificity using WDBC are not reported in the literature. The values presented in Table 8 using the proposed method may be used as a reference for comparison.

Classification accuracy alone does not provide detailed information about the performance of a classifier. Consider a situation where the number of malignant cases is much higher than the number of benign cases. Even if all the benign cases are misclassified, the accuracy of the classifier is more. A performance measure like confusion matrix provides a detailed information about the classifier. A confusion matrix for a binary classifier contains two rows and two columns. Each element of the matrix denotes four possible outcomes, i.e. true positive (TP), true negative (TN), false positive (FP) and false negative (FN). In this matrix, elements along the main diagonal indicates the correct classification. Whereas, elements along the off-diagonal indicates an error in classification. The comparison of classification accuracies in terms of confusion matrix for BPNN, Koza's, GONN and GNRBA is presented in Table 9, for all training – testing partition ratios using WBCD. In the table, $B$ represents the output of benign cases and $M$ represents the output of malignant cases. From Table 9, it is observed that the sum of TP (true positive) and TN (true negative) increases with the increase in the number of training samples. It is also observed that TP and TN for 70 – 30 partition and for 10 – fold cross validation is 100%, which shows the superiority of the proposed GNRBA.

The confusion matrix for GNRBA obtained using WDBC is presented in Table 10. In the literature, confusion matrix evaluation using WDBC is not reported. Thus, the confusion matrix values presented in Table 10 may be used as reference values for comparison. From Table 10, it is observed that the sum of TP (true positive) and TN (true negative) increases with the increase in the number of training samples. It is also observed that TP and TN for 10 – fold cross validation is 100%, which shows the accuracy of the proposed GNRBA.

A comparison of AUC of different methods like BPNN, Koza's, GONN and the proposed GNRBA is presented in Table 11, using WBCD. The AUC values are obtained from the ROC curves using the trapezoidal rule. From Table 11, it is observed that AUC values of the proposed GNRBA are higher as compared to the existing models. However, AUC value of the GNRBA is very close to that of GONN model in case of 10 – fold cross validation.

A comparison of AUC values of methods like PCA-KNN, PCA-SVM, knowledge-based method (Nilashi et al., 2017) and the proposed GNRBA is presented in Table 12, using WDBC. The table represents the average AUC values using 10 – fold cross validation. The AUC values of the above mentioned methods using different partition ratios is not reported in the literature. The data for GNRBA on 50 – 50, 60 – 40 and 70 – 30 training – testing partition presented in Table 12 may serve as a reference for future work in this area. From the table, it is observed that the results of GNRBA for 10 – fold cross validation outperform all other methods. The results represented in Table 12 prove that GNRBA has better prediction accuracy in breast cancer classification.

To show the statistical significance, GNRBA is evaluated using the Friedman statistical test (Demšar, 2006; Garcia &

**Table 9**
Comparison of confusion matrix for WBCD.

| Name of the classifier | Desired result | Result of 50–50 partition | | Result of 60–40 partition | | Result of 70–30 partition | | Result of 10-fold cross validation | |
|---|---|---|---|---|---|---|---|---|---|
| | | *B* | *M* | *B* | *M* | *B* | *M* | *B* | *M* |
| BPNN | Benign samples | 209 | 13 | 172 | 6 | 129 | 4 | 43 | 1 |
| | Malignant samples | 21 | 99 | 14 | 81 | 8 | 64 | 4 | 20 |
| Koza's | Benign samples | 210 | 12 | 173 | 5 | 130 | 3 | 43 | 1 |
| | Malignant samples | 20 | 100 | 12 | 83 | 8 | 64 | 3 | 21 |
| GONN | Benign samples | 219 | 3 | 178 | 0 | 133 | 0 | 44 | 0 |
| | Malignant samples | 6 | 114 | 1 | 94 | 0 | 72 | 0 | 24 |
| GNRBA | Benign samples | **218** | **1** | **177** | **1** | **133** | **0** | **46** | **0** |
| | Malignant samples | **4** | **119** | **1** | **94** | **0** | **72** | **0** | **22** |

**Table 10**
Confusion matrix for WDBC.

| Name of the classifier | Desired result | Result of 50–50 partition | | Result of 60–40 partition | | Result of 70–30 partition | | Result of 10-fold cross validation | |
|---|---|---|---|---|---|---|---|---|---|
| | | *B* | *M* | *B* | *M* | *B* | *M* | *B* | *M* |
| GNRBA | Benign samples | 215 | 4 | 170 | 0 | 131 | 0 | 41 | 0 |
| | Malignant samples | 3 | 63 | 4 | 55 | 2 | 39 | 0 | 16 |

**Table 11**
Comparison of AUC on WBCD.

| Name of the method | AUC for different training – testing partition | | | |
|---|---|---|---|---|
| | 50–50 | 60–40 | 70–30 | 10-fold cross validation |
| BPNN | 0.883 | 0.909 | 0.929 | 0.873 |
| Koza model | 0.889 | 0.922 | 0.932 | 0.932 |
| GONN | 0.978 | 0.989 | 0.998 | 1.0 |
| GNRBA | **0.9982** | **0.9995** | **1.0** | **0.9977** |

**Table 12**
Comparison AUC on WDBC.

| Name of the method | Training-testing ratio | AUC |
|---|---|---|
| PCA-KNN | 10-fold cross validation | 0.8230 |
| PCA-SVM | 10-fold cross validation | 0.8670 |
| Knowledge-based method | 10-fold cross validation | 0.9320 |
| GNRBA | 50–50 | **0.9895** |
| | 60–40 | **0.9980** |
| | 70–30 | **0.9981** |
| | 10-fold cross validation | **0.9993** |

**Table 13**
*p* – value comparison of the selected samples using Friedman test on WBCD.

| Training-testing partition ratio | GNRBA | |
|---|---|---|
| | *p*-value | Significance |
| 50–50 | 0.6174 | Not significant |
| 60–40 | 0.5909 | Not significant |
| 70–30 | 0.1336 | Not significant |
| 10-fold cross validation | 0.2013 | Not significant |

**Table 14**
*p* – value comparison of the selected features using Friedman test on WDBC.

| Training-testing partition ratio | GNRBA | |
|---|---|---|
| | *p*-value | Significance |
| 50–50 | 0.6052 | Not significant |
| 60–40 | 0.8463 | Not significant |
| 70–30 | 0.1270 | Not significant |
| 10-fold cross validation | 0.5290 | Not significant |

Herrera, 2008). It is a non-parametric test used to show the statistical significance among multiple classifiers. Under the null hypothesis, two algorithms are significantly different for $p < 0.05$. We have used it to show the statistical significance of sample selection before classification as presented in Table 13 and Table 14, respectively. From Table 13 and Table 14, it is observed that there is no statistical significance in the sample selection before classification. It is concluded that the Euclidean distance measure effectively selects the training samples resulting in correct classification.

**Table 15**
*p*-value comparison of GNRBA with other methods using Friedman test on WBCD.

| | Training-testing partition ratio | *p* – value | | |
|---|---|---|---|---|
| | | BPNN | Koza's model | GONN |
| GNRBA | 50–50 | 0.001 | 0.001 | 0.0273 |
| | 60–40 | 0.001 | 0.001 | 0.0273 |
| | 70–30 | 0.001 | 0.001 | 0.0273 |
| | 10-fold cross validation | 0.001 | 0.001 | 0.0273 |

To show the statistical significance among classification methods, the Friedman test is also performed on WBCD only. Table 15, presents the *p* – values on all training – testing partition with a significance level of 0.05 between the GNRBA and the other methods. A higher statistical difference is observed by comparing the GNRBA with BPNN and the Koza's model. Although it performs slightly better as compared to GONN, the GNRBA in general outperforms all other methods, as confirmed in Table 15.

To justify the robustness of the GNRBA, a systematic evaluation is investigated against noisy data. As far as noise in classification is concerned, it is of two types: attribute noise and class noise. From the study it is found that, the two datasets (WBCD and WDBC) do not have any class imbalance (Nettleton, Orriols-Puig, & Fornells, 2010). In this paper, we have considered only attribute noise to show the performance of the GNRBA.

Table 16 presents the CA obtained by the GNRBA with the noisy data on both the datasets. Specifically, for each dataset, the experiment is carried out using training data with (1) 10% added attribute noise (TR10) and (2) 50% added attribute noise (TR50). It is also repeated using test data with (1) 10% added attribute noise and (2) 50% added attribute noise, as suggested in (Nettleton et al., 2010; Zhu & Wu, 2004). In addition, we have calculated the mean CA to measure the overall performance of GNRBA to a particular training – testing partition ratio against noise.

Table 17 presents the comparison of the proposed GNRBA with other techniques reported in the literature for the breast cancer classification problem using WBCD. In this comparison different methods based on feature selection, fuzzy logic, neural network, hybrid methods based on evolutionary techniques etc. are investigated. In most of the methods, the authors have not mentioned details about the result, such as classification accuracy achieved by

**Table 16**
CA (%) of GNRBA in noisy data on WBCD and WDBC.

| | Training-testing partition ratio | Noise in train | | Noise in test | | Noise in train and test | | Mean CA |
|---|---|---|---|---|---|---|---|---|
| | | TR10 TS0* | TR50 TS0 | TR0# TS10 | TR0 TS50 | TR10 TS10 | TR50 TS50 | |
| WBCD | 50–50 | 98.57 | 95.68 | 98.03 | 91.43 | 98.66 | 96.99 | 96.56 |
| | 60–40 | 99.71 | 95.78 | 98.74 | 91.92 | 98.99 | 97.13 | 97.05 |
| | 70–30 | 99.27 | 96.37 | 99.11 | 94.97 | 100 | 97.93 | **97.94** |
| | 10–fold cross validation | 98.55 | 95.87 | 99.75 | 88.08 | 99.58 | 97.33 | 96.53 |
| WDBC | 50–50 | 95.26 | 81.39 | 66.49 | 65.77 | 87.82 | 69.84 | 77.76 |
| | 60–40 | 95.57 | 88.3 | 70.81 | 67.79 | 90.31 | 83.46 | 82.71 |
| | 70–30 | 97.35 | 90.4 | 72.36 | 72.13 | 91.81 | 84.2 | **84.71** |
| | 10 – fold cross validation | 96.27 | 83.22 | 71.79 | 70.15 | 95.24 | 85.17 | 83.64 |

# indicates 0% noise in training data
* indicates 0% noise in test data

**Table 17**
Comparison of classification accuracy with breast cancer classification methods on WBCD.

| Name of the method | CA (%) | Reference |
|---|---|---|
| C4.5 | 94.74 | Quinlan (1996) |
| RAIC | 95.00 | Hamilton et al. (1996) |
| Neuro-fuzzy | 95.06 | Nauck and Kruse (1999) |
| Fuzzy-GA | 97.36 | Pena-Reyes and Sipper (1999) |
| LSA machine | 98.80 | Albrecht, Lappas, Vinterbo, Wong, and Ohno-Machado (2002) |
| Supervised fuzzy clustering | 95.57 | Abonyi and Szeifert (2003) |
| Fuzzy-AIRS | 98.51 | Polat and Güneş (2007a) |
| SVM | 99.54 | Übeyli (2007) |
| LS-SVM | 98.53 | Polat and Güneş (2007b) |
| CFW | 99.50 | Peng et al. (2010) |
| Real coded GA | 96.5 | Örkcü and Bal (2011) |
| AMMLP | 99.26 | Marcano-Cedeño et al. (2011) |
| Decision tree algorithm | 92.97 | Lavanya and Rani (2011) |
| RS_SVM[a] | 96.55 | Chen et al. (2011) |
| RS_SVM[c] | 96.72 | Chen et al. (2011) |
| ICA | 97.75 | Malmir et al. (2013) |
| RF-ANN | 98.05 | Koyuncu and Ceylan (2013) |
| PSO (4 – 2) | 94.74 | Xue et al. (2014) |
| GONN[a] | 97.73 | Bhardwaj and Tiwari (2015) |
| GONN[b] | 99.11 | Bhardwaj and Tiwari (2015) |
| GONN[c] | 99.21 | Bhardwaj and Tiwari (2015) |
| GONN[d] | 99.26 | Bhardwaj and Tiwari (2015) |
| **GNRBA[a]** | **98.54** | |
| **GNRBA[b]** | **99.27** | |
| **GNRBA[c]** | **100** | |
| **GNRBA[d]** | **99.23** | |

[a] Result for 50–50 training-testing partition.
[b] Result for 60–40 training-testing partition.
[c] Result for 70–30 training-testing partition.
[d] Result for 10-fold cross validation.

them is the maximum or mean value and the type of training-testing ratio used to partition their data. It is observed from the Table 17 that the GNRBA is better as compared to other methods. Our improvised results are due to the consideration of a subset of significant training samples. In addition, the optimal values of weighting coefficients are used to calculate the class contribution ability of each class to represent test sample.

In many classification methods, the computational time must be considered a key performance index to decide their real-time application. We have calculated the response time of different methods, as shown in Table 18. To the best of our knowledge, in the literature response time is one of the measures used in this context (Zheng et al., 2014). From Table 18, it is observed that the proposed GNRBA performs breast cancer classification with minimum response time. To measure the computational complexity of the GNRBA, an asymptotic measure like Big O notation is used. The computational complexity of the proposed GNRBA is bounded by $O(K)$, where $K$ is the number of training samples selected using the Euclidean distance. From the study it is found that, the computational complexity of ANN completely depends on its architecture. For a polynomial neural network, the computational complexity is $O(nlog(n))$, where $n$ is the number of operations. Neural network based classification methods such as BPNN, Koza's model and GONN, the computational complexity can be $O(n^2)$, (Orponen, 1994; Williams, & Zipser, 1995). From the above discussion, we can conclude that the proposed GNRBA outperforms all other methods in terms of minimum computational time as well as minimum computational complexity.

## 5. Conclusion

In this paper, a novel expert system (GNRBA) is introduced for breast cancer classification. The key to our success is the use of Gauss-Newton learning to find the optimal weighting coefficients of the significant training samples. The benefit of our research work is that the computational complexity and response time are reduced as compared to the conventional sparse representation methods. In addition, experiments on different datasets (WBCD and WDBC) demonstrate that GNRBA outperforms the existing models. It is observed that the proposed method achieved the highest classification accuracies of 98.54% for 50–50 partition, 99.27% for 60–40 partition and 100% for 70–30 partition for a sub-

**Table 18**
Response time of different methods.

| Method | Dataset | Training-testing ratio | Response time (seconds) |
|---|---|---|---|
| BPNN | WBCD | 50–50 | 3.7922 |
| | | 60–40 | 4.6146 |
| | | 70–30 | 0.3492 |
| | | 10-fold cross validation | 1.9668 |
| Koza's model | WBCD | 50–50 | 3.6449 |
| | | 60–40 | 4.4353 |
| | | 70–30 | 0.3356 |
| | | 10-fold cross validation | 1.8904 |
| GONN | WBCD | 50–50 | 3.4859 |
| | | 60–40 | 4.2419 |
| | | 70–30 | 0.3210 |
| | | 10-fold cross validation | 1.8080 |
| GNRBA | WBCD | 50–50 (K = 74) | **2.9003** |
| | | 60–40 (K = 92) | **3.5293** |
| | | 70–30 (K = 2) | **0.2670** |
| | | 10-fold cross validation (K = 2) | **1.5042** |
| | WDBC | 50–50 (K = 160) | **5.1786** |
| | | 60–40 (K = 54) | **1.7889** |
| | | 70–30 (K = 43) | **1.0952** |
| | | 10-fold cross validation (K = 43) | **3.4367** |

set (WBCD) that contains 74, 92 and 02 training samples, respectively. Results also show the superiority of the proposed method in terms of sensitivity, specificity, confusion matrix, statistical test and AUC. Further, to show the significance difference in the selected subset, the Friedman statistical test is carried out before classification. It is evident that no significance difference between the selected subset for correct classification. The robustness of the proposed method to noisy data has been verified. Additionally, an experiment is also conducted using WDBC. From the results, it is seen that the proposed algorithm achieved a highest classification accuracy of 97.54%, 98.25%. 98.86% and 98.46% for 50–50 partition, 60–40 partition, 70–30 partition and 10 – fold cross validation, respectively. We believe that the promising results (with WDBC) demonstrated by the proposed method will certainly serve as a reference data to carry out further research in this direction. Till date, in the breast cancer classification problem, tumor samples are considered for two classes only i.e. benign or malignant. In the future, a pre-malignant class may be introduced for early detection of breast cancer for effective treatment.

## References


Abdel-Nasser, M., Melendez, J., Moreno, A., Omer, O. A., & Puig, D. (2017). Breast tumor classification in ultrasound images using texture analysis and super-resolution methods. *Engineering Applications of Artificial Intelligence, 59*, 84–92.

Abdel-Zaher, A. M., & Eldeib, A. M. (2016). Breast cancer classification using deep belief networks. *Expert Systems with Applications, 46*, 139–144.

Abonyi, J., & Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Letters, 24*(14), 2195–2207.

Acharya, U. R., Ng, W. L., Rahmat, K., Sudarshan, V. K., Koh, J. E. W., Tan, J. H., Hagiwara, Y., Yeong, C. H., & Ng, K. H. (2017). Data mining framework for breast lesion classification in shear wave ultrasound: A hybrid feature paradigm. *Biomedical Signal Processing and Control, 33*, 400–410.

Albrecht, A. A., Lappas, G., Vinterbo, S. A., Wong, C., & Ohno-Machado, L. (2002). Two applications of the LSA machine. In *Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on: 1* (pp. 184–189). IEEE.

Bache, K., & Lichman, M.. *Retrieved from UCI machine learning repository* http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/.

Bennett, K. P., & Mangasarian, O. L. (1992). Robust linear programming discrimination of two linearly inseparable sets. *Optimization methods and software, 1*(1), 23–34.

Bhardwaj, A., & Tiwari, A. (2015). Breast cancer diagnosis using genetically optimized neural network model. *Expert Systems with Applications, 42*(10), 4611–4620.

Chen, C. H. (2014). A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection. *Applied Soft Computing, 20*, 4–14.

Chen, H. L., Yang, B., Liu, J., & Liu, D. Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications, 38*(7), 9014–9022.

Christoyianni, I., Dermatas, E., & Kokkinakis, G. (2000). Fast detection of masses in computer-aided mammography. *IEEE Signal Processing Magazine, 17*(1), 54–64.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research, 7*, 1–30 Jan.

Elmore, J. G., Wells, C. K., Lee, C. H., Howard, D. H., & Feinstein, A. R. (1994). Variability in radiologists' interpretations of mammograms. *New England Journal of Medicine, 331*(22), 1493–1499.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861–874.

Garcia, S., & Herrera, F. (2008). An extension on"statistical comparisons of classifiers over multiple data sets"for all pairwise comparisons. *Journal of Machine Learning Research, 9*, 2677–2694 Dec.

Gill, P. E., Murray, W., & Wright, M. H. (1981). *Practical optimization*. University of Michigan, USA: Academic Press.

Hagan, M. T., Demuth, H. B., Beale, M. H., & De Jesús, O. (1996). *Neural network design*: 20. Boston: PWS publishing company.

Hamilton, H. J., Shan, N., & Cercone, N. (1996). *RIAC: A rule induction algorithm based on approximate classification*. Computer Science Department, University of Regina.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell, 144*(5), 646–674.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer Science & Business Media.

Jang, J. S. R., Sun, C. T., & Mizutani, E. (1997). *Neuro-fuzzy and soft computing; a computational approach to learning and machine intelligence*. Prentice-Hall.

Koyuncu, H., & Ceylan, R. (2013). Artificial neural network based on rotation forest for biomedical pattern classification. In *Telecommunications and Signal Processing (TSP), 2013 36th International Conference on* (pp. 581–585). IEEE.

Koza, J. R., & Rice, J. P. (1991). Genetic generation of both the weights and architecture for a neural network. In *Neural Networks, 1991., IJCNN-91-Seattle International Joint Conference on: 2* (pp. 397–404). IEEE.

Lavanya, D., & Rani, D. K. U. (2011). Analysis of feature selection with classification: Breast cancer datasets. *Indian Journal of Computer Science and Engineering (IJCSE), 2*(5), 756–763.

Magna, G., Casti, P., Jayaraman, S. V., Salmeri, M., Mencattini, A., & Martinelli, E. (2016). Identification of mammography anomalies for breast cancer detection by an ensemble of classification models based on artificial immune system. *Knowledge-Based Systems, 101*, 60–70.

Mahmoudi, M., & Sapiro, G. (2012). Sparse representations for range data restoration. *IEEE transactions on image processing: A publication of the IEEE Signal Processing Society, 21*(5), 2909–2915.

Malmir, H., Farokhi, F., & Sabbaghi-Nadooshan, R. (2013). Optimization of data mining with evolutionary algorithms for cloud computing application. In *Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on* (pp. 343–347). IEEE.

Mangasarian, O. L., & Wolberg, W. H. (1990). Cancer diagnosis via linear programming. *SIAM News, 23*(5) pp 1 & 18.

Mangasarian, O. L., Setiono, R., & Wolberg, W. H. (1990). Pattern recognition via linear programming: Theory and application to medical diagnosis. *Large-scale numerical optimization* (pp. 22–31). Philadelphia: Siam Publications.

Marcano-Cedeño, A., Quintanilla-Domínguez, J., & Andina, D. (2011). WBCD breast cancer database classification applying artificial metaplasticity neural network. *Expert Systems with Applications, 38*(8), 9573–9579.

McPherson, K., Steel, C., & Dixon, J. M. (2000). Breast cancer-epidemiology, risk factors, and genetics. *BMJ: British Medical Journal, 321*(7261), 624.

Mei, X., Ling, H., & Jacobs, D. W. (2011). Illumination recovery from image with cast shadows via sparse representation. *IEEE Transactions on Image Processing, 20*(8), 2366–2377.

Moon, W. K., Chen, I. L., Chang, J. M., Shin, S. U., Lo, C. M., & Chang, R. F. (2016). The Adaptive Computer-aided Diagnosis System based on Tumor Sizes for the Classification of Breast Tumors detected at Screening Ultrasound. *Ultrasonics, 76*, 70–77.

Muto, T., Bussey, H. J. R., & Morson, B. C. (1975). The evolution of cancer of the colon and rectum. *Cancer, 36*(6), 2251–2270.

Nauck, D., & Kruse, R. (1999). Obtaining interpretable fuzzy classification rules from medical data. *Artificial intelligence in medicine, 16*(2), 149–169.

Nettleton, D. F., Orriols-Puig, A., & Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial intelligence review, 33*(4), 275–306.

Nilashi, M., Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). A knowledge-based system for breast cancer classification using fuzzy logic method. *Telematics and Informatics, 34*(4), 133–144.

Örkcü, H. H., & Bal, H. (2011). Comparing performances of backpropagation and genetic algorithms in the data classification. *Expert systems with applications, 38*(4), 3703–3709.

Orponen, P. (1994). Computational complexity of neural networks: A survey. *Nordic Journal of Computing, 1*(1), 94–110.

Pena-Reyes, C. A., & Sipper, M. (1999). A fuzzy-genetic approach to breast cancer diagnosis. *Artificial intelligence in medicine, 17*(2), 131–155.

Peng, Y., Wu, Z., & Jiang, J. (2010). A novel feature selection approach for biomedical data classification. *Journal of Biomedical Informatics, 43*(1), 15–23.

Polat, K., & Güneş, S. (2007a). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing, 17*(4), 694–701.

Polat, K., & Güneş, S. (2007b). A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS. *Computer methods and programs in biomedicine, 88*(2), 164–174.

Prasad, Y., Biswas, K. K., & Jain, C. K. (2010). SVM classifier based feature selection using GA, ACO and PSO for siRNA design. In *International conference in swarm intelligence* (pp. 307–314). Berlin Heidelberg: Springer.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4. 5. *Journal of Artificial Intelligence Research, 4*, 77–90.

Rodrigues, P. S., Chang, R. F., & Suri, J. S. (2006). Non-extensive entropy for cad systems of breast cancer images. In *2006 19th Brazilian Symposium on Computer Graphics and Image Processing* (pp. 121–128). IEEE.

Şahan, S., Polat, K., Kodaz, H., & Güneş, S. (2007). A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. *Computers in Biology and Medicine, 37*(3), 415–423.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437.

Übeyli, E. D. (2007). Implementing automated diagnostic systems for breast cancer detection. *Expert systems with Applications, 33*(4), 1054–1062.

Wang, P., Hu, X., Li, Y., Liu, Q., & Zhu, X. (2016). Automatic cell nuclei segmentation and classification of breast cancer histopathology images. *Signal Processing, 122*, 1–13.

Williams, R. J., & Zipser, D. (1995). Gradient-based learning algorithms for recurrent networks and their computational complexity. In *Backpropagation: Theory, Architectures, and Applications: 1* (pp. 433–486).

Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences, 87*(23), 9193–9196.

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995). Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analytical and Quantitative cytology and histology, 17*(2), 77–87.

Xu, Y., Zhang, D., Yang, J., & Yang, J. Y. (2011). A two-phase test sample sparse representation method for use with face recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 21*(9), 1255–1262.

Xue, B., Zhang, M., & Browne, W. N. (2014). Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. *Applied Soft Computing, 18*, 261–276.

Yuan, X. T., Liu, X., & Yan, S. (2012). Visual classification with multitask joint sparse representation. *IEEE Transactions on Image Processing, 21*(10), 4349–4360.

Zhang, L., Yang, M., & Feng, X. (2011). Sparse representation or collaborative representation: Which helps face recognition? In *2011 International Conference on Computer Vision* (pp. 471–478). IEEE.

Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications, 41*(4), 1476–1482.

Zhu, X., & Wu, X. (2004). Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review, 22*(3), 177–210.