A Diagnostic Model of Breast Cancer Based on Digital Mammogram Images Using Machine Learning Techniques

Farouk A. K. Al-Fahaidy, Belal Al-Fuhaidi, Ishaq AL-Darouby, Faheem AL-Abady, Mohammed AL-Qadry, and Abdurhman AL-Gamal

1 Faculty of Engineering, Electrical Engineering Department, IBB University, IBB, Yemen
2 Faculty of Computing and IT, University of Science and Technology Sana’a, Taiz, Yemen

Correspondence should be addressed to Belal Al-Fuhaidi; belalarh@gmail.com

Received 19 May 2022; Accepted 10 August 2022; Published 20 September 2022

Academic Editor: V. E. Sathishkumar

Copyright © 2022 Farouk A. K. Al-Fahaidy et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Breast cancer disease is one of the most recorded cancers that lead to morbidity and maybe death among women around the world. Recent research statistics have exposed that one from 8 females in the USA and one from 10 females in Europe are contaminated by breast cancer. Hence, breast cancer is the main problem in public health. So, the early detection of this disease is the best stratagem for fighting it. Mammography diagnosis is the most commonly precise tool that can be employed for the early detection of breast cancer [2, 4, 5]. The preprocessing step of mammogram images is crucial. In research [6], the objective of the Breast Imaging-Reporting and Data System (BI-RADS) of the American College of Radiology (ACR) is to provide a consistent classification system for commentary mammographic breast concentrations. Different research makes the determination either to detect breast lesions using computer-aided detection systems (CADe) or to understand mammograms through Computer-Aided Diagnostic Systems (CADx). These systems are working as a supplement to the radiologists’ valuation. In general, there are four steps to be tracked...
in the development of a CAD system for diagnosing suspicious regions in mammograms. The first step is the preprocessing for preparing mammograms to be applied in the following steps, without noise. The second one is the identification of regions of interest (ROI) for selecting the desired mammogram information. The third step is the extraction and optimum selection of the features from the identified ROIs. Finally, the classification step of ROIs aims to classify mammograms and decide if it is a normal mammogram or abnormal mammogram which is either a malignant mammogram or a benign mammogram [7, 8].

As it was mentioned above, there are many variant recent works that aimed to propose CAD systems for breast cancer diagnosing, using different feature extraction techniques, such as wavelet processing, statistical methods, and different classification methods like machine learning, neural networks, and deep learning. Recently, AI based on machine learning plays a promising and attractive branch in classification purposes [from tables in related works]. Most of the related works are proposing a classification methodology for benign and malignant while the classification of normal and abnormal is not identified clearly. This work aims to explore a flexible and effective machine learning method for breast cancer diagnosing of normal/abnormal and benign/malignant classifications, with systematic steps and identified methods/techniques. A mammography dataset of the Mammographic Image Analysis Society (MIAS) database [9] is chosen. The proposed CAD system is illustrated in the following steps. The first step, the preprocessing step of the MAIS dataset images, is carried out by applying different image processing techniques such as noise removal, artifacts and labels suppressing, and image enhancement. The 2d median filter is used for noise removal from a mammogram image then, a morphological operation is carried for suppressing artifacts and labels of mammogram images, and finally, the contrast enhancement of mammography images is performed. The second step, the segmentation step, is carried out by taking away the pectoral muscle, using the seeded region growing (SRG) algorithm [10–12]. Subsequently, the identification and extraction of the region of interest (ROI) are done to the result of the segmentation. In the third step, the feature extraction step, several features are extracted from the ROI using different feature extraction classes such as the first-order statistical class, the second-order statistical class based on the gray level cooccurrence matrix (GLCM), the fractal dimension class, the shape class, and the wavelet features class. After that, the step of the optimized feature selection is involved picking the optimum and effective subset of features, which show clear effects on the classification accuracy. This is performed using the Sequential Forward Selection (SFS) technique. Finally, the classification of either normal or cancer by its two different types, benign and malignant, is decided. More details on these steps are presented and discussed in Section 3. The main contributions of this work can be summarized as follows:

(i) A proposed method based on machine learning for diagnosing breast cancer with mammography images is introduced and illustrated in systematic steps, studied, and investigated.

(ii) The MAIS dataset of mammography images is used after carrying out the preprocessing step of the dataset in three substeps, noise removal using 2d median filter, thresholding and contrast enhancement, and pectoral muscle removal using SRG technique.

(iii) 307 features are extracted using different statistical methods and wavelet feature classes in the suggested method, and only 21-features with an effective impact on the accuracy are chosen during the training phase.

(iv) The SVM algorithm is used for classification in two levels, one for normal and abnormal classification and the other for benign and malignant classification.

(v) A CAD system with GUI manipulation is implemented for flexible manipulation of mammography.

(vi) The accuracy of the proposed method has reached 100% accuracy in the case of normal and abnormal classification of mammography.

The rest of this paper is organized as follows: Section 2 introduces a survey on related works; Section 3 presents and discusses the proposed system. Section 4 demonstrates the practical implementation of the proposed CAD systems and presents and discusses experiments and results. Finally, Section 5 summarizes the conclusions of this work.

2. Literature Review and Related Works

There are numerous approaches that have been anticipated for mammogram diagnosing. In general, they are can be grouped into statistical-based methods [13, 14], wavelets-based methods [15–17], Markovian-based models [18], machine-learning-based methods [19], etc. Numerous investigations have been issued on computer breast cancer diagnosis. In [20], the author presents an overview of the recent advances in the field of CAD breast cancer diagnosis based on mammogram image analysis. In [21], this research presents an outline of procedures that have been suggested in analyzing breast cancer images’ histopathology.

For the development of a CAD system for breast cancer, the process of distinguishing the ROIs is an important step that plays a key challenge. Many suggestions have been introduced by researchers for breast tissue/muscle region segmentation according to the density and the texture variances [22]. A proposal provided by [23] uses the Bayesian techniques with a Markovian random field to partition mammogram images into three diverse regions, the pectoral muscle, the fatty, and the fibroglandular regions. Other approaches were using the LBP, the K-means, SVM, and the GLCM algorithms for identifying the ROI regions from mammogram images like in [24–26]. The proposals presented by [2–4] introduce adaptive thresholding methods based on multiresolution for detecting suspicious lesions in mammogram images. For the feature extraction/
classification step, there are many different types of research suggestions, and the Automatic CAD detection/classification system of suspicious lesions was presented by [27]. In [28], the SVM classification was applied to the development of a classification algorithm for breast masses. The author in [29] presents a different machine learning technique for classifying breast cancer as malignant via cytological imaginings of fine-needle aspiration. The work presented by [30] proposes an automated breast cancer diagnosing system by employing the GVFSnake Segmentation method, the wavelet-based feature extraction, and the fuzzy-based classification. A hybrid optimization algorithm-based feature selection for mammogram images and hybrid transfer learning for detecting the breast masses of mammographs are presented by [31, 32]. The authors in [33] proposed an automated breast cancer diagnosis system using the wavelet-based feature extraction, and the fuzzy-based classification. A hybrid optimization algorithm-based feature selection for mammogram images and hybrid transfer learning for detecting the breast masses of mammographs are presented by [31, 32]. The authors in [33] proposed a novel computer-aided diagnosis (CAD) system based on one of the regional deep learning techniques, an ROI-based convolutional neural network for simultaneous detection and classification of breast masses in digital mammograms. In [34, 35], the research is developing a CAD system that employs a temporal analysis for improving radiologists. From the previous studies in the field of breast cancer detection and classification, it is clear to us that the CAD system is an attractive way that could lead to good marks in diagnosing a breast cancer disease. The systematic review in [36] provides a comprehensive description and analysis of existing CAD systems that make use of machine learning techniques, as well as an assessment of how they now stand in relation to various categorization schemes and mammography image modalities. All CAD phases, including preprocessing, segmentation, feature extraction, feature selection, and classification, were covered in this systematic study. The systematic review outlined suggestions for the next study and identified research gaps.

Table 1 presents some related works and their shortcomings discussed in the following paragraph. In this table, there are different proposed methods with different considerations. Some were working with machine learning like K-mean, fuzzy C-means, and neural networks (NN) and compared them with SVM, and others using multiclassifiers, artificial neural networks (ANNs), or convolutional neural networks (CNNs). On another side, there are many different datasets and different feature extraction and selection techniques. Overall, the SVM classifier has approved its practice effectiveness, and the wavelet and GLCM feature extraction classes are performed well. Moreover, there is no clear classification will be identified for normal versus abnormal classification. Hence, we make our decision for carrying out effective preprocessing of the MIAS dataset which is the only available dataset to us in 2018 when we start the work. The preprocessing of the MIAS mammograms is carried out based on the proposed work in [10]. Then, for

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Classifier method</th>
<th>Dataset</th>
<th>Features</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaleel et al. [37]</td>
<td>ANN</td>
<td>MiniMIAS</td>
<td>Discrete wavelet transform and GLCM</td>
<td>Accuracy, sensitivity, specificity</td>
</tr>
<tr>
<td>Kaur et al. [38]</td>
<td>SVM, KNN, LDA and DT</td>
<td>MiniMIAS</td>
<td>SURF</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Kamil et al. [39]</td>
<td>KNN</td>
<td>MiniMIAS</td>
<td>Gray level Cooccurrence matrix</td>
<td>Accuracy, sensitivity, specificity</td>
</tr>
<tr>
<td>Al-Azzam and Shatnawi [40]</td>
<td>Compare supervised learning (SL) with semisupervised learning (SSL) for 9 algorithms</td>
<td>Wisconsin diagnostic breast cancer (WDBC) dataset</td>
<td>30 attributes of 569 patients with 569 instances</td>
<td>Accuracy, sensitivity, specificity</td>
</tr>
<tr>
<td>Huang et al. [41]</td>
<td>SVM with three functions and two features: bagging and boosting</td>
<td>Two datasets</td>
<td>Accuracy, sensitivity, specificity</td>
<td>Accuracy, specificity</td>
</tr>
<tr>
<td>Asri et al. [42]</td>
<td>SVM, C4.5, NB, k-NN</td>
<td>Wisconsin breast cancer (original) dataset (WBCD)</td>
<td>11 attributes of 699 instances</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Rawal [43]</td>
<td>Compare (K-means, EM, PAM, and fuzzy C-means) with SVM and C5.0</td>
<td>Wisconsin prognostic breast cancer dataset</td>
<td>32 attributes with 194 instances</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Kourdifi and Bahaj [44]</td>
<td>Random forest, naive Bayes, SVM, KNN</td>
<td>Wisconsin breast cancer dataset</td>
<td>30 attributes with 699 instances</td>
<td>Accuracy, sensitivity, specificity</td>
</tr>
<tr>
<td>Cai et al. [45]</td>
<td>Conventional neural network (CNN)</td>
<td>The datasets were collected at two medical institutions</td>
<td>990 images, 540 malignant masses, and 450 benign lesions</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Ionescu et al. [46]</td>
<td>CNN</td>
<td>Private dataset</td>
<td>67,520 mammographic images from 16,968 women</td>
<td>Accuracy, sensitivity, specificity</td>
</tr>
<tr>
<td>Zebari et al. [47]</td>
<td>Five classifiers with ANN</td>
<td>Mini-MIAS, and other datasets</td>
<td>(i) LBP (ii) FD (iii) Proposed M-FD</td>
<td>Accuracy, sensitivity, specificity</td>
</tr>
</tbody>
</table>

Table 1: Related works that are carried out for proposing CAD systems for breast cancer.
feature extraction, we try to use GLCM classes and wavelet classes followed by feature selection using the SFS technique. Finally, based on the literature, the SVM is chosen as a classifier.

3. The Proposed CAD System Methodology

An abstract view of the proposed CAD system processes is shown in Figure 1. The flow of processes of the suggested CAD system is starting by preprocessing the mammogram images for removing noise. The next step, image segmentation, is carried out for detecting and identifying the ROIs of mammogram images. Then, the ROIs are used in the features extracting step. After that, the Sequential Forward Selection (SFS) technique is used for the important and relevant feature selection. The final step, the classification, is applied to mammogram’s features to classify mammograms as either normal or abnormal, with further classification of abnormal images as benign or malignant. A detailed description of these steps is presented in the followed subsections.

3.1. Digital Mammography Image Dataset. As an initial step, before the processes of the proposed CAD system for breast cancer diagnosing system, data collection will be obtained.

3.1.1. Data Collection. We obtain a well-known digital mammogram dataset from a known data acquisition society called the mini-MIAS database. The obtained images include left and right breast images of breasts that are fatty, fatty-glandular, and dense-glandular. The three basic cases of the obtained mammogram images are malignant, benign, and normal that are each further categorized into five groups as follows:

(1) Constrained masses
(2) Spoilage-like masses
(3) Irregular masses
(4) Masses with deformed architecture
(5) Asymmetries in the masses

A well-known available mammogram images database which represents 322 digital mammogram images of 161 pairs at a resolution of 50 μ as a portable gray map style with associated truth data, from the MIAS, UK [9], is selected. The MIAS data is used to feed as standard inputs to the proposed CAD system in this work. Images are associated with annotation labels.

3.2. The First Step: The Preprocessing of Database. Naturally, different medical images, such as mammogram images, are hard to understand or interpret; therefore, preprocessing is required to enhance the quality of images by removing noise and to make better results in the segmentation step. Image preprocessing includes noise removal, artifact suppression, and background separation [10].

3.2.1. Noise Removal. The majority of obtained mammography images contain digitization noises like straight lines, which are filtered using a two-dimensional (2D) median filtering technique in a 3-by-3 neighborhood connection. Each output pixel includes the median value for the 3-by-3 area surrounding the corresponding input pixel for removing artifacts, labels, and unwanted image borders by using morphological operations. However, the images’ edges are changed to zeros. A mammography image is shown in Figure 2(a) with the digitization noise present, and the same image is shown in Figure 2(b) with the noise removed.

3.2.2. Artifact Suppression and Background Separation. Using threshold and morphological techniques, shadow artifacts in the mammography images such as wedges and labels are eliminated. Figure 2(a) displays a mammography image with a shadow artifact that was taken from the MIAS mammogram database. A global threshold with a value of $\bar{I}_{\text{normalized}} = 18$ normalized value is discovered to be the most suitable threshold for converting the grayscale images into binary [0, 1] format through manual inspection of all mammography images acquired. The binary mammography pictures are subjected to morphological operations such as dilation, erosion, opening, and closing after the grayscale mammogram images are converted to binary, as shown in Figure 2(c) for the image in Figure 2(a). The breast profile region is likewise separated from the background during artifact, wedge, and label suppression (see Figure 2(d)).
Figure 3 presents a sample mammogram image with different types of elements before applying the preprocessing and segmentation steps.

3.3. The Image Segmentation Step. The main objective of image segmentation is to identify and extract the ROI, i.e., the pectoral muscle, which has similar characteristics that may match the tumor in the breast profile if exist.

Initially, for obtaining the best outcomes of segmentation, we do the following steps:

1. Prior to completing seeded region growing (SRG), it is necessary to determine the breast orientation in each mammography image. The binary image in Figure 2(c) is used to determine the breast profile orientation (left or right) using an automated approach. The binary image is chopped from top to bottom and from left to right so that the breast profile hits the image’s four boundaries (top, left, right, and bottom). In the cropped binary images, the sum of the first and final five columns of binary values is then determined. The breast profiles are oriented by simply comparing the sums of the first and final five columns; if the first sum is bigger than the last, the breast is right-oriented; otherwise, it is left-oriented.

2. The image contrast of the breast profile images is enhanced using image adjusting and stretching limits techniques in MATLAB.

3. Then, the image segmentation is carried out via the SRG algorithm for extracting the pectoral muscle from a processed mammogram image. For more details on the SRG algorithm, see [11, 12].
3.3.1. The Region of Interest (ROI) Selection. Features necessity is to be calculated as of abnormal region, benign and malignant, in breast profile without no including all other insignificant fragments of the breast tissue. So, features could not be directly calculated from the segmented mammogram images. That is due to the fact that the segmented image willpower biases the exposure results. For such cases, the Ground Truth data assimilated along with the mammography’s dataset information are used for identifying as well as for extracting the ROIs [9]. According to the size of ROI, a suitable ROI window size that is adequate for encircling the majority of benign and malignant abnormalities is selected to be $64 \times 64$ pixels [23, 24]. So, the proposed CAD system uses an ROI block of size $64 \times 64$ pixels. Figure 4 shows examples of ROIs within 3-different mammogram images, Figure 4(a) is an example of benign ROI, Figure 4(b) is an example of malignant ROI, and Figure 4(c) is an example of normal ROI.

3.4. The Feature Extraction Step. Since the conventional images are complicated to manipulate and are highly textured, they led to difficult interpretation. So, the digital mammogram images are used in the feature extracting phase of the proposed system based on their ability for visualizing masses. This is because the spatial resolution of digital mammogram images resulting from an X-ray is in the order of limited microns. Consequently, it is necessary to enhance the performance of the diagnosis in terms of accuracy and reliability by extracting features from digital mammograms. Figure 5 shows five classes of feature extraction methods that are applied in this work.

<table>
<thead>
<tr>
<th>Table 2: List of eighteen first-order statistical features.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Standard deviation</td>
</tr>
<tr>
<td>Kurtosis</td>
</tr>
<tr>
<td>Skewness</td>
</tr>
<tr>
<td>Variance</td>
</tr>
</tbody>
</table>

An explanation of these five main classes of feature extraction is presented in the following subsections.

3.4.1. The Features of First-Order Statistics Class. The first-order statistics feature class offers variant statistical properties representing the image’s intensity histogram. The first-order statistic depends on the intensity values of each pixel only. Table 2 lists 18-features representing the first-order textural features that could be extracted from ROI [48, 49].

3.4.2. The Features of Second-Order Statistics Class. The most commonly used second-order statistics feature is the GLCM feature, which can be applied in the extraction of 2nd order texture information on images based on well-known features having a strong statistical tool. Table 3 presents a list of GLCM features [50–52].

3.4.3. The Features of the Shape Class. To notice the ROI’s shape of images, the shape features’ class could be employed. Many shape features can be extracted from mammogram images; the quantitative analysis of these shape’s properties permits us to distinguish between malignant lockups from those of benign as well as the normal cells. Eight shape
features could be derived from each ROI, the spread-ness, and the seven invariant moments [48, 53].

3.4.4. The Features of the Fractal Dimension Class. The fractal is an asymmetrical geometric entity with infinite nested structures on its whole scale. The most significant properties of fractals are chaos, self-similarity, and fractal dimension. The fractal dimension class provides a calculable measure of the similarity-on-self and the scaling features. The two ways which are used to estimate the fractal dimension features and used by the proposed system in this work are as follows:

- The piecewise triangular prism surface area (PTPSA) technique,
- The piecewise modified box-counting (PMBC) technique.

The exponent of the number of similarity-on-self fragments, \( N \), to the magnification factor, \( 1/R \) into which a figure may be fragmented defines the fractal dimension [54], as listed by

\[
FD = \frac{\ln(N)}{\ln(1/R)}
\]  

(1)

3.4.5. The Wavelet Features’ Class. One of the most commonly used signal/image processing/representation techniques is wavelet transformation. Typically, wavelet transformation of images results in two types of decomposition, the detail, and the approximation coefficients. The detailed coefficients represent high-frequency components of the input image; the approximations represent low-frequency components of the input image. Wavelet transform is good in extracting the characteristics’ vectors. In this numerical signature, the vector of characteristics represents masses and calcifications through the smaller-scaled image with a minimum amount of values; this depends on the resolution of the input image [55, 56]. In this work, the wavelet transform is used for detecting and segmenting masses and calcifications, by evaluating the practice of the Haar and the Daubechies wavelet types for breast lesions, mass, and calcification and characterization. These two wavelet transforms can characterize the difference between masses and calcifications over the gray levels. Figure 6 demonstrates the wavelet processing of the image as a low-frequency filter is used to pass low-frequency components of the input image while a high-pass filter that passes high-frequency components of the image results in detailed components.

3.5. The Important Feature Selection Step. The critical role in providing the best performance of the classifier of the image is how accurately the features are extracted and selected by the classifier. The optimum selection of the effective features from the whole feature set provides the best accuracy of the classifier. Roughly, there are three different categories of feature selection methods which are as follows: the filtering methods, the Wrapper methods, and the embedded methods. In this work, the Sequential Forward Selection (SFS) technique, which belongs to the Wrapper methods, is used for the selection of the important features from the extracted features [57, 58]. In the SFS selection method, an empty set of features is growing sequentially, by adding variant features to the feature set, training the model with these added features, and dropping out features with the lowest effects, until the addition of extra features does not reduce the criterion.

3.6. The Classification Step. The final stage of the CAD system is the classification which is used to categorize breast cancer to be either normal or abnormal as malignant or benign. It aids in predicting features using the Support Vector Machine (SVM) algorithm. SVM was established by the machine learning community [59]. The SVM is a classifier defined by a separation of hyperplanes. The hyperplane is defined by the distance from such hyperplane to the nearest data points on each side, termed maximal support vectors [60]. It could be extended to non-linearly separable data with the help of kernel function application on the data.
to make them linearly separable [61]. An approach based on wavelet SVM was discussed by [62]. More details on SVM and its application to the diagnosing of breast tumor was discussed in [63, 64].

4. Computer Experiment Study and Investigation

This section presents the experimental results as well as the implementation of the proposed computer-aided detection and diagnosis (CADD) system. Different experiments have been carried out in the sequential stages followed by the CADD system. The next subsections are providing the experimental results of the preprocessing, image segmentation, feature extraction and selection, and classification stages. Then, the implementation of the proposed CADD system in the working environment, the MATLAB software, is presented, discussed, and tested.

4.1. The Preprocessing. The preprocessing is carried out by applying a mammogram image to be passed through the nonlinear median filter for noise removal; then, the filtered binary image is separated from the background. Finally, the only breast profile is taken as an output from this phase. Figure 2 shows the outputs of these consecutive experiments that are carried to the input processed mammogram image.

The input image is shown in Figure 2(a), and the filtered image after noise removal is presented in Figure 2(b), which shows a significant improvement against the input image in its clearance. Figure 2(c) demonstrates the image’s background separation. Finally, the breast profile of the filtered image after the separation of background is presented in Figure 2(d).

The output of this stage, the breast profile, is applied as input to the segmentation step for extracting the ROI.

4.2. The Image Segmentation. The objective of the segmentation phase is to identify and extract the pectoral muscle, i.e., the ROI. Initially, in the output of the preprocessing step, the breast profile of the input image’s contrast is enhanced for obtaining the finest results from segmentation. Then, the image segmentation is carried out by applying the SRG algorithm for extracting the pectoral muscle from a processed mammogram image. Figure 8(a) shows the input image after contrast enhancement with identified seed point while Figure 8(b) presents the pectoral muscle extraction.

4.3. The Feature Extraction. In this phase, the pectoral muscle is fed as input to the five classes of feature extraction presented in Figure 5. The output of feature extraction is 307 features that are extracted from each ROI. These features are organized as follows:

Figure 7: A view of Haar wavelet features.
The first-order statistics features that are extracted are 10 features.

The second-order statistics features that are extracted are 62 features.

The shape features that are extracted are 9 features.

The fractal dimension features that are extracted are 2 features.

Finally, the wavelet features that are extracted are 224 features.

From these features, the optimum and effective small number of features are selected by the next step, the feature selection, to be used in the final stage, the classification stage.

A snapshot of a sample of statistical features which represent some different first- and second-order statistical features matrix values is shown in Figure 9. Figure 10 shows a portion of wavelet features matrix values extracted by feature experiments. It should be noted that each row in both of the previous figures numbered Figures 9 and 10 represents a sample image and the columns contain the values of the extracted features.
4.4. **The Important Feature Selection.** As shown in the previous Section 4.3, there are a total number of 307 features that are extracted based on different types of features which is a large number of features. This phase aims to optimize the number of features by selecting the optimum number of effective features. This is performed using the SFS technique with exhaustive experiments carried out and examined in the two steps of the training and the testing phases using samples around 70% and 30%, respectively. More explanation on learning and testing samples ratios and results is shown in the next subsection, the classification. Exhaustive experiments are carried out in training and testing phases for selecting the effective features based on the steps illustrated in Figure 11. As a result of these experiments, 5 features are selected from the total of 83 statistical features, and 16 features are selected from the total of 224 wavelet features. The selected features show their importance thru the training phase; i.e., there are a total of 21 features that are selected from the total of 307 features.
effective features that are selected during the training phase. These 21 features are then used for the final version of the proposed CADD system in a testing phase.

4.5. The Classification. A well-known classifier algorithm, the SVM, is used for the classification stage of the proposed CADD system. Since the SVM is a binary classifier, in this work, two SVM classifiers are used to avoid this problem as indicated in Figure 12.

In the operation of these two SVM classifiers as illustrated in Figure 12, the first SVM classifier is used to identify if the entered data, the features, indicate cancer or no cancer, i.e., abnormal class or normal class. Then, based on the resulting class from the first classifier, the second SVM classifier may be activated or inactivated. Intuitively, if the first classifier indicates a normal case, the two outputs of the second classifier will be 0. Otherwise, the second SVM classifier is activated to decide on the type of cancer, either a malignant class or a benign class.

4.5.1. Performance Evaluation of the SVM Classifier. To implement the SVM classifier, the selected features from the mammogram image are separated into two distinct sets, the learning set, and the testing set. Table 4 shows the total number of samples, 324 samples which are separated into two main groups, around 70% of total samples are used for the learning/training set and the remainder, around 30% of samples, are used for the testing set. From these two sets, there are 208 samples of normal classes and 116 of abnormal classes with 49 samples belonging to the cancer class and 67 samples belonging to the benign cancer class.

The training accuracy of the SVM classification machine could be estimated by regulating the error parameter. Our proposed system uses the RBF (Gaussian) with nonlinear SVM classification. The RBF kernel’s parameter is employed to control the width of the Gaussian and needs to be optimized for the SVM hyperparameter. Therefore, the determination of two SVM hyperparameters is needed for constructing an optimum classifier that balances its generalization and memorization capabilities.

Now, the results of testing of the two working SVM classifiers are obtained and presented for dialogue in Tables 5 and 6, of the two binary confusion matrices.

4.5.2. Performance Evaluation of the SVM Classifier. The results of the first SVM classifier, in the testing of normal and abnormal classes, show that the classifier’s Sensitivity, Specificity, and Accuracy are all 100% which is a great performance while the second SVM classifier, in the testing of malignant and benign classes, shows that the classifier’s Sensitivity is equal to 90%, the Specificity is 85.714%, and the Accuracy is equal to 87.1%; this is due to the use of few numbers of images for diagnosing malignant and benign cancer types. However, the second SVM classifier is

---

**Table 4: The training and testing ratios distribution of samples from the MIAS database.**

<table>
<thead>
<tr>
<th>The class type</th>
<th>Number of samples</th>
<th>Training set (around 70% of samples)</th>
<th>Testing set (30% of samples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal (no cancer)</td>
<td>208</td>
<td>157</td>
<td>51</td>
</tr>
<tr>
<td>Abnormal (malignant)</td>
<td>49</td>
<td>37</td>
<td>12</td>
</tr>
<tr>
<td>Abnormal (benign)</td>
<td>67</td>
<td>48</td>
<td>19</td>
</tr>
<tr>
<td>Total number</td>
<td>324</td>
<td>242</td>
<td>82</td>
</tr>
</tbody>
</table>
Table 5: The confusion matrix of the first SVM classifier, in the testing of normal and abnormal.

<table>
<thead>
<tr>
<th>The class type</th>
<th>Correct classified</th>
<th>False classified</th>
<th>Total number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal (positive class)</td>
<td>TP = 51</td>
<td>FP = 0</td>
<td>51</td>
</tr>
<tr>
<td>Abnormal (negative class)</td>
<td>TN = 31</td>
<td>FN = 0</td>
<td>31</td>
</tr>
</tbody>
</table>

Table 6: The confusion matrix of the second SVM classifier, in the testing of malignant and benign.

<table>
<thead>
<tr>
<th>The class type</th>
<th>Correct classified</th>
<th>False classified</th>
<th>Total number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive class (malignant)</td>
<td>TP = 9</td>
<td>FP = 3</td>
<td>12</td>
</tr>
<tr>
<td>Negative class (benign)</td>
<td>TN = 18</td>
<td>FN = 1</td>
<td>19</td>
</tr>
</tbody>
</table>

providing significantly high performance. By looking at the results presented in Table 6, it is noticed that the malignant class provides a bad performance compared to all other classes; this maybe due to the malignant class being trained and tested with the smallest number of samples of 37 samples in training and 12 samples in testing. So, it can be noted that the performance of the second classifier could be improved by applying an extended dataset of mammogram images which may be available soon.

The suggested system has shown its capability for diagnosing breast cancers with the MIAS dataset. However, this dataset consists of a few hundred images. Therefore, it is noticeable for improving the performance of the suggested system in terms of its accuracy. For dataset size, firstly, we increase the number of available MIAS database images, by applying an augmentation process to images. This is carried out via constructing many forms of images from a given image by applying some augmentation operations, such as different angel rotations and flipping. Secondly, we use other datasets with a larger size and even use many different datasets available on the Internet. Another way for the system model is that one can use a deep learning model based on convolution neural networks or any pretrained convolution neural networks.

4.5.2. Comparative Study and Limitations. This section compares the implemented algorithm in this work with related works to assess how well it performs. The comparison is based on the accuracy metric. The studies that assessed the MIAS database and different databases are used with SVM classifiers and different classifiers.

According to the related works in Table 7, especially those that used few images of the MIAS dataset with the GLCM feature extraction, the proposed method gives a good accuracy compared to the other methods that used KNN [39] and SVM [39] classifiers. Other methods used a large number of images in many datasets along with MIAS for training the model. The authors in [40, 41, 45, 47] trained the proposed methods with different classes of features and different classifiers algorithms, and they outperform our proposed method in terms of accuracy metrics, due to the large numbers of images and different datasets. Our proposed method is characterized by three properties. The first property, the use of two classifiers, has trained the suggested method for classifying cancer as normal/abnormal and benign/malignant while they are only trained and classified cancer as benign versus malignant. The second property is that our suggested method has reached 100% accuracy in classifying normal and abnormal mammograms. And the third property is characterized by the implementation of a computer program with GUI. However, the second-level classifier of benign versus normal of the suggested method reached 87.1% in terms of accuracy. This is due to the minimum number of images that are used. Henceforward, for improving the accuracy of the second classifier, one can either retrain the suggested model with more mammograms either by using of additional dataset or by extending the existed images using augmentation methods to these mammograms.

4.6. The Implementation of the Proposed Computer-Aided Detection and Diagnosis (CADD) System. This work is finalized by implementing an application of the proposed CADD system using MATLAB software environment, version R2013a, on the windows 8 operating system, by employing MATLAB’s image processing tools and statistical tools. All experiments are implemented on a Dell laptop of Intel Core i7 with the processing power of 2.4 GHz CPU with 12 GB RAM. Figure 13 illustrates the graphical user interface working environment of the proposed CADD system.

Figure 14 shows the relationship between different files used in the development of the proposed CADD system application. The dashed spherical shape representation means there are additional functions to this component/file code. These files/components and their operations are summarized as follows:

(i) Image reading.m: this file allows the user to enter the mammographic image from a specific file directory to be diagnosed by the proposed CADD system.

(ii) Image preprocessing.m: This file improves the quality of the input image and removes the noise. It includes the following MATLAB files:

(i) Binary.m which eliminates the noise and converts the image into a binary image.

(ii) Breast.m which removes the labels and identifies the orientation of the image.
### Table 7: The comparative study with related works.

<table>
<thead>
<tr>
<th>Related work</th>
<th>Classifier method</th>
<th>Dataset</th>
<th>Features</th>
<th>Accuracy of normal/abnormal classifier</th>
<th>Accuracy of benign/malignant classifier</th>
<th>CAD tool based-GUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kamil Jassam [39]</td>
<td>KNN</td>
<td>MIAS</td>
<td>Gray level cooccurrence matrix (GLCM)</td>
<td>None</td>
<td>86.1%</td>
<td>No</td>
</tr>
<tr>
<td>Al-Azzam and Shatnawi [40]</td>
<td>Compare supervised learning (SL) with semisupervised learning (SSL) for 9 algorithms</td>
<td>Wisconsin diagnostic breast cancer (WDBC) dataset</td>
<td>30 attributes of 569 patients with 569 instances</td>
<td>None</td>
<td>K-NN (SL = 98% &amp; SSL = 97%) and logistics regression (SL = 97% &amp; SSL = 98%)</td>
<td>None</td>
</tr>
<tr>
<td>Huang et al. [41]</td>
<td>SVM with three functions and two features: bagging and boosting</td>
<td>Two datasets</td>
<td>First dataset: 11 attributes with 699 instances Second dataset: 117 attributes with 102294 instances</td>
<td>None</td>
<td>96.85%, 95%</td>
<td>None</td>
</tr>
<tr>
<td>Cai et al. [45]</td>
<td>Conventional neural network (CNN)</td>
<td>The datasets were collected at two medical institutions</td>
<td>990 images, 540 malignant masses, and 450 benign lesions</td>
<td>None</td>
<td>87.68%</td>
<td>No</td>
</tr>
<tr>
<td>Zebari et al. [47]</td>
<td>Five classifiers with ANN</td>
<td>Mini-MIAS, and other datasets</td>
<td>(i) LBP (ii) FD (iii) Propose M-FD</td>
<td>None</td>
<td>Train, test 79.11, 83.61, 80.06, 89.55, 96.2, 100</td>
<td>No</td>
</tr>
<tr>
<td>Kamil and Jassam [39]</td>
<td>SVM</td>
<td>MIAS</td>
<td>GLCM</td>
<td>None</td>
<td>83.5%</td>
<td>No</td>
</tr>
<tr>
<td>Our work</td>
<td>Two-level SVM classifiers</td>
<td>MIAS dataset</td>
<td>Statistical classes, wavelet class, and SFS technique: 21 attributes with 307 instances</td>
<td>100%</td>
<td>87.1%</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Figure 13: The proposed CADD system application.**
(iii) *Artifact.m* which removes the artifacts from the image and separates the background from the breast profile.

(iii) *Image segmentation.m*: this file implements the contrast enhancement method and removes the pectoral muscle from the preprocessed image. It includes the *region.m* file which implements the pectoral muscle removal by using the SRG algorithm (implemented in the *region-growing.m* file) and enhances the image after pectoral muscle removal.

(iv) *Texture feature.m*: this file extracts 83 texture features from the segmented image.

(v) *Wavelet-feature.m*: this file extracts 224 wavelet features from the segmented image.

(vi) *Feature-selection.m*: it selects the optimum texture and wavelet features for providing better classification accuracy. It employs the SFS algorithm which selects 5 of 83 texture features and 16 from 224 wavelet features.

(vii) *First-classifier.m*: it implements the first SVM classifier by using the selected 5 texture features to classify cases as normal and cancer cases. It includes the following files:

(i) *svmtrain.c* which trains the first SVM classifier.

(ii) *svmpreduct.c* which tests the first trained SVM model.

(iii) *svmtrain.c* and *svmpreduct.c* which are implemented by using the visual studio compiler.

(viii) *secondclassifier.m*: it implements the second SVM classifier by using the selected 16 wavelet features to classify cases as benign and malignant cases. This file includes the following files:

(i) *svmtrain.c* which trains the second SVM classifier.

(ii) *svmpreduct.c* which tests the second trained SVM model.

(iii) *project.m*: which contains the graphical user interface which gathers all the above files into this file.

5. Conclusions

This work has proposed a CAD system for breast cancer identification and diagnosis from mammography images. The proposed method is passed throughout followed and ordered systematic steps for manipulating the mammography images to identify the ROI, feature extraction and optimization, and then for breast cancer classification using a machine learning algorithm, the SVM algorithm. The experimental results of the proposed method have shown its effectiveness, especially in the case of normal and abnormal tumor classifications which reach 100% accuracy. After the investigation of the proposed method with training and testing phases, a CAD software system with GUI is implemented using the proposed method. This CAD tool can play as an efficient tool for early identification and diagnosis of breast cancer from mammogram images. However, the second-level classifier accuracy reached 87% in classifying the cancer class as either benign or malignant. This is due to the small size of the available MAIS dataset.
images, especially mammogram images with tumors. So, we suggest different future works be considered. One of the future works to be suggested is to extend the size of the MAIS dataset using augmentation. Another suggestion is to use another dataset with a larger size. Moreover, researchers may have to use a suitable deep learning machine like the conventional neural networks.

**Data Availability**

The mini MIAS images data set used to support the findings of this study are available from the corresponding author upon request.

**Conflicts of Interest**

The authors declare that they have no conflicts of interest.

**References**


