Int. J. Nonlinear Anal. Appl. 15 (2024) 1, 17–30
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2023.27818.3726

Study and evaluation of feature vector optimization and classic methods in automatic breast cancer detection

Roozbeh Rahmani, Shahin Akbarpour*, Ali Farzan

Department of Computer Engineering, Shabestar Branch, Islamic Azad University, Shabestar, Iran

(Communicated by Javad Vahidi)

Abstract

Breast cancer is known to be among the most prevalent cause of mortality among women. Since early breast cancer diagnosis increases survival chances, the development of a system with a highly accurate output to detect suspicious masses in mammographic images is of great significance. Thus, many studies have focused on the development of methods with favorable performance and acceptable accuracy to detect cancerous masses, proposed various techniques to diagnose breast cancer, and compared their accuracies. Most previous studies have used composite selection and feature reduction techniques to detect breast cancer and accelerate its treatment; however, most have failed to reach the desired accuracy due to the selection of ineffective features and the lack of a proper analytical method for the features. The present study reviews the methods proposed to detect breast cancer so far and analyzes the process of feature vector optimization techniques as well as the normal/abnormal and benign/malignant mass classification.

Keywords: Breast cancer detection, Feature extraction, Classification, Mammographic images
2020 MSC: 68Txx

1 Introduction

Breast cancer is the most prevalent cancer among women and the fifth leading cause of mortality due to cancer. Early and an accurate breast cancer diagnosis can keep the patients alive for a prolonged period. Despite the increase in the prevalence of this disease, statistics suggest a decline in the mortality rates associated with it. This could be due to the new therapeutic methods and diagnostic techniques such as mammography systems. Mammographic images can be used to detect various abnormalities such as breast cancer. Similar to other medical images, mammographic images have specific features that make them difficult to interpret and reduce their performance in distinguishing between malignant and benign masses. Moreover, the detection of this type of cancer is also difficult due to the presence of small cancerous patents in the whole image. Many studies have been conducted on mammographic images over the recent years to detect cancerous masses without diagnostician intervention to reduce the errors due to carelessness, personal mistakes, and fatigue (1, 2). Various features have been presented to define breast masses [30, 22]. The performance of each feature is associated with its ability to detect masses from various classes. The feature space might contain a large number of unfavorable items taking up large storage space and reducing the classification accuracy. Thus, a method needs to be proposed to improve the detection accuracy as well as the extraction of more effective features

*Corresponding author
Email addresses: roozbeh_ra75@yahoo.com (Roozbeh Rahmani), sh.akbarpour@gmail.com (Shahin Akbarpour), alifarzanam@gmail.com (Ali Farzan)

Received: July 2022 *Accepted:* January 2023

[27]. Therefore, it would appear that an accurate system capable of extracting effective features for the early detection of benign and malignant breast tumors is necessary. The present study seeks to investigate feature vector optimization and classic methods in early breast cancer detection using mammographic images. Figure 1 demonstrates the general diagram block of the process of breast cancer detection. In the detection process, mammographic images from each sample are categorized into either the normal/abnormal or benign/malignant classes. The main stages investigated in the present study include the extraction of the areas of interest, extraction of the effective features, feature vector generation, and implementation of the classifiers.

2 Data collection

Using a standard image database is imperative to investigate the improvement of breast cancer detection accuracy. Various databases are thus used in detecting breast cancer. The mammographic images of each patient are available from various angles in each of these databases some of which are reviewed in the following.

2.1 The Wisconsin Diagnostic Breast Cancer (WDBC) database

The WDBC breast cancer database is available on the UCU website. This database contains 699 samples with 32 features, among which one is the ID number and another is the class label determining the type of the sample (malignant or benign). The other 30 features include mean and standard deviations and a maximum of 10 features including diameter, tissue, circumference, area, smoothness, concentration, concavity, concavity points, symmetry, and fractal dimensions. The dataset contains a total of 32 features in 10 categories. The three indicators of mean, standard deviation, and maximum are measured for each category [46].

2.2 The UC-Irvine database

This dataset [34] includes the risk factors of bulk thickness, cell shape uniformity, cell size uniformity, edge adhesion, naked nuclei, epithelial tissue cell volume, bland chromatin, normal nucleus, and cell division, and contains data collected in Wisconsin, USA.

2.3 The MIAS database

The Mammographic Image Analysis Association (MIAS) database contains 322 mammographic images of the left and right breasts of 61 various ladies. The images were 1024 by 1024 in terms of dimensions and were digitized with a micron pixel edge 200 resolution. Digital mammographic images are grayscale images with a depth of eight bits. These images are asymmetrical and structurally distorted in terms of damage, containing normal breasts, containing masses, and containing micro-classification clusters. The database contains 209 normal breasts, 67 ROIs with benign masses, and 54 ROIs with malignant masses [37, 47].

2.4 The DDSM database

The complete version of the database contains around 2,500 mammographic images classified into the three groups of normal, benign, and malignant. The normal group contains mammographic images from patients with no mass observed in their breasts, the benign group contains mammographic images of patients with benign masses in their breasts, and the malignant group contains mammographic images from patients whose breast masses were diagnosed to be malignant [48].

3 Extracting the area of interest

The area of interest (the area in which the tumor is observed) is extracted from the mammographic images before the feature extraction process. Most databases contain information on the presence or absence of masses from the images in addition to the mammographic images themselves. Thus, feature extraction is conducted on the area of interest in the following.

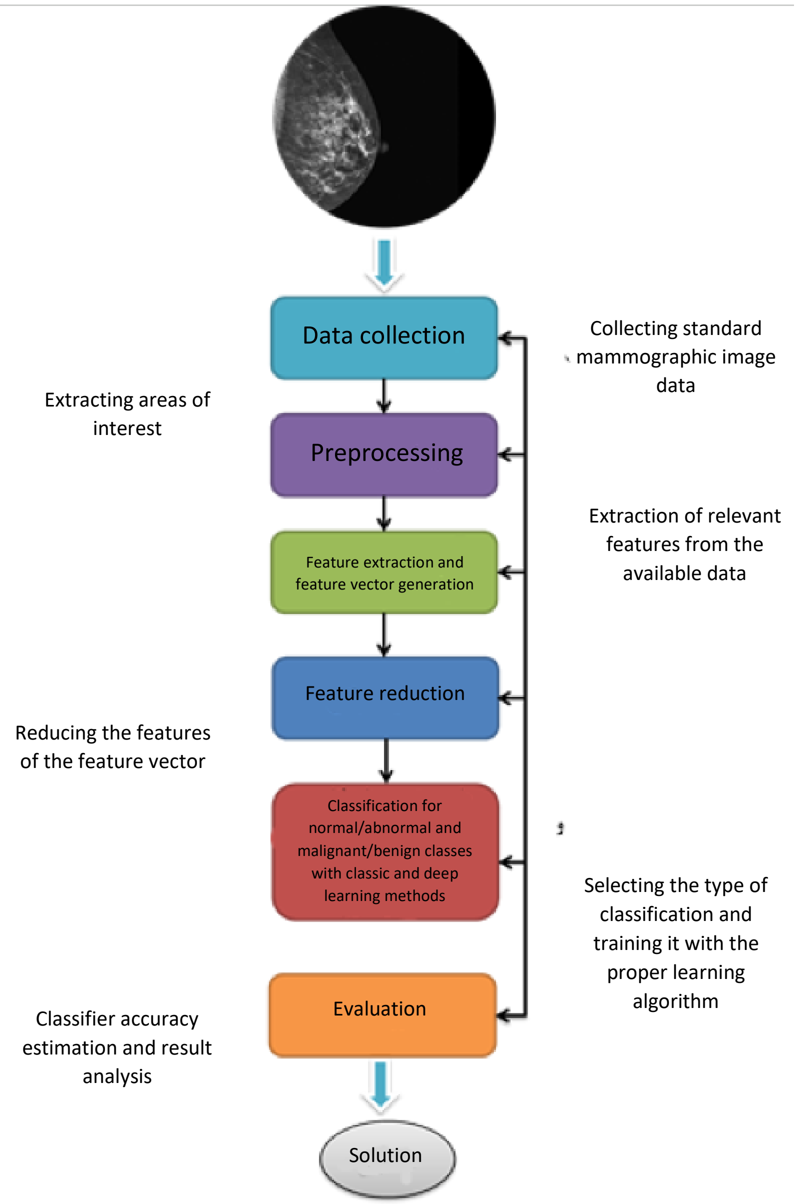

Figure 1: the general scheme of the automated breast cancer detection process

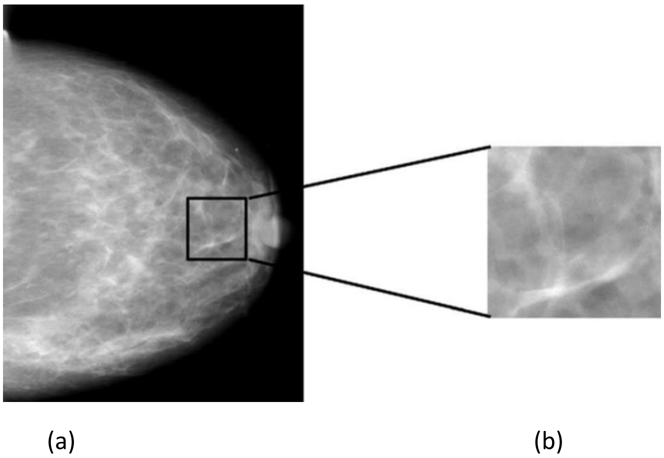

Figure 2: Sample images from the database extracted with dimensions of 200 by 200 (a: the main image, b: the ROI image)

4 Feature vector generation and extraction process

Feature extraction is a process in which the effective and determinant features of the data are determined through a set of operations. The features that are capable of distinguishing patterns are determined at this stage. Thus, the features are the specifications of the objects used as the input of the classifiers and make up various classes. The feature of an object is in fact what distinguished one input pattern from the other. Some of the features extracted from mammographic images are mentioned as follows [10, 7]:

$$\text{Contrast} = \sum_{1}^{i}\sum_{1}^{j} |i-j|^2 p(i,j) \tag{4.1}$$

The contrast feature is a criterion of diversity and spatial difference of an image. i and j are the indices of the image pixel and $P(I,j)$ is a random matrix.

$$\text{Homogeneity} = \sum_{i}\sum_{j} \frac{p(i,j)}{1+(i-j)^2} \tag{4.2}$$

Homogeneity determines the closeness of the distribution of matrix elements compared to the matrix diameter. In the equation above, i and j are the coordinates of the horizontal and vertical pixels, and P is the value of the pixel.

$$\text{skewness} = \frac{\mu_3}{(\mu_2)^{3/2}} \tag{4.3}$$

Skewness indicates cancerous masses with abnormal cavities and bumps. This feature demonstrates grade I cavities and lumps. Mean μ demonstrates the estimation of the location where clustering occurs.

$$\text{kurtosis} = \frac{\mu_4}{(\mu_2)^2} - 3 \tag{4.4}$$

Kurtosis indicates cancerous masses with abnormal cavities and bumps. This feature demonstrates grade II cavities and lumps.

$$\text{histogramvariance} = \frac{\sum (X_i - \bar{X}_l)^2}{N} \tag{4.5}$$

Variance is another index measuring the data dispersion from mean-variance. $(X_i - \bar{X}_l)^2$ is the squared distance of the data from the mean, and N is the number of pieces of data.

$$\text{entropy} = \sum_{1}^{i}\sum_{1}^{j} C(i,j)\log C(i,j) \tag{4.6}$$

Cancerous masses have different information from normal masses, which is revealed by this feature. i and j stand for the indices of the image pixel.

$$\text{inertia} = \sum_{1}^{i}\sum_{1}^{j}(i-j)^2 C(i,j) \tag{4.7}$$

Cancerous masses have elongations and a continuum of light which is revealed by this feature. i and j stand for the indices of the image pixel.

A feature vector including as many rows as the images available in the dataset and as many columns as the extracted features are generated based on the obtained features. After the feature vector is obtained, the more effective features can be selected through dimensionality reduction techniques.

5 Feature dimensionality reduction

Feature selection requires a large space for inquiry to select a proper subset of the features based on one or several quality criteria without any conversions. A better subset would have a higher ability in expressing the specifications of input data and predicting new samples. The main goal of feature selection is the selection of the best subset containing the relevant and non-redundant features. There are various feature dimensionality reduction methods, among which the most popular in breast cancer detection studies ate PCA and t-test.

5.1 Feature dimensionality reduction through t-test

Equation (5.1) is considered the best feature selection for the t-test method. The equation is applied to each column of the feature table and the t value is obtained. The sum of the values from normal images is first calculated and the respective mean is obtained. The obtained value is m_{ij}. The sum of the values from abnormal images is also calculated and the respective mean is obtained. The obtained value is m_{ik}. s_{ij}^2 is the standard deviation of zero values or normal images, s_{ik}^2 is the standard deviation of values that are equal to zero or one, N_j is the number of normal/abnormal classes, and N_k is the number of benign/malignant classes. A figure indicating the value of each feature is obtained according to the aforementioned. The same is repeated for the other columns of the table so that the value of every feature is determined.

$$t = \frac{m_{ij} - m_{ik}}{\sqrt{(s_{ij}^2/N_j) + (s_{ik}^2/N_k)}} \tag{5.1}$$

$$df = \frac{[s_{ij}/N_j + s_{ik}^2/N_k]^2}{\frac{(s_{ij}^2/N_j)^2}{N_j} + \frac{(s_{ik}^2/N_k)^2}{N_k}} - 2 \tag{5.2}$$

5.2 Feature dimensionality reduction through the PCA technique

The PCA technique is the best way for linear data dimensionality reduction. This technique eliminates the less significant coefficients obtained from the diversion and thus has less missing information compared to the other techniques. In this technique, new axes of coordinates are defined for the data, based on which they are expressed. The first axis must be placed in a direction that maximizes data variance. The second axis must be perpendicular to the first so that the data variance is maximized. All the next axes are perpendicular to the previous axis in such a way that the data have the highest scatter in that direction.

6 Classification

After the feature vector is created and dimensionality is reduced, the final vector is considered as the classifier input. Most studies in the field of breast cancer detection using supervised classification. The input and output are specified in this type of classification, and there is a so-called supervisor that provides the learner with information. Thus, the system tries to learn a function from the input to the output. Figure 3 demonstrates some of the supervised classification algorithms used in breast cancer detection.

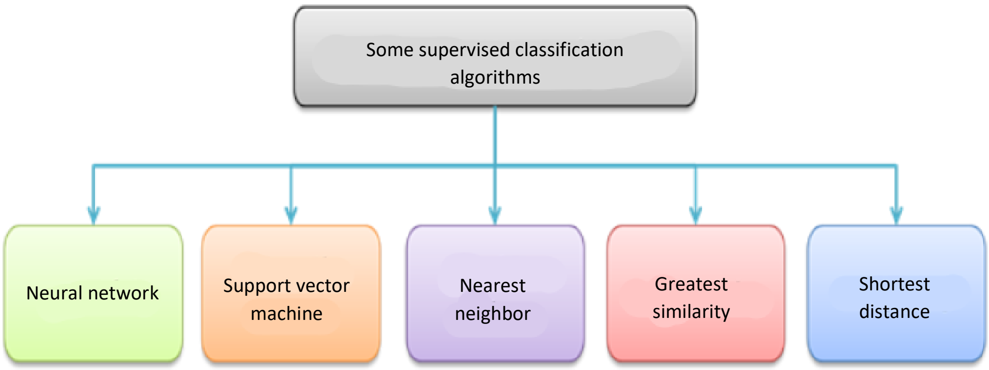

Figure 3: Supervised classification algorithms

Table 1: The advantages and disadvantages of the KNN, NB, and SVM classifier algorithms

Classifier	Features	Limitations
KNN	Classes are not linearly separable. No cost to the learning process. Suitable for multi-purpose classes.	Findings the nearest neighbor can take too long in large training data Sensitive to irrelevant or noisy features Algorithm performance depends on the number of dimensions used.
NB	Easy to implement Excellent computational efficiency and classification rate	Reduced algorithm accuracy in smaller datasets
SVM	High accuracy Works well when the data are not linearly separable in the property space	Needs a larger size and greater speed in both train and test sets High complexity and extensive memory requirements for classification in many cases

7 Analysis and evaluation of classic methods' performance

The correctness of a test –particularly in breast cancer detection- is expressed through the three main indices of sensitivity, specificity, and accuracy. Indices such as sensitivity, specificity, accuracy, PPV, and NPV were used to evaluate the proposed method after implementation.

TP (True Positive): indicates the number of correct predictions for the current class

TN (True Negative): indicates the number of correct predictions for another class

FP (False Positive): indicates the number of incorrect predictions for the current class

FN (False Negative): indicates the number of incorrect predictions for another class

In addition to the indices above, two other criteria called Positive Prediction Value (PPV) and Negative Prediction Value (NPV) were also used to evaluate the results.

The calculation of the evaluation criteria for the normal/abnormal class:

```
Accuracy NA = (NA_TP+NA_TN)/(NA_TP+NA_TN+NA_FP+NA_FN);
Sensitivity NA = NA_TP/(NA_TP+NA_FN);
Specificity NA = NA_TN/(NA_FP+NA_TN);
PPV_NA = NA_TP/(NA_TP+NA_FP);
NPV_NA = NA_TN/(NA_FN+NA_TN)
```

The calculation of the evaluation criteria for the benign/malignant class:

```
Accuracy MB = (MB_TP+MB_TN)/(MB_TP+MB_TN+MB_FP+MB_FN);
```

```
Sensitivity MB = MB_TP/(MB_TP+MB_FN);
Specificity MB = MB_TN/(MB_FP+MB_TN);
PPV_MB = MB_TP/(MB_TP+MB_FP);
NPV_MB = MB_TN/(MB_FN+MB_TN)
```

Tables 2 and 3 demonstrate the results of implementing KNN, NB, and SVM algorithms without feature dimensionality reduction for the normal/abnormal and benign/malignant classes. Results indicated that the SVM algorithm performed the best for normal/abnormal classes and the NB algorithm performed the best for the malignant/benign classes, both of which had significantly better performances compared to KNN in terms of the results of indices.

Table 2: Comparison of the results of SVM, NB, and KNN classifiers for the normal/abnormal classes

Classifier	Accuracy	Sensitivity	Specificity	PPV	NPV
SVM	0.9303	0.8992	0.9479	0.9068	0.9434
NB	0.7107	0.5610	0.7875	0.5750	0.7778
K-NN	0.9212	0.8926	0.9378	0.8926	0.9378

Table 3: Comparison of the results of SVM, NB, and KNN classifiers for the benign/malignant classes

Classifier	Accuracy	Sensitivity	Specificity	PPV	NPV
SVM	0.4050	0.4815	0.3433	0.3714	0.4510
NB	0.6860	0.8061	0.1739	0.8061	0.1737
K-NN	0.5537	1	0	0.5537	0

Table 4 compares some of the proposed methods for breast cancer detection. As the table demonstrates, most methods have used support vector machines to classify the benign and malignant tumor classes.

Table 4: The detection accuracy of some classic methods using various databases

Reference	Technique	Database	Estimated accuracy
[18]	Simple Bayesian	Wisconsin	98%
[40]	Structure support vector machine	DDSM	91%
[15]	Support vector machine	UCI	93%
[32]	Feature selection	Wisconsin	69%
[33]	Two-stage SVM	UCI-WBC	99%
[3]	Bayesian network and support vector machine	Chicago	74% and 67%
[35]	Comparative study	-	Highest accuracy for SVM at 97%
[23]	Kernel	UCI	96%
[42]	Group algorithm based on a support vector machine	Wisconsin	94%
[45]	Combination algorithm of K-means and support vector machine	WDBC	97%
[25]	SVM and KNN	DDSM	96%
[4]	Predictive algorithm	SEER	77%

8 Comparative study of some classic methods with feature vector optimizer

Many studies have been conducted over the recent year seeking to reduce the error of breast cancer detection and increase its accuracy through various techniques, some of which are reviewed in the following.

Xiao et al. [44] proposed a novel unsupervised feature extraction method based on deep learning with a support vector machine model to detect breast cancer in 2018. Their proposed method comprises the use of a support vector machine to classify the samples with a new feature for benign and malignant tumors. Salama et al. [36] proposed a computer detection system to detect breast cancer in digital mammography in 2018. They used an improved method to extract the features based on a contourlet transform to obtain the features of the areas of interest which could improve accuracy compared to other methods. They also proposed a composite method based on a support vector machine and genetic algorithm for feature dimensionality reduction.

Liu et al. [26] proposed a Bayesian model to explore the potential correlation between the cancer data features in 2018. They also used a learning algorithm and statistical computational method to build and evaluate the Bayesian method. The data they used was collected from a clinical sonogram dataset from a local Chinese hospital and needle aspiration cytology from the machine learning database.

Wang et al. [41] proposed a mammographic screening strategy compatible with breast cancer risk in 2018. Their proposed strategy comprised of the two stages of estimating the breast cancer risk based on age and deciding on the healthy mammography screening based on the estimated risk. Results indicated that an optimal combination of the independent variable used in risk estimation was not the same across various age groups. The optimized decision-making in this strategy was the mammography screening decision in the case of losing the better mean life expectancy.

Abdar proposed a novel data mining technique considering artificial neural networks and support vector machines for breast cancer detection in 2019. The proposed method mainly sought to develop an automated expert system for breast cancer detection. This study first used the support vector machine with various values for the parameters and then introduced a new breast cancer detection method using the two techniques of collective learning, the weighted voting approach, and the boosting technique.

Liu et al. [24] proposed an intelligent breast cancer detection approach in 2019. The proposed method first used a genetic algorithm and simulated annealing for feature selection and ranked the features, and proceeded to use the support vector machine to extract the optimal features. Not only did the feature selection approach proposed in this study help reduce complexity and extract the optimized features, but it also obtained the highest classification accuracy and the lowest classification costs.

Qui et al. [34] proposed an automated breast cancer detection model for ultrasonography images in 2019. The conventional ultrasonic image analysis methods use manipulated features to classify images, and the inability to change the size, shape, and tissue of breast masses results in the low sensitivity of the clinical application programs. This study proposed a method to detect ultrasonic breast images using deep convolution neural networks with multiscale kernels and jump joints to overcome these deficiencies.

Pramanik et al. [31] proposed a new framework for early breast cancer detection in 2019. Their proposed framework included two stages. The first stage was to segment the suspected areas automatically, and the second stage was to classify the segments into benign and malignant cases. An area-based surface set method was proposed to segment the suspicious areas. This study also used adaptive thresholding to estimate the suspicious areas.

Benzebouchi et al. [8] proposed a convolution neural network method for automatic breast cancer detection in 2019 using segmented data from the digital database for mammographic screening. The study developed a network with convolutional neural network architecture. The proposed method provided better classification rates and yielded more accurate breast cancer detections.

Wang et al. [39] proposed a computer breast cancer detection system based on a convolutional neural network in 2019. The proposed method first conducted a mass detection based on convolutional neural network features and unsupervised clustering and then created a set of features combining the deep, morphological, tissue, and density features. At the final stage, a backpropagation error classification has been used using the set of the composite features to classify breast tumors into benign and malignant masses.

Khan et al. [21] proposed a deep learning framework for the detection and classification of breast cancer using the concept of transfer learning in 2019. Their proposed framework extracted the features from images using pre-trained convolutional neural structures. The tests were conducted on standard datasets to evaluate the performance of the proposed framework.

Alickovic and Subasi [2] proposed a novel model based on a multilayer perceptron neural network to classify breast cancer with high accuracy in 2019. The proposed method WAS TESTED ON THE Wisconsin data set and revealed a classification accuracy of 99%.

Matos et al. [29] proposed a method to detect the benign and malignant patterns of tumors observed in digital mammographic images based on local feature analysis in 2018. This study used scale-invariant feature transform (SIFT) definers to extract the local features, Speeded-Up Robust Features (SURF), extracted features as input for support vector machine classifiers, and adaptive boosting and random forest to distinguish between benign and malignant tumors.

Gherghout et al. [11] used a framework to classify the normal, benign, and malignant tumors in 2019. This study first considered a set of rules to pre-classify the mammographic images based on the created tissue which divided various shapes of the breast based on the abnormality. The key point in this study was the use of the error backpropagation neural network model to demonstrate the tissue and morphological features of the tumors. The Mias database was eventually used to validate the proposed method.

Chaieb and Kalti [9] studied an ideal subset of features to improve tumor classification performance in 2018. The authors first reviewed the various definers that are often used in studying breast cancer and conducted a comparative study between the selected features to test their ability in detecting benign and malignant tumors.

Wang et al. [42] proposed a group learning algorithm to detect breast cancer based on a support vector machine to reduce the detection variance and increase accuracy in 2018. The Wisconsin breast cancer and the research protocols of the National Cancer Institute of the United States were studied to evaluate the performance of the proposed model. Experimental results indicated that the proposed model has higher accuracy and lower significant variance for breast cancer detection compared to the mechanisms of the other groups and two common organizational models of adaptation and mass classification tree.

Kaymak et al. [20] proposed a method for automatic image classification for breast cancer detection. Image classification was conducted through a backpropagation neural network (BPN) in 2019. Backpropagation error neural network and radial basis function networks had accuracies of 59.0% and 70.4%, respectively.

Vijayarajeswari et al. [38] conducted feature extraction and classification using a support vector machine and Hough transform for rapid breast cancer detection in 2016. Their proposed method used the Hough transform to extract certain features from mammographic images. Results of this study indicated that the proposed model effectively classified the abnormal class.

Jitaree et al. [17] studied the classification of breast cancer areas in microscopic images using tissue features in 2016. The authors evaluated the application of two types of classification (neural network and decision tree) in the classification of three regions (cancer, lymphocytes, and stroma) in their study. This study combined tissue features based on energy information and fractal dimension for feature selection.

Karthiga et al. [19] detected breast cancer using curvelet and regional features in 2019. Their study used the feature extraction method using the curvelet transform in digital mammography to detect normal and abnormal breast cancer. Preprocessing is essential to improve the contrast in mammographic images. This study used upper and lower curve transforms. The features (contrast, correlation, homogeneity, and energy) were extracted from the curvelet coefficients using the gray-level surface co-occurrence matrix.

Hussain et al. [16] studied automatic breast cancer detection using machine learning techniques by extracting various feature extracting strategies in 2018. This study used several strategies for feature selection. Moreover, they used the SIFT technique, tissue features, and descriptive features and obtained acceptable final results.

Avinash et al. [24] proposed a rapid breast cancer detection technique using a support vector machine using sequential minimal optimization in 2020. The support vector machine was revealed to have a better performance compared to the other classifiers when tested on the Wisconsin dataset.

Melekoodappattu et al. proposed an automatic breast cancer detection using an extreme machine learning classifier in 2020. This study used the fruit fly optimization algorithm to adjust the input weight to obtain the favorable output in the hidden extreme machine learning node [34].

Assegie proposed a method based on the optimized K-NN algorithm to detect breast cancer in 2021. The proposed method used grid search to find the best K value that could create the highest breast cancer detection accuracy.

HAQ et al. proposed a method to detect breast cancer through clinical data using supervised and unsupervised feature selection techniques in 2021. The proposed method used the supervised technique of the rescue algorithm and the unsupervised technique of Autoencoder, PCA algorithms, to select the relevant features from the dataset.

Table 5: Demonstrates a comparative study of the methods proposed for breast cancer detection mentioned in this section

Technique	Year	Feature extraction	Classification	Detection accuracy	Advantages	Disadvantages	Reference
Deep learning	2018	Based on deep learning	Support vector machine	98%	Excellent detection accuracy improvement	On preprocessing	[44]
Computer-aided detection	2018	Contourlet transform	Support vector machine	97%	The use of preprocessing and feature dimensionality reduction in the proposed method	Low accuracy in detecting benign/malignant classes	[36]
Bayesian modeling	2018	No feature extraction	Naïve Bayes	95%	Increase adaptation and tumor classification tree	No use of standard breast cancer databases	[26]
Adaptive mammographic screening	2018	No feature extraction	-	-	Close study of risk estimation in screening	-	[41]
Collective learning	2019	Based on the features specified in the database	Composite	100%	The accurate distinction between normal and abnormal tissue	No feature dimensionality reduction	[1]
Smart classification	2019	Feature selection using genetic algorithm	Neural network and support vector machine	95%	Reduced complexity and optimal feature selection	Reduced accuracy and sensitivity	[24]
Deep neural network	2019	Deep network	Deep network	98%	Excellent detection accuracy improvement	Computational complexity	[34]
Composite neural network	2019	Statistical moments	Artificial neural network	89%	More effective feature selection	Fewer evaluation criteria	[31]
Deep neural network	2019	Deep network	Deep network	98%	The accurate distinction between normal and abnormal tissue	The computational complexity of the proposed algorithm	[8]
Machine learning	2019	Deep network	Error back-propagation network	86%	Algorithm running speed	Low accuracy in detecting benign/malignant classes	[39]
Deep learning	2019	Deep network	Proposed architecture	98%	Excellent detection accuracy improvement	Computational complexity	[21]
Multilayer perception neural network	2019	No feature extraction	Multilayer perception neural network	99%	High accuracy breast cancer classification	More accurate evaluation criteria have not been mentioned	[2]
Local feature analysis	2018	Feature-scale-independent transform modifiers	Adaptive boosting and support vector machine	99%	The accurate distinction of normal and abnormal tissue	Reduce accuracy and sensitivity	[29]

Breast tissue classification	2019	Correlation matrix	Neural network	98%	Accurate separation and cancerous from healthy breast tissue	Complex and slow due to the use of various algorithms	[11]
Ideal set of features	2018	Correlation matrix	Multilayer perceptron, support vector machine, and KNN	-	More effective feature extraction	Computational complexity	[9]
Support vector machine	2018	-	Improved support vector machine	97%	Algorithm sunning speed and accurate evaluation criteria	Lower significant variance	[42]
Artificial neural network	2017	No feature extraction	Artificial neural network	70%	Low accuracy in detecting the classes	More accurate evaluation criteria have not been mentioned	[20]
Hough transform and support vector machine	2019	Hough transform	support vector machine classifier	94%	The accurate distinction between the classes	More accurate evaluation criteria have not been mentioned	[38]
Tissue features	2016	Energy information and fractal dimension	Neural network and decision tree	-	High accuracy	No feature dimensionality reduction	[17]
Curvelet	2019	Regional features	-	98%	More effective feature extraction	Not considering a classifier to divide the malignant/benign classes	[19]
Machine learning techniques	2018	Descriptive and tissue features	Support vector machine	97%	Optimal feature extraction	No shape feature extraction	[16]
Support vector machine	2020	Based on the features specified in the dataset	Support vector machine	93%	Algorithm running speed	Reduced accuracy	[6]
Extreme machine learning	2020	Extreme machine learning	Particle Swarm Optimization	99%	High detection accuracy	Computational complexity	[28]
Optimized K-NN algorithm	2021	Shape and tissue features	K-NN algorithm	94%	Finding the best K value to increase the classifier's accuracy	Unspecified extraction and selection process	[5]
Machine learning algorithms and clinical data	2021	Principal component analysis	The use of various classifier algorithms	99%	High detection accuracy	No use of various databases for accurate evaluation	[12]

9 Conclusion and future work

Many studies gave been conducted to detect breast cancer over the reason years, but have failed to obtain an adequate accuracy due to the selection of ineffective features and not using an efficient classifier algorithm. The present study reviewed and compared the feature vector optimization and classic methods and analyzed the process of breast cancer detection. Results indicated that selecting more effective features and proper classifier algorithms can improve the accuracy of breast cancer detection. Results of using a support vector machine indicated an accuracy of over 80% through optimizing the obtained features in most studies. Thus, despite the striking progress over the recent years, more work needs to be done to expand the breast cancer detection systems and use precision methods. The use of effective and efficient methods must lead to early disease diagnosis and advanced disease prediction. Thus, future works can focus on the following to increase the accuracy of breast cancer detection:

1) Accurate analysis of the features and extracting more effective features
2) Using algorithms such as linear separators to select the proper features
3) Improving the classifiers through feature purification
4) Improving the classification through various training algorithms

References

[1] M. Abdar and V. Makarenkov, *CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer*, Measurement **146** (2019), 557–570.

[2] E. Alickovic and A. Subasi, *Normalized neural networks for breast cancer classification*, CMBEBIH 2019 Int. Conf. Med. Bio. Engin., 16-18 May 2019, Banja Luka, Bosnia and Herzegovina. Springer International Publishing, 2020.

[3] D.A. Aljawad, E. Alqahtani, A.L.K Ghaidaa, N. Qamhan, N. Alghamdi, S. Alrashed, J. Alhiyafi and S.O. Olatunji, *Breast cancer surgery survivability prediction using Bayesian network and support vector machines*, Int. Conf. Inf. Health Technol. (ICIHT), IEEE, 2017, pp. 1–6.

[4] L.M. Alnemer, L. Rajab and I. Aljarah, *Conformal prediction technique to predict breast cancer survivability*, Int. J. Adv. Sci. Technol. **96** (2016), 1–10.

[5] T.A. Assegie, *An optimized K-Nearest Neighbor based breast cancer detection*, J. Robotics Control **2** (2021), no. 3, 115–118.

[6] K. Avinash, M.B. Bijoy and P.B. Jayaraj, *Early detection of breast cancer using support vector machine with sequential minimal optimization*, Adv. Comput. Intel. Engin.: Proc. ICACIE 2018, Volume 1. Springer Singapore, 2020, pp. 13–24.

[7] W. Ayadi, W. Elhamzi, I. Charfi and M. Atri, *A hybrid feature extraction approach for brain MRI classification based on Bag-of-words*, Biomed. Signal Process. Control **48** (2019), 144–152.

[8] N.E. Benzebouchi, N. Azizi and K. Ayadi, *A computer-aided diagnosis system for breast cancer using deep convolutional neural networks*, Comput. Intel. Data Min.: Proc. Int. Conf. CIDM 2017, Springer Singapore, 2019.

[9] R. Chaieb and K. Kalti, *Feature subset selection for classification of malignant and benign breast masses in digital mammography*, Pattern Anal. Appl. **22** (2019), 803–829.

[10] S. Dhahbi, W. Barhoumi and E. Zagrouba, *Breast cancer diagnosis in digitized mammograms using curvelet moments*, Comput. Bio. Med. **64** (2015), 79–90.

[11] Y. Gherghout, Y. Tlili and L. Souici, *Classification of breast mass in mammography using anisotropic diffusion filter by selecting and aggregating morphological and textural features*. Evolv. Syst. **12** (2021), 273–302.

[12] A.U. Haq, J.P. Li, A. Saboor, J. Khan, S. Wali, S. Ahmad, A. Ali, G.A. Khan and W. Zhou, *Detection of breast cancer through clinical data using supervised and unsupervised feature selection techniques*, IEEE Access **9** (2021), 22090–22105.

[13] H. Hosseinzadeh, *Automated skin lesion division utilizing Gabor filters based on shark smell optimizing method*, Evolv. Syst. **11** (2020), no. 4, 589–598.

[14] H. Hosseinzadeh and M. Sedaghat, *Brain image clustering by wavelet energy and CBSSO optimization algorithm*, J. Mind Med. Sci. **6** (2019), no. 1, 110–120.

[15] M.W. Huang, C.W. Chen, W.C. Lin, S.W. Ke and C.F. Tsai, *SVM and SVM ensembles in breast cancer prediction*, PloS one **12** (2019), no. 1, e0161501.

[16] L. Hussain, W. Aziz, S. Saeed, S. Rathore and M. Rafique, *Automated breast cancer detection using machine learning techniques by extracting different feature extracting strategies*, 17th IEEE Int. Conf. Trust Secur. Privacy Comput. Commun. /12th IEEE Int. Conf. Big Data Sci. Engin. (TrustCom/BigDataSE). IEEE, 2018, pp. 327–331.

[17] S. Jitaree, A. Phinyomark, P. Phukpattaranont and P. Boonyapiphat, *Classifying breast cancer regions in microscopic image using texture features*, 13th Int. Conf. Electric. Engin. Electronics Comput. Telecommun. Inf. Technol. (ECTI-CON), IEEE, 2016.

[18] M. Karabatak, *A new classifier for breast cancer detection based on Naïve Bayesian*, Measurement **72** (2015), 32–36.

[19] R. Karthiga, K. Narasimhan and G. Usha, *Breast cancer diagnosis using curvelet and regional features*, Int. Conf. Comput. Commun. Inf. (ICCCI). IEEE, 2019, pp. 1–5.

[20] S. Kaymak, A. Helwan and D. Uzun, *Breast cancer image classification using artificial neural networks*, Procedia Comput. Sci. **120** (2017), 126–131.

[21] S. Khan, N. Islam, Z. Jan, I.U. Din and J.J.C. Rodrigues, *A novel deep learning based framework for the detection and classification of breast cancer using transfer learning*, Pattern Recogn. Lett. **125** (2019), 1–6.

[22] Y.C. Kuo, W.C. Lin, S.C. Hsu and A.C. Cheng, *Mass detection in digital mammograms system based on PSO algorithm*, Int. Symp. Comput. Consumer Control IEEE, 2014, pp. 663–668.

[23] J.N.K. Liu, Y.L. He, X.Z. Wang and Y.X. Hu, *A comparative study among different kernel functions in flexible naïve Bayesian classification*, Proc. Int. Conf. Machine Learn. Cybernet. **2** (2011), 638—643.

[24] N. Liu, E.S. Qi, M. Xu, B. Gao and G.Q. Liu, *A novel intelligent classification model for breast cancer diagnosis*, Inf. Process. Manag. **56** (2019), no. 3, 609–623.

[25] X. Liu, J. Tang, *Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method*, IEEE Syst. J. **8** (2013), no. 3, 910–920.

[26] S. Liu, J. Zeng, H. Gong, H. Yang, J. Zhai, Y. Cao and X. Ding, *Quantitative analysis of breast cancer diagnosis using a probabilistic modelling approach*, Comput. Bio. Med. **92** (2018), 168–175.

[27] A.O.I. Malagelada, *Automatic Mass Segmentation in Mammographic Images*, Ph.D. Thesis, Universitat de Girona, 2007.

[28] J.G. Melekoodappattu and P.S. Subbian, *Automated breast cancer detection using hybrid extreme learning machine classifier*, J Ambient Intell Human Comput (2020). https://doi.org/10.1007/s12652-020-02359-3

[29] C.E.F. Matos, J.C. Souza, J.O.B. Diniz, G.B. Junior, A.C. de Paiva, J.D.S. de Almeida, S.V. da Rocha and A.C. Silva, *Diagnosis of breast tissue in mammography images based local feature descriptors*, Multimedia Tools Appl. **78** (2019), 12961–12986.

[30] S.H. Mohan and T.R. Mahesh, *Particle swarm optimization based contrast limited enhancement for mammogram images*, 7th Int. Conf. Intell. Syst. Control (ISCO), 2013, pp. 384–388.

[31] S. Pramanik, D. Banik, D. Bhattacharjee and M. Nasipuri, *AA computer-aided hybrid framework for early diagnosis of breast cancer*, Adv. Comput. Syst. Secur. **8** (2019), 111–124.

[32] A.I. Pritom, M.A.R. Munshi, S.A. Sabab and S. Shihab, *Predicting breast cancer recurrence using effective classification and feature selection technique*, 19th Int. Conf. Comput. Inf. Technol. (ICCIT), IEEE, 2016, pp. 310–314.

[33] A.H. Osman, *An enhanced breast cancer diagnosis scheme based on two-step-SVM technique*, Int. J. Adv. Comput. Sci. Appl. **8** (2017), no. 4, 158–165.

[34] X. Qi, L. Zhang, Y. Chen, Y. Pi, Y. Chen, Q. Lv and Z. Yi, *Automated diagnosis of breast ultrasonography images using deep neural networks*, Med. Image Anal. **52** (2019), 185–198.

[35] S. Sádi, A. Maleki, R. Hashemi, Z. Panbechi and K. Chalabi, *Comparison of data mining algorithms in the diagnosis of type II diabetes*, Int. J. Comput. Sci. Appl. **5** (2015), no. 5, 1–12.

[36] M.S. Salama, A.S. Eltrass and H.M. Elkamchouchi, *An improved approach for computer-aided diagnosis of breast cancer in digital mammography*, IEEE Int. Symp. Med. Measur. Appl. (MeMeA), IEEE, 2018, pp. 1–5.

[37] J. Suckling, *The mammographic image analysis society digital mammogram database*, Exerpta Medical Int. Cong. Ser. **1069** (1994), 375–378.

[38] R. Vijayarajeswari, P. Parthasarathy, S. Vivekanandan and A. Alavudeen Bash, *Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform*, Measurement **146** (2019), 800–805.

[39] Z. Wang, M. Li, H. Wang, H. Jiang, Y. Yao, H. Zhang and J. Xin, *Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features*, IEEE Access **7** (2019), 105146–105158.

[40] D. Wang and Ph.H. Lin Shi, *Automatic detection of breast cancers in mammograms using structured support vector machines*, Neurocomputing **72** (2009), 13–15.

[41] F. Wang, S. Zhang and L.M. Henderson, *Adaptive decision-making of breast cancer mammography screening: A heuristic-based regression model*, Omega **76** (2018), 70–84.

[42] H. Wang, B. Zheng, S.W. Yoon and H. Sang Ko, *A support vector machine-based ensemble algorithm for breast cancer diagnosis*, Eur. J. Oper. Res. **267** (2018), no. 2, 687–699.

[43] J.L. Weidong Tangb and H. Hosseinzadeh, *Developed multiple-layer perceptron neural network based on developed search and rescue optimizer to predict iron ore price volatility: A case study*, ISA Trans. **130** (2022), 420–432.

[44] Y. Xiao, J. Wu, Z. Lin and X. Zhao, *Breast cancer diagnosis using an unsupervised feature extraction algorithm based on deep learning*, 37th Chinese Control Conf. (CCC), IEEE, 2018, pp. 9428–9433.

[45] B. Zheng, S.W. Yoon and S.S. Lam, *Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms*, Expert Syst. Appl. **41** (2014), 1476—1482.

[46] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[47] *The Mammographic Image Analysis Society (MIAS)*, Internet site address: http://peipa.essex.ac.uk/info/mias.html.

[48] *University of South Florida Digital Mammography Home Page*, Available At: http://marathon.csee.usf.edu/Mammograph y/Dhatabase.html.