Chapter 4

Machine Learning Methods for Breast Cancer Diagnostic

Shahnorbanun Sahran, Ashwaq Qasem,
Khairuddin Omar, Dheeb Albashih, Afzan Adam,
Siti Norul Huda Sheikh Abdullah, Azizi Abdullah,
Rizuana Iqbal Hussain, Fuad Ismail, Norlia Abdullah,
Suria Hayati Md Pauzi and Nurdashima Abd Shukor

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.79446

Abstract

This chapter discusses radio-pathological correlation with recent imaging advances such as machine learning (ML) with the use of technical methods such as mammography and histopathology. Although criteria for diagnostic categories for radiology and pathology are well established, manual detection and grading, respectively, are tedious and subjective processes and thus suffer from inter-observer and intra-observer variations. Two most popular techniques that use ML, computer aided detection (CADe) and computer aided diagnosis (CADx), are presented. CADe is a rejection model based on SVM algorithm which is used to reduce the False Positive (FP) of the output of the Chan-Vese segmentation algorithm that was initialized by the marker controller watershed (MCWS) algorithm. CADx method applies the ensemble framework, consisting of four-base SVM (RBF) classifiers, where each base classifier is a specialist and is trained to use the selected features of a particular tissue component. In general, both proposed methods offer alternative decision-making ability and are able to assist the medical expert in giving second opinion on more precise nodule detection. Hence, it reduces FP rate that causes over segmentation and improves the performance for detection and diagnosis of the breast cancer and is able to create a platform that integrates diagnostic reporting system.

Keywords: computer-aided detection, computer-aided diagnosis, support vector machine, false positive, grading

1. Introduction

Breast cancer is one of the most dangerous and common reproductive cancers that affect mostly women. The oldest documented cases of breast cancer were in Egypt in 3000 BC [1]. Breast tumor is an abnormal growth of tissues in the breast, and it may be felt as a lump or nipple discharge or change of skin texture around the nipple region. Cancers are abnormal cells that divide uncontrollably and are able to invade other tissues. Cancer cells have the ability to spread to other parts of the body through the blood and lymphatic systems [1]. It is the leading cause of death among middle aged and older women [1]. According to cancer statistics, breast cancer is the second most common and the leading cause of cancer deaths among women, second only to lung cancer [1]. Around 1 in 36 (3%) women dies due to breast cancer [2]. It has become a major health issue in the past 50 years, and its incidence has increased in recent years [1]; in Malaysia, breast cancer is the most frequent type of cancer among women. It has an incidence rate of about 26% (more than 4400 women) among cancer affecting women. Around 40% of the women who suffered from breast cancer in Malaysia have died (IARC). Hence, determining the right decision from a right diagnosis is crucial.

In today's world with the advent of personalized medicine, it increases the workload and complexity of the doctors in cancer diagnosis. Radiologic and pathology are the key players in making decision for cancer diagnosis. Based on the radiology diagnosis, the results will be submitted to pathology for further diagnosis. Pathology and radiology form the core of cancer diagnosis, yet based on our observation at our studied hospital and under current process of diagnostic medicine, the communication among them remained on papers. That paper contains their respective report of the case on the same patient. This scenario is in parallel with what James et al. [3] had highlighted in their paper. The working flows of both specialties remain ad hoc and occur in separate "silos," with no direct linkage between their case accessioning and/or reporting systems, even when both departments belong to the same host institution. Since both radiologists' and pathologists' data are essential to make correct diagnoses and appropriate patient management and treatment decisions, the isolation of radiology and pathology work flows can be detrimental to the quality and outcomes of patient care. These detrimental effects underscore the need for pathology and radiology work flow integration and for systems that facilitate the synthesis of all data produced by both specialties. With the enormous technological advances currently occurring in both fields, the opportunity has emerged to develop an integrated diagnostic reporting system that supports both specialties and, therefore, improves the overall quality of patient care. In this chapter, we are focusing on breast cancer diagnostic for data collected from UKMMC. Hence, breast radio-pathological correlation is essential. The covered topics would include radio-pathological correlation with recent imaging advances such as machine learning with use of technical methods such as mammography and histopathology.

As a standard, the current diagnostic screening consists of a mammography to identify suspicious regions of the breast, followed by a biopsy of potentially cancerous areas. A breast biopsy is a diagnostic procedure that can determine if the suspicious area is malignant or benign [4–6]. Although criteria for diagnostic categories of radiologic and pathology are well established, manually detection and grading respectively is a tedious and subjective process and thus suffers from inter-observer and intra-observer variations. Early detection via mammography increases breast cancer treatment options and the survival rate. However, mammography is not perfect. Detection of suspicious abnormalities is a repetitive and fatiguing task. For every thousand cases analyzed by a radiologist, only three to four are cancerous, and thus an abnormality may be overlooked. As a result, radiologists fail to detect 10–30% of cancers. Approximately two thirds of these false-negative results are due to missed lesions that are evident retrospectively. Due to the considerable amount of overlap in the appearance of malignant and benign abnormalities, mammography has a positive predictive value (PPV) of less than 35%, where the PPV is defined as the percentage of lesions subjected to biopsy that were found to be cancer. Thus, a high proportion of biopsies are performed on benign lesions. Avoiding benign biopsies would spare women anxiety, discomfort, and expense [7]. As mentioned earlier, with the advent of personalized medicine, the process becomes more complex. Not only that, the emerging of 4th Industrial Revolution (4IR) technology allowed huge amount of data to be captured, and this contributes to the complexity of the radiology and pathology workload. To address these challenges, many researchers are leveraging artificial intelligence to improve medical diagnostics. Machine learning is a sub discipline in the field of artificial intelligence (AI) that explores the study and design of algorithms that can learn from data [8].

2. Machine learning

ML comprises a broad class of statistical analysis algorithms that iteratively improve in response to training data to build models for autonomous predictions. In other words, computer program performance improves automatically with experience [9]. ML algorithm's aim is to develop a mathematical model that fits the data. It comprises of two types of learning which are supervised and unsupervised. Supervised learning algorithm required the data to be labeled for training purposes. For example, in training a set of medical images to identify a specific breast tumor type, the label would be tumor pathologic results or genomic information. These labels, also known as ground truth, can be as specific or general as needed to answer the question. The ML algorithm is exposed to enough of these labeled data to allow them to move into a model designed to answer the question of interest. Because of the large number of well-labeled images required to train models, curating these data sets is often laborious and expensive [10]. Unsupervised ML clusters the data that have similar characteristics, and the unlabeled data are exposed to the algorithm with the goal of generating labels that will meaningfully organize the data. This is typically done by identifying useful clusters of data based on one or more dimensions. Compared with supervised techniques, unsupervised learning sometimes requires much larger training data sets. Unsupervised learning is useful in identifying meaningful clustering labels that can then be used in supervised training to develop a useful ML algorithm. This blend of supervised and unsupervised learning is known as semi-supervised.

ML algorithms are to analyze any data set to extract data-driven model, prediction rule, or decision rule from the data set. Generally, in order to ensure the ML behave intelligently without human intervention, the system learns or extracts knowledge such as rules or patterns from a collection of input data or past experience. So the steps involved can be described as firstly, the system must acquire features from data. Elaboration of features is well explained in our previous work [11, 12]. Feature selection is very important as it contains information that can be used to train the system to identify specific patterns. The pixels are rich with qualitative abstractions or values of the input. Second step is analyzing all these features for detecting and classifying possible pattern or abnormality. Finally, the step is involving a ML algorithm to determine a best suitable model to represent the behavior or the pattern of the data [13].

Various machine learning algorithms are now used to develop high-performance medical image processing systems such as computer-aided detection (CADe) system that detects clinically significant objects from medical images and computer-aided diagnosis (CADx) system that quantifies malignancy of manually or automatically detected clinical objects [14]. Therefore, CADe for mass in mammogram detects the suspicious region in the mammogram then tries to reduce the false positive and finally classifies this region to a mass or nonmass. In CADx for mass in a mammogram, most researchers use a region of interest (ROI) that contains the mass as an input to the CADx. Then, CADx tries to classify it into benign or malignant and gives the appropriate recommendation to do biopsy or follow-up screening [15]. Recent studies have shown that CAD systems, when used as an aid, have improved radiologists' accuracy of detection of breast cancer and also pathology decision [1, 7, 16]. It is worthwhile to distinguish ML from traditional computer-aided detection (CAD) algorithms. Traditional CAD algorithms are mathematical models that identify the presence or absence of image features known to be associated with a disease state. One of the examples is a microcalcification on a mammogram. Traditional CAD allows the developer to identify a feature explicitly and attempts to determine the presence or absence of that feature within a set of images. In contrast, ML techniques focus on a particular labeled outcome (ductal adenocarcinoma), and in the process of training, clusters of nodes evolve into algorithms for identifying features. The power and promise of the ML approach over traditional CAD is that useful features can exist

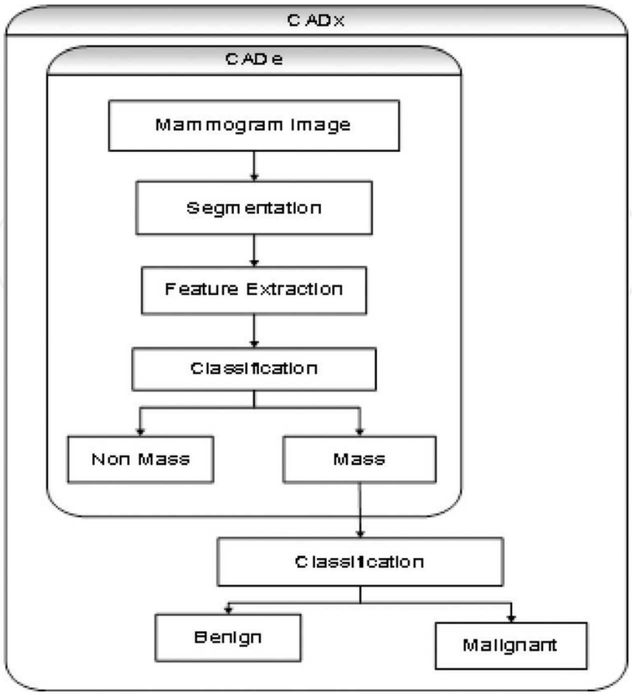


Figure 1. CADe vs. CADx. Source: Sampat et al. [7].

that are not currently known or are beyond the limit of human detection [10]. **Figure 1** shows the difference between CADe and CADx.

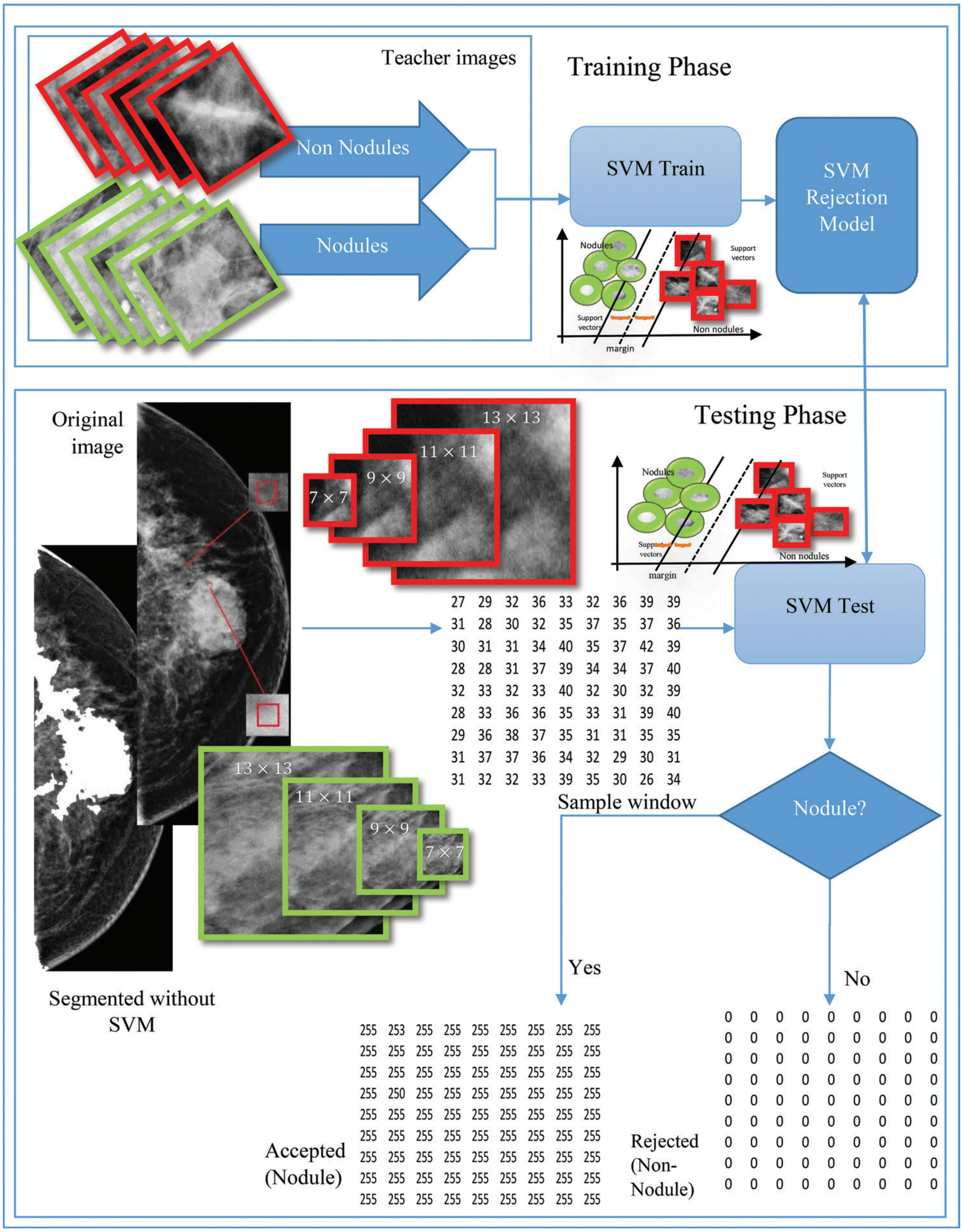
In **Figure 1**, ML algorithm is implemented at the segmentation, feature extraction, and classification steps. One of the most popular and powerful ML algorithm for all the steps is support vector machine (SVM). SVMs are useful for taking a large number of features and discriminating inputs into one of two classes. SVMs, once trained, show the line or border that provides the greatest margin of separation. This concept can be extrapolated to a larger number of features (or dimensions), whereby the line of separation becomes an irregular plane known as a hyperplane. Because of the large number of features that can be combined mathematically, SVMs have been found useful for image processing. This chapter is focusing on SVMs for both CADe for radiology and CADx for pathology diagnostics.

3. Computer-aided detection

Digital medical image recognition (DMIR) might give a promising solution. DMIR is considered as an essential aspect of artificial intelligence. DMIR techniques aim to extract specific information from medical images to assist doctors in diagnosing certain diseases and follow their progress. Many image processing techniques have been utilized in DMIR, such as segmentation, object detection, and classification. DMIR is concerned with numerous imaging modalities in the field of diagnosis including computed tomography (CT), digital mammography, magnetic resonance imaging (MRI), and microscopic histopathological images [16, 17]. Depending on the type of breast tissue, breast mass appears different in a mammogram. While it appears as solid block in dense breast, it appears as a roundish pie in a fatty breast. The mass may be alone or with microcalcifications [1]. In some cases, healthy breasts are also diagnosed as suspicious of cancer by the radiologist, and unfortunately, unnecessary biopsy is performed on them. Knowing that there are many possibilities of masses in breast cancer, detecting these features and localizing them are important. In general, localizing the mass is important in computer-aided detection, where it searches for the location in the mammogram images and segments it. Refs. [1, 18] examine the most important approaches used for mass segmentation in mammogram. In general, localizing the mass is important in computer-aided detection where it searches for the location in the mammogram images and do segmentation. Cheng et al. [18] examine the most important approaches used for mass segmentation in mammogram. Image segmentation using thresholding is the simplest way to isolate the object from its background when the image has a distinct gray level distribution. Segmentation separates the regions by assuming that the region that have gray levels below a specific value, called the threshold, as a background and the region with gray levels higher than the threshold as the object or vice versa. Identifying the threshold value is the key point in this algorithm. By selecting a representable threshold, object extraction will be more accurate. Mostly, image histogram is used to identify the threshold value. Mass localization method is discussed in this chapter. This section is based on our previous work on SVM rejection model for breast cancer. This method is a rejection model based on SVM algorithm used to reduce the FP of the output of the Chan-Vese segmentation algorithm that was initialized by the MCWS algorithm.

Abnormal findings on screening mammograms lead to recall for further assessment, which includes additional imaging procedures and if considered necessary fine needle aspiration cytology, core needle biopsy, or surgical biopsy. Women recalled for further assessment without having a breast cancer diagnosed are considered to have had a FP screening result. FP results are a concern of mammographic screening as they might cause distress, anxiety, and other psychological problems to the women [19, 20]. It also implies additional hospital visits and diagnostic tests, as well as additional costs [21, 22]. The rates of FP screening results depend on the screening performance and organization, such as the screening interval, single versus double reading, participation patterns, sensitivity of the radiologists performance, equipment, and characteristics related to the screening population [22–26]. From image segmentation perspective, the FP is an over-segment result where the noncancerous pixel is segmented as a cancer pixel. The FP rate is considered a challenge in localizing masses in mammogram images. Hence, in this section, a rejection model is proposed by using SVM.

The goal of the rejection model which is based on SVM is the reduction of FP rate in segmenting mammogram through the Chan-Vese method, which is initialized by the MCWS algorithm. The MCWS algorithm is utilized for segmentation of a mammogram image. The segmentation is subsequently refined through the Chan-Vese method, followed by the development of the proposed SVM rejection model with different window size as well as its application in eliminating incorrect segmented nodules MCWS algorithm. SVM rejection model consists of three important stages: (i) initial segmentation, (ii) segmentation using Chan-Vese, and (iii) refined segmentation using SVM rejection model. First, the source image is cropped to remove any unnecessary parts in an image. Based on the high dimensionality in digital mammogram images, the image is then resized to speed up the subsequent processes. Second, completing the pre-processing stage, the SVM rejection model is built to reduce the FP rate. Presegmentation and postsegmentation enhancement for Chan-Vese level set algorithm is then proposed to localize masse in the mammogram. The key to achieve a good segmentation result using Chan-Vese is the initial contour. Instead of getting the initial contour from the expert, here, MCWS algorithm is used to obtain the initial contour, as well as to eliminate the noise. This makes the proposed method fully automated and reduces the time of interference. Lastly, localization of mass in mammogram, Chan-Vese active contour-based algorithm was used. Chan-Vese can find and maximize the convergence ranges, as well as treat the topological change. This ensures that Chan-Vese performs well in image segmentation. Support vector machine is a learning machine algorithm expounded by Cortes and Vapnik [15] at the AT&T Bell Laboratories that strives to address the issues pertaining to a two-group classification. The underlying working principle of this algorithm is to search for the optimal hyperplane that sets positive classes (+1) apart from negative classes (−1). In this context, the two classes are the nodules and the nonnodules of breast images, of which the provided training data were used for the SVM to build a model in predicting the target values of the two test data attributes. In this work, the radial basis function (RBF) kernel is employed in complementary with the SVM. The two best parameters, C and γ, are prerequisites for the generation of an accurate breast nodule and nonnodule classification by the RBF kernel. The SVM rejection model has three phases: extracting teacher image, training, and testing as shown in **Figure 2**. The grid has been used as a straightforward search on the training data to find the best parameters, and the reason for using the grid search instead of other



Sample window

27	29	32	36	33	32	36	39	39
31	28	30	32	35	37	35	37	36
30	31	31	34	40	35	37	42	39
28	28	31	37	39	34	34	37	40
32	33	32	33	40	32	30	32	39
28	33	36	36	35	33	31	39	40
29	36	38	37	35	31	31	35	35
31	37	37	36	34	32	29	30	31
31	32	32	33	39	35	30	26	34

Accepted (Nodule)

255	253	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255
255	250	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255

Rejected (Non-Nodule)

0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

Figure 2. The process of SVM rejection model.

search algorithms is because of its short computation time. Additionally, the grid search can be easily parallelized because it is independent. The search spaces used in this research are {2^{-5}, 2^{10}}. It is important to note that this study used the strategy of dividing the data set into two parts, of which one is considered unknown. The prediction accuracy obtained from the unknown set will reflect on the classification performance of the independent data set. This procedure is known as cross validation. Its goal is to divide the training set into v subsets of equal size. One subset will be tested using the classifier trained on the remaining subsets. Subsequently, each instance of the training set will be predicted once. This is to ensure that the cross-validation accuracy is the percentage of data that have been correctly classified. The training data (teacher images) for the rejection model were manually extracted from the mammogram images by analyzing the false positives (FP) and true positives (TP) of the Chan-Vese segmentation result. After the teacher images were extracted, they were resized using the same factor for the original image. Next, depending on the window size that considered the number of inputs to SVM rejection model, the teacher image was resized. Based on the experiment, either a window size of (7 × 7), (9 × 9), (11 × 11), or (13 × 13) was taken into consideration. After that, the image was transferred to a vector and then written into the training data file. This file contained two variables, x and y. The first variable x is a matrix containing rows of window pixel values for the teacher images. Each row represented one image. The length of the rows depended on the window size. The number of rows in this variable depended on the number of teacher images. The other variable y is a vector containing the class for each image. The class may be “1” for nodule images or “0” for nonnodule images. Before proceeding with the SVM rejection training, training data were used to obtain the best values for parameters C, γ. As previously mentioned, the grid search was used as a straightforward search on the training data to obtain these values. Cross validation was also applied to spill the training data 10-fold into training and testing. Depending on the best accuracy value returned by SVM, the best C and best γ values were chosen. The SVM rejection model was built using the selected C and γ values and the training data set.

Based on model in **Figure 2**, each row in the training data (x_i) represents an observation, and each column represents features. Class labels (y_i) represent the class label for the corresponding row in the training data.

3.1. Results and evaluation

About 170 mammogram images from 109 patients were collected from the UKM Medical Centre (UKMMC). **Table 1** and **Figure 3** show training and testing data that have been used in the experiment. The teacher images extracted from the training data based on the segmentation result contained 35 nodule images and 35 nonnodule images extracted from the training data set. The SVM rejection model was run 10 times with a standard deviation of 0.0001, and the results showed the effectiveness of using the rejection model compared with the ground

	Training data		**Testing data**	
	Nodule	**Nonnodule**	**Nodule**	**Nonnodule**
Number of images	11	17	46	96
Total number of images	28		142	

Table 1. Data set for training and testing.

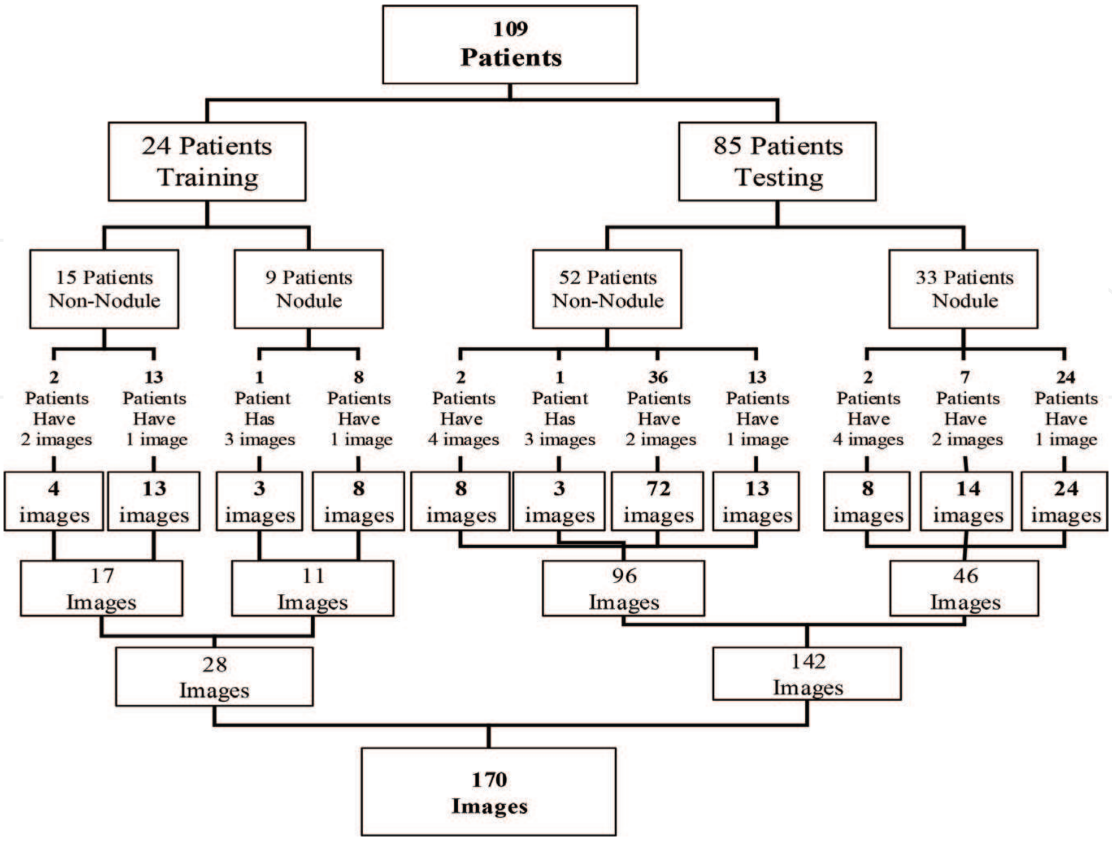


Figure 3. Hierarchy of UKMMC data set.

truth. As mentioned earlier, the grid search was used as a straightforward search on the training data to determine the best parameters C, γ. **Table 2** shows values of C, γ using various window sizes (7 × 7), (9 × 9), (11 × 11), and (13 × 13).

Accuracy denotes the proportion of the correct result and it can be calculated as shown in the following Eqs. (1)–(7), where TP is true positives, TN is true negatives, FP is false positives (type 1 error), and FN is false negatives (type 2 error). In mass localization, the concept of the confusion matrix that is in **Table 2** represents the correctly segmented nodule and nonnodule with the miss segment. TP and TN are the correctly localized nodule and nonnodule, respectively, while FP is the incorrectly segmented nonnodule as a nodule and FN is incorrectly segmented nodule as a nonnodule.

	Result (predicted)	
	Nodule pixel	**Nonnodule pixel**
Ground truth (actual)		
Nodule pixel	TP	FN
Nonnodule pixel	FP	TN

Table 2. Confusion matrix.

Specificity is also known as TN rate, and it represents the ability of the method to identify the nonnodule and avoiding false positives.

Sensitivity, which is also known as TP rate or recall, represents the ability to identify the nodule and avoid false negatives.

The FP rate shows the nonnodule pixel, which is segmented as nodule. It is an over segmented pixel. The FN rate shows the nodule pixel, which is segmented as nonnodule. It is the miss segmented.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

$$Specificity \;\; (SP) = \frac{TN}{TN+FP} \tag{2}$$

$$Sensitivity\,(SE) = \frac{TP}{TP+FN} \tag{3}$$

$$FP \;\; rate = \frac{FP}{FP+TN} \tag{4}$$

$$FN \;\, rate = \frac{FN}{FN+TN} \tag{5}$$

$$Negative \;\; Rate \;\, Matric \;\, (NRM) = \frac{FP \;\; rate + FN \;\, rate}{2} \tag{6}$$

The NRM shows the mismatch between the predicted results and the actual ground truth. Our method was evaluated by comparing the segmented images to the ground truth. To show the effectiveness of the method, a comparison was done before and after the rejection model, as shown in **Figure 4**. This process was performed first by comparing each pixel in the resulting image with the corresponding pixel in the ground truth image. Then, objective evaluation was used to evaluate the method by calculating the confusion matrix as in **Table 2**, based on the prediction result and the actual ground truth. **Table 3** and **Figure 4** show the quantitative analysis of the results and sample of the result. The effectiveness of our method can be proven by comparing the result before and after using the rejection model. **Table 3** shows the FP rate of the rejection model is inversely proportionate to the window size. On the other hand, the specificity rate of the rejection model is linearly proportional to window size.

This section discussed on reducing the FP rate based on SVM machine learning. The SVM rejection model was built to reduce the FP rate after segmentation. Our method has three steps in the segmentation phase: first, MCWS was used to obtain the initial contour by segmenting the mammogram image. Then, the output of MCWS was used as an initial contour to the Chan-Vese algorithm. Finally, the rejection model based on SVM was used in order to reduce the FP rate. The SVM rejection model has three steps in the following order: extracting teacher images, training the rejection model, and testing the model. The FP rate reduction by means of SVM machine learning been put forth, wherein the FP rate, upon segmentation, had been reduced by the developed SVM rejection model. The segmentation of the mass in mammogram

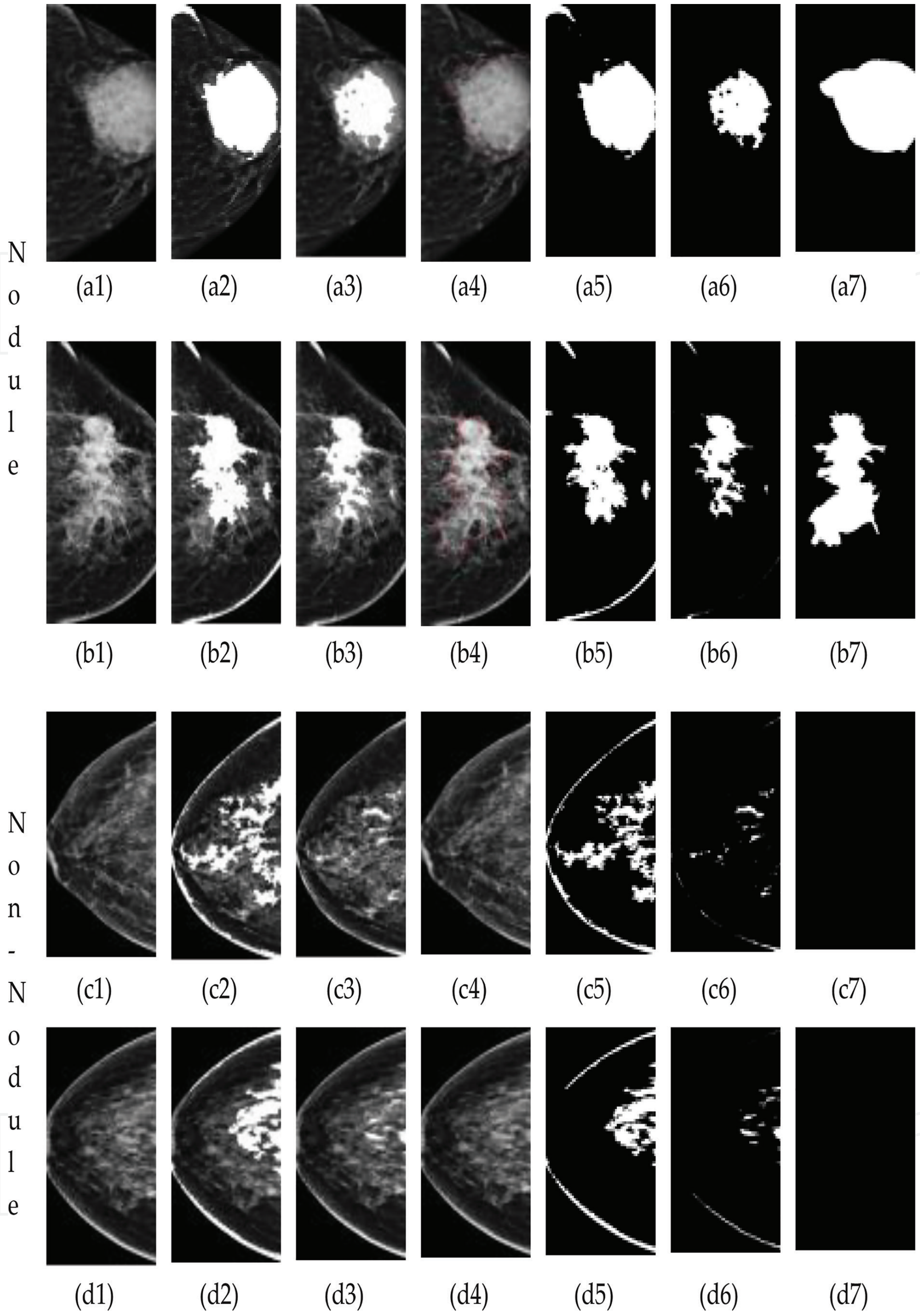


Figure 4. Result before and after using SVM model. (a1, b1, c1, and d1) original nonnodule and nodule images. (a2, b2, c2, and d2) segmentation result without using SVM rejection model, (a3, b3, c3, and d3) segmentation result after reducing the FP rate using SVM rejection model, (a4, b4, c4, and d4) ground truth images,. (a5, b5, c5, and d5) binary segmentation result without using SVM rejection model (a6, b6, c6, and d6) binary segmentation result after reducing the FP rate using SVM rejection model (a7, b7, c7, and d7) ground truth images.

images as well as the extraction of the initial contour was performed through MCWS, of which the proposed method comprises. The Chan-Vese algorithm is employed as the initial contour to enhance the result of the segmentation. The three steps of the SVM rejection model are in

	FP rate	SP	AC	NRM	OVERLAP
Without the rejection model	0.196	0.803	0.803	0.099	0.800
With the rejection Model (7 × 7 × 7 × 7)	0.058	0.941	0.938	0.031	0.933
With the rejection model (9 × 9 × 9 × 9)	0.051	0.948	0.944	0.028	0.940
With the rejection model (11 × 11 × 11 × 11)	0.044	0.955	0.950	0.025	0.946
With the rejection model (13 × 13 × 13 × 13)	0.040	0.959	0.954	0.023	0.950

Table 3. Quantitative analysis.

the following sequence: extracting teacher images, training the rejection model, and testing the model. Credence can be given to the MCWS algorithm in surmounting the challenges associated with the Chan-Vase algorithm. The Chan-Vese algorithm can be made more autonomous and converge faster by using a good initialization generated by MCWS.

Nevertheless, the reliance mammogram segmentation on the divergence and convergence of the intensity value of the image pixels is the constraint factor for this algorithm. The tendency has been toward segmenting the outlier component as part of the contour component, resulting in an incremental FP rate of the selected contour pixels. Accordingly, to overcome this issue, the SVM rejection model is geared toward reducing the FP rate. T-test was performed to determine the mean difference of two samples, that is, the accuracy before and after using rejection model with the best window size, which is (13 × 13). The T-test was applied to determine if there was a difference before and after applying the rejection model. The hypothesized mean difference of T-test was set to value 0, also named as null hypothesis. That means, assuming that there was no difference in the result whether using the rejection model. The alpha was set to value 0.05. The concept of T-test states that if the P value is less than the assumed alpha, the null hypothesis is not correct and there is a difference between the mean of the two samples. T-test result shows that the proposed method is considered statistically significant with ($P = 0.00001 < 0.05$). Furthermore, the proposed rejection models also showed less standard deviation (0.0001) and yields to stability in its performance. In general, this proposed method offers alternative decision-making ability and is able to assist the medical expert in giving second opinion on more precise nodule detection. Hence, it reduces FP rate that causes over segmentation.

4. Computer aided diagnosis for pathology

This section focuses on the histopathological grading step in the breast diagnosis, the procedure used to grade a certain tissue by examining the tissue slide biopsy, which must undergo a preparation step prior to the grading.

4.1. Tissue preparation

Breast tissue biopsy is a piece of tumorous tissue taken from the breast to investigate the occurrence of cancer. After the biopsy is extracted, it is enclosed in a fixative to prevent it from decaying. Then, the tissue is sectioned into fragile slices (e.g., 2–15 μm) using a microtome machine, which creates very thin slices. The slices are then arranged on the glass slide before being stained. The tissue is stained using certain pigments to reveal the tissue components (e.g., lumen, nuclei, cytoplasm, and stroma). This helps the pathologist to view the individual tissue component more clearly. This procedure is called cells marker. The pathologists use different methods of staining depending on the diagnostic process at hand. Among the common staining types, Hematoxylin and Eosin combination H&E is the most popular for diagnosis and grading. After staining the tissue slide, the pathologist evaluates the tissue slide using the microscope as in UKMMC or through a digital scanner used to produce digital pathology images. In UKMMC, a specific type of microscope (Olympus BX50 microscope) is used for the diagnosis [16]. This microscope has a camera to capture images of the region of interest. The next subsection will explain the image acquisition steps involved in the creation of the prostate and breast cancer data sets required for this study. Subsequent subsections will present a brief overview of the devices required for the image acquisition and image acquisition flow.

4.2. Image acquisition devices

In this study, prostate histological images were captured from tissue slides. All the images were viewed using an Olympus BX50 microscope (Olympus Corporation, Japan), and images were captured using a DP72 digital camera (Olympus Corporation) and cellSens Life Science imaging software, version 1.6 (Olympus Corporation) [16]. The sensitivity of the illumination source and camera's intensity were kept constant. The microscopes were adjusted manually to form clear magnified images, and the cameras were controlled through desktop computers to capture color digital images. Before image acquisition, the pathologists in UKMMC had selected the ROIs under the microscope. However, this requires substantial time and effort from pathologists, and more importantly, a subjective choice of the ROIs could introduce biases into the database and harm the generalizability of the developed computer CAD system.

4.3. Image acquisition work flow

Prior to acquiring the images, the microscope components, such as the light condenser, diffusing screen, and objective lens, were properly cleaned to remove any dust in the light path, which might badly affect the clarity of the acquired image. The focal plane was adjusted manually for clear images and was readjusted before every new image was taken. A light condenser was used to increase the light intensity for high-resolution image acquisition. To acquire an image from an ROI, the pathologist in UKMMC first reviewed the tissue section at a low magnification (e.g., 1× or 4×) to locate the ROI at the center of the image's field of view [16]. Usually, fine tuning is needed at higher magnification (40× magnification) to ensure a region with a typical Gleason pattern in the ROI is selected. The focal plane was then adjusted to produce a sharp image, and the light intensity was tuned so that the largest pixel value was slightly lower than the upper limit of the pixel's dynamic range. When all those adjustments were satisfactory, a still image was captured and saved onto the desktop computer as a color RGB digital image with a (tiff) extension. This process was repeated for all images that were captured for breast pathologists.

4.4. Self-collected data set from UKMMC

This data set contains self-collected breast tissue region images stained using the H&E procedure and captured from tissue slides of needle biopsies taken from 32 breast carcinoma cases. These tissue region images were digitized at 40× magnification, yielding high resolution images (4140 × 3096 pixels) in (tiff) format. The diagnosis assigned to each region image is based on the Bloom–Richardson grading system [16]. Each image was annotated as low grade (Grade 1) or high grade (Grade 3) by three expert pathologists from the HUKM center [16]. The total number of collected images is 100. These can be classified into 56 low-grade cases and 54 high-grade cases. **Figure 5** shows some sample images taken from this data set.

4.5. Ensemble learning of tissue components for histopathology image grading

This section explains the ensemble framework that we used for the classification of breast cancer and Gleason grading using the tissue components of the H&E histopathological region images. This project has been carried out from our previous work [16]. The framework is based on the ensemble learning approach from machine learning and medical tissue components (lumen, nuclei, cytoplasm, and stroma), both of which are of semantic meanings to pathologists. The framework extracts a set of textural features for each tissue component, which creates four independent sub data sets, and the diversity demonstrated by these data sets is then used to create an ensemble framework that is able to classify and grade breast cancer. Our framework consists of five phases: segmentation of four tissue components, feature extraction, feature selection, base classifiers of the framework, and ensemble fusion phase, as per **Figure 5**.

The typical CAD for breast cancer grading extracts features directly from histopathological images. Then, a single classifier is used to train these features to classify unknown patterns (e.g., image). Unlike this typical CAD, our project uses the concept ensemble learning (**Figures 5** and **6**).

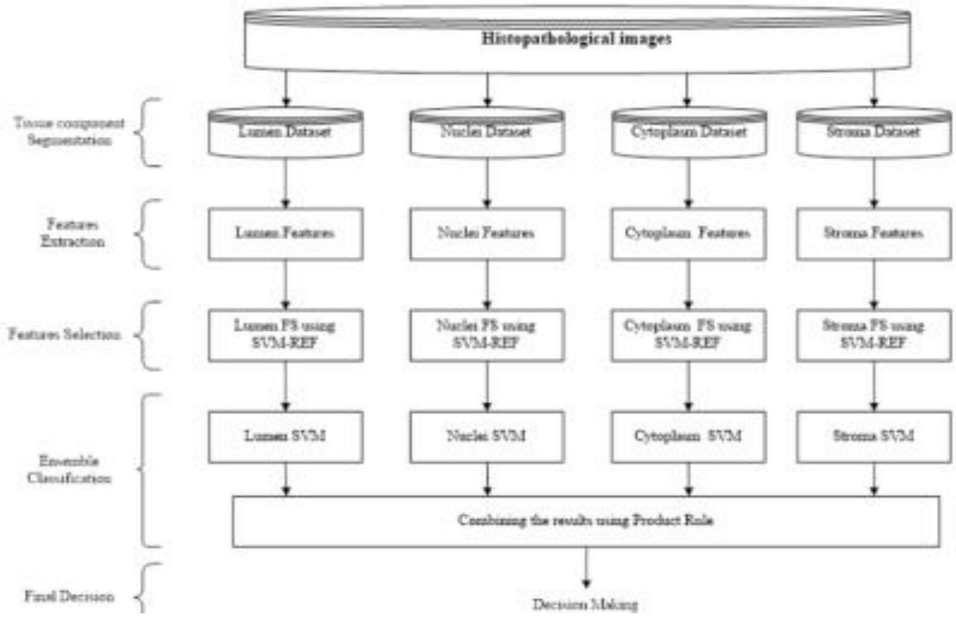


Figure 5. Ensemble framework for breast tissue image diagnosis and grading.

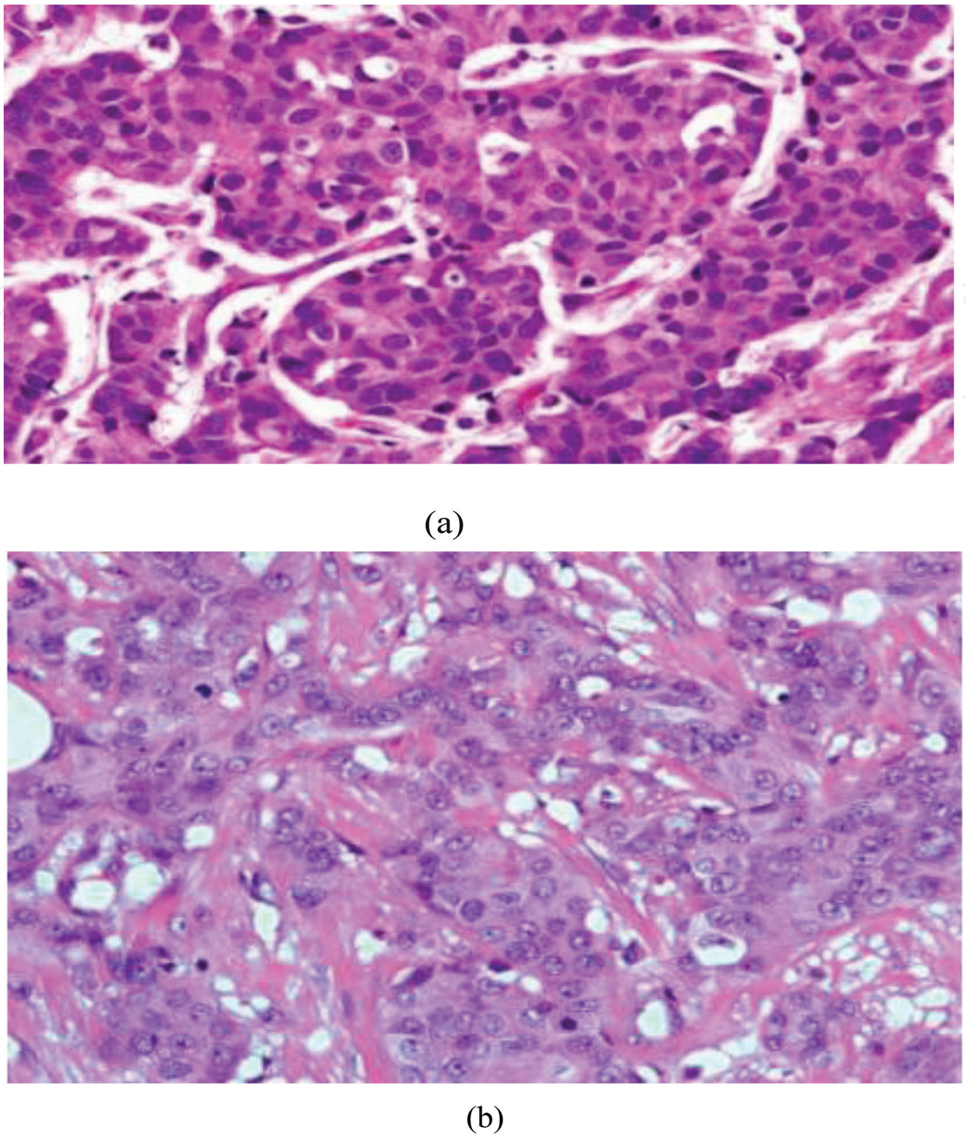


Figure 6. Two types of tissue classes of interest for the breast grading problem: (a) Grade 1 (low grade) tissue and (b) Grade 3 (high grade) tissue.

Due to the diversity of the tissue components, four different training data sets are created for the corresponding tissue components (lumen, nuclei, cytoplasm, and stroma). Thus, the diversity of the tissue components in ensemble learning is utilized to improve prostate diagnosis and grading. In this project, the ensemble framework, consisting of four-base SVM (RBF) classifiers, where each base classifier is a specialist, is trained to use the selected features of a particular tissue component. The decision function of SVM (RBF) with the top selected features (Ω) in the training model is defined as per (Eq. (7)):

$$(x\Omega) \;=\; \mathrm{sgn}\left\{(\mathrm{w}\;.\;x\Omega)\right\} \;=\; \mathrm{sgn}\left\{\propto i\;yi\;k(x\Omega,i,\,x\Omega)+b\;ni\;=\;1\right\}, \tag{7}$$

where $x\Omega$ is the test sample with only Ω corresponding features, $x\Omega$, ii is that of sample i in the training set ($i = 1, 2, \ldots, n$) with only Ω features, $\mathrm{y}_{\mathrm{i}} \in$ {1 $bening$, 0 $malignant$ (low $Grade\ or$ $high\ Grade$)} is the class label of the training sample $x\Omega$, i, and k is the kernel function that is used to calculate the inner product between the $\Phi\ x\Omega$, i and $\Phi(x\Omega)$ in the transformed space using nonlinear mapping Φ. The product rule Eq. (8) is utilized to produce the final decision for the proposed ensemble framework to combine the prediction outputs of all four base classifiers. The product rule is preferred in the ensemble when the single classifiers posterior probabilities are correctly estimated [16]. The final prediction (x) for the test image (x) based on product rule is computed using (Eq. (8))

$$class(x) = \max_{j=1}^{c=2} \prod \prod_{t=1}^{t=4} p_j^t(x) \tag{8}$$

4.6. Results and evaluation

In the ensemble framework, the stages of feature selection and classification are executed 50 times for each classification task. In each run, the data set of each base classifier (i.e., tissue component) is randomly divided into 50% training and 50% testing) after normalizing, as per [16]. It should be pointed out that in each run of the ensemble framework, similar numbers of selected features are used with all base classifiers. The base classifiers utilize the SVM with Radial-Basis-Function (RBF) kernel, while the SVM-RFE utilizes the linear SVM. To deploy RBF, one needs to set an appropriate value of the cost penalty, c, and gamma, γ. The grid search tool is one of the most common methods to identify suitable values for c and γ [1, 16]. The SVM implementation is utilized by the LibSVM toolbox [1, 16], while the C and γ in the SVM are estimated using a grid search with different internal threefold cross-validations on the training data set only from $\{2^{-20}, 2^{20}\}$. In this data set, the low vs. high grades classification task is dealt with, which is the most well-known task in state-of-the-art breast cancer analyses [1]. The results reported by this data set are shown in **Table 4**. As shown in **Table 4**, the proposed ensemble framework can effectively classify the low vs. high grades breast images. The AUC of low vs. high grade reached an average of 90.7%, which was greater than both the naïve and typical CAD. Moreover, when comparing the structure-method, the proposed method was far more superior. In using the proposed ensemble CAD, classification performance in the context of AUC can be substantially improved by 15% for the structure-based method. The results in **Figure 5** show that the ensemble framework was significantly quite accurate (90.8%) compared to the accuracy of each individual tissue components in the low vs. high grades in breast histopathology images. This framework has also been

Classification task Breast UKM	Measure	Proposed ensemble framework	Naïve approach	Typical CAD [22]	Significant of ensemble with	
					Naive	Typical CAD [22]
Low vs. high grade	AUC	90.7 ± 5.0	89.9 ± 4.8	89.8 ± 3.9	—	—
	Accuracy	90.8 ± 5.0	89.9 ± 4.8	89.8 ± 3.9	—	—
	Sensitivity	87.11 ± 8.4	87.1 ± 8.8	88.5 ± 7.7	—	—
	Specificity	94.3 ± 5.3	92.7 ± 6.3	91.1 ± 6.9	—	—

Table 4. The performance of the proposed ensemble framework on breast histopathology images data set.

validated using prostate and colon data set. Results proved that the ensemble framework can be utilized with other types of histopathology images if the main tissue components are visible in the image [7].

5. Discussion and conclusion

This chapter discusses how machine learning, particularly SVM can improve the performance for detection and diagnosing of breast cancer. SVM for now is one of the most powerful machine learning techniques that is able to model the human understanding of classifying data. It can find the relationship between data and segregates them accordingly. Using pixel values in mammogram images, SVM helps to improve the mass detection and segmentation of Chan-Vese algorithms by classifying correctly the false positive pixels. As a result, a sharper mass was detected with better estimation of its shapes and sizes. Hence, radiologist can give better diagnosis and biopsy location. Then, images of cell structure or tissue textures from the biopsy sample were examine by the pathologist. These pathology slides were analyzed under the pathologist sharp eyes to locate and identify any abnormal pattern of tissue texture or architecture. The process is tiring and subjective to the pathologist experience in interpreting the tissue condition. Thus, inter-observer and intra-observer variations exist. However, the proposed SVM algorithm can identify the different tissue component and model the pattern of relationship between these components spatially and statistically. The model is then used to grade any new pathology slides into its modified Bloom-Richardson grading, according to what the SVMs have learned from previous examples. Using the technique, it helps the radiologist and pathologist reducing their work load by automating the automation for decision making, especially for common and mundane cases. Radiologist and pathologist would have more time to spend on special or rare cases. The learning curve for young apprentice can

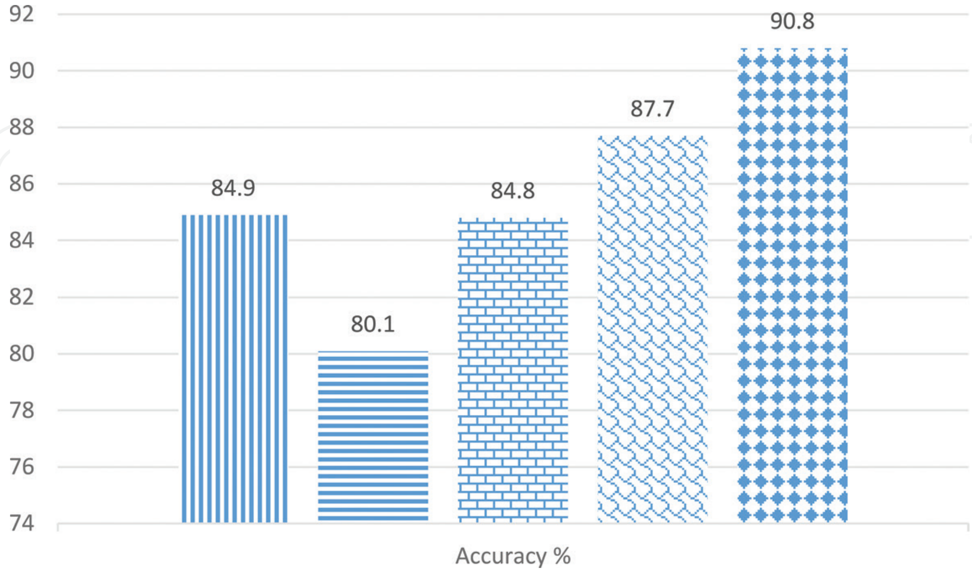


Figure 7. Single vs. ensemble classification results for low vs. high grade.

be steeper. The automate grading of breast cancer helps to reduce the variation of inter- and intra-observation by the pathologist. In our work, it should be noted that we are not using the identical patient data of mammogram and pathology due to some limitation. However, in the future it is possible to take the identical patient. Via the automatic decision making we are able to create a platform that integrate diagnostic reporting system that supports both specialties and, therefore, improves the overall quality of patient care (**Figure 7**).

However, combining these tissue components' features resulted in dense feature vectors, which suffers from overfitting. The use of the ensemble learning framework that allows prediction using several training subsets could help mitigate this problem. These different subsets are clearly shown in the proposed ensemble framework. The results indicate that proposed ensemble framework significantly outperformed the typical CAD, naïve approach, and structure-based method.

Acknowledgements

The authors thank the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, for providing the facilities and financial support under "ETP-2013-053 diagnostic services nexus for breast cancer." Gratitude is also due to Human Life Advancement Foundation (HLAF) for its financial support. Ethics approval was obtained entitled "FF-338-2012 Breast Cancer Diagnostic Imaging System Using Mammogram and Ultrasound" from UKM Medical Center, Malaysia, for collecting and studying breast cancer patient records.

Author details

Shahnorbanun Sahran[1]*, Ashwaq Qasem[1], Khairuddin Omar[1], Dheeb Albashih[2], Afzan Adam[1], Siti Norul Huda Sheikh Abdullah[1], Azizi Abdullah[1], Rizuana Iqbal Hussain[3], Fuad Ismail[4], Norlia Abdullah[5], Suria Hayati Md Pauzi[6] and Nurdashima Abd Shukor[6]

*Address all correspondence to: shahnorbanun@ukm.edu.my

1 Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia

2 Computer Science Department, Prince Abdullah Bin Ghazi Faculty of Information Technology, Al-Balqa Applied University, Jordan

3 Department of Radiology, Universiti Kebangsaan Malaysia Medical Center, Malaysia

4 Department of Oncology, Universiti Kebangsaan Malaysia Medical Center, Malaysia

5 Department of Surgeon, Universiti Kebangsaan Malaysia Medical center, Malaysia

6 Department of Pathology, Universiti Kebangsaan Malaysia Medical Center, Malaysia

References

[1] Ashwaq Q, Siti Norul Huda SA, Shahnorbanun S, Rizuana IH, Fuad I. An accurate rejection model for false positive reduction of mass localisation in mammogram. Pertanika Journal of Science and Technology. 2017;**25**(S):49-62

[2] The American Cancer Society. How Common Is Breast Cancer? 2017. Available from: https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html [Accessed: 22-01-2018]

[3] James S, Denise RA, Dena E, Silvana L, Ossama T, Dean WW. Integrating pathology and radiology disciplines: An emerging opportunity? BMC Medicine. 2012;**10**:100

[4] Ebert J, Xu Y, Smith G, Shen Y, Jiang J, Buchholz T, Hunt K, Black D, Giordano GW, Yang W, Shen C, Elting L, Smith B. Surgeon influence on use of needle biopsy in patient with breast cancer: A national medicare study. Journal of Clinical Oncology. 2014; **32**(21):2206-2216

[5] Adepoju L, Qu W, Kazan V, Nazzal M, Williams M, Sferra J. The evaluation of national time trends, quality of care and factors affecting the use of minimally invasive breast biopsy and open biopsy for diagnosis breast lesions. American Journal of Surgery. 2014;**208**(3):382-390

[6] Wan T, Cao J, Chen J, Qin Z. Automated grading of breast cancer histopathology using cascaded ensemble with combination of multi-level image features. Journal of Neurocomputing. 2017;**229**(C):34-44

[7] Sampat MP, Markey MK, Bovik AC. Computer-aided detection and diagnosis in mammography. Handbook of Image and Video Processing. 2005;**2**(1):1195-1217

[8] Anju J. Machine learning techniques for medical diagnosis: A review. In: 2nd International Conference on Science, Technology and Management. New Delhi; 27 September 2015

[9] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015;**349**:255-260

[10] Marc K, Luciano MP, Ross WF, Geis JR. Implementing machine learning in radiology practice and research. American Journal of Roentgenology. 2017;**208**(4):754-760

[11] Afzan A,Khairuddin O. Computerized breast cancer diagnosis with Genetic Algorithm and Neural Network. In: Proceedings of the 3rd International Conference on Artificial Intelligence and Engineering Technology (ICAIET), Universiti Malaysia Sabah; 22-24 November 2006. pp. 533-538

[12] Shahnorbanun S, Albashish D, Azizi A, Nordashima AS, Suria HMP. Absolute cosine-based SVM-RFE feature selection method for prostate histopathological grading. Artificial Intelligence in Medicine. 2018;**87**:78-90

[13] Azizi A. Supervised Learning Algorithms for Visual Object Categorization. Netherlands: Universiteit Utrecht; 2010. ISBN: 978-90-393-5440-7

[14] Nemoto M, Masutani Y, Nomura Y, Hanaoka S, Miki S, Yoshikawa T, Hayashi N, Ootomo K. Machine learning for computer-aided diagnosis. Igaku Butsuri. 2016;**36**(1):29-34

[15] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;**20**(3):273-297

[16] Dheeb AA. Thesis of Embedded feature selection methods based on support vector machine for histopathology grading. Malaysia: Universiti Kebangsaan; 2017

[17] Pham DL, Xu C, Prince J. Current methods in medical image segmentation. Annual Review of Biomedical Engineering. 2000;**2**(1):315-337

[18] Cheng H, Shi X, Min R, Hu L, Cai X, Du H. Approaches for automated detection and classification of masses in mammograms. Pattern Recognition. 2006;**39**(4):646-668

[19] Brett J, Bankhead C, Henderson B, Watson E, Austoker J. The psychological impact of mammographic screening. A systematic review. Psychooncology. 2005;**14**:917-938

[20] Bond M, Pavey T, Welch K, Cooper C, Garside R, Dean S, et al. Systematic review of the psychological consequences of false-positive screening mammograms. Health Tech_nology Assessment (Winchester). 2013;**17**:1-170, v-vi

[21] Lerman C, Trock B, Rimer BK, Boyce A, Jepson C, Engstrom PF. Psychological and behavioral implications of abnormal mammograms. Annals of Internal Medicine. 1991; **114**:657-661

[22] Román M, Castells X, Hofvind S, von Euler-Chelpin M. Risk of breast cancer after false-positive results in mammographic screening. Cancer Medicine. 2016;**5**(6):1298-1306

[23] Roman R, Sala M, Salas D, Ascunce N, Zubizarreta R, Castells X. Effect of protocol-related variables and women's characteristics on the cumulative false-positive risk in breast cancer screening. Annals of Oncology. 2012;**23**:104-111

[24] Elmore JG, Miglioretti DL, Reisch LM, et al. Screening mammograms by community radiologists: Variability in false-positive rates. Journal of the National Cancer Institute. 2002;**94**:1373-1380

[25] Sala M, Salas D, Belvis F, et al. Reduction in false-positive results after introduction of digital mammography: Analysis from four population-based breast cancer screening programs in Spain. Radiology. 2011;**258**:388-395

[26] Utzon-Frank N, Vejborg I, von Euler-Chelpin M, Lynge E. Balancing sensitivity and specificity: Sixteen years of experience from the mammography screening programme in Copenhagen, Denmark. Cancer Epidemiology. 2011;**35**:393-398