Expert Systems With Applications 119 (2019) 90–103

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Review

Is mass classification in mammograms a solved problem? - A critical review over the last 20 years

Ricardo Wandré Dias Pedro[a,b,*], Ariane Machado-Lima[c,d], Fátima L.S. Nunes[c,d,a,b]

[a] Electrical Engineering, Polytechnic School, University of São Paulo, Brazil
[b] Av. Professor Luciano Gualberto, travessa 3, 158 - Cidade Universitária - São Paulo/SP - 05508-010, Brasil
[c] Information Systems, School of Arts, Sciences and Humanities, University of São Paulo, Brazil
[d] Av. Arlindo Béttio, 1000 - Ermelindo Matarazzo - São Paulo/SP - 03810-000, Brasil

A R T I C L E I N F O

Article history:
Received 14 July 2018
Revised 29 September 2018
Accepted 15 October 2018
Available online 16 October 2018

Keywords:
Mammography
Mammogram
Breast cancer
Classification
Diagnosis
Pattern recognition

A B S T R A C T

Breast cancer is one of the most common and deadliest cancers that affect mainly women worldwide, and mammography examination is one of the main tools to help early detection. Several papers have been published in the last decades reporting on techniques to automatically recognize breast cancer by analyzing mammograms. These techniques were used to create computer systems to help physicians and radiologists obtain a more precise diagnosis. The objective of this paper is to present an overview regarding the use of machine learning and pattern recognition techniques to discriminate masses in digitized mammograms. The main differences we found in the literature between the present paper and the other reviews are: 1) we used a systematic review method to create this survey; 2) we focused on mass classification problems; 3) the broad scope and spectrum used to investigate this theme, as 129 papers were analyzed to find out whether mass classification in mammograms is a problem solved. In order to achieve this objective, we performed a systematic review process to analyze papers found in the most important digital libraries in the area. We noticed that the three most common techniques used to classify mammographic masses are artificial neural network, support vector machine and k-nearest neighbors. Furthermore, we noticed that mass shape and texture are the most used features in classification, although some papers presented the usage of features provided by specialists, such as BI-RADS descriptors. Moreover, several feature selection techniques were used to reduce the complexity of the classifiers or to increase their accuracies. Additionally, the survey conducted points out some still unexplored research opportunities in this area, for example, we identified that some techniques such as random forest and logistic regression are little explored, while others, such as grammars or syntactic approaches, are not being used to perform this task.

1. Introduction

According to the World Health Organization (WHO)[1] breast cancer is one of the most common type of cancers among women. Each year, more than 1.5 million women suffer from this disease, which causes the greatest number of cancer-related deaths. In 2015, about 570,000 women died due to breast cancer, which represents 15% of all cancer deaths among women. Surveillance, Epidemiology, and Results Program (SEER)[2] provides statistics based on the US population, and estimates that in 2017 there were more than 250,000 new cases, representing 15% of all new cancer cases. Also, SEER estimates more than 40,000 deaths are related to breast cancer, which represents 6.8% of all cancer deaths in 2017.

Mammogram X-ray is considered the most reliable and effective method in early detection of breast cancer at an early stage (Li, Meng, Wang, Tang, & Yin, 2017). Although there are some rules to differentiate between benign and malignant cases, only 15 - 30% of the masses referred to surgical biopsy are malignant (Mohanty, Senapati, Beberta, & Lenka, 2013). Performing biopsies in unnecessary situations can lead to several problems such as the financial cost of the procedure, the physical pain women are submitted to, and the severe anxiety until the final diagnosis is confirmed (Keleş, Keleş, & Yavuz, 2013; Todd & Naghdy, 2004).

To help physicians compose a more precise diagnosis, several computer-aided detection (CAD) and computer-aided diagnosis (CADx) systems have been proposed over the last decades to,

* Corresponding author.
E-mail addresses: rwandre@usp.br (R.W.D. Pedro), ariane.machado@usp.br (A. Machado-Lima), fatima.nunes@usp.br (F.L.S. Nunes).

1 World Health Organization: http://www.who.int/en/.

2 Surveillance, Epidemiology, and Results Program: https://seer.cancer.gov.

https://doi.org/10.1016/j.eswa.2018.10.032
0957-4174/

Table 1
Digital libraries used to find the studies.

Library	String	Date	Number of papers
PubMed	mammogra* AND (classi* OR recognition) AND (mass OR nodule)	September, 2016	945
Periódicos Capes	mammogra* AND (classi* OR recognition) AND (mass OR nodule)	September, 2016	1020
PubMed	mammogra* AND (classi* OR recognition) AND (mass OR nodule)	May, 2017	62
IEEE Xplorer	("Document Title":mammogra* AND (classification) AND (mass OR nodule) NOT (calcification OR microcalcification))	July, 2017	182
Elsevier	title(mammogra* AND (classification) AND (mass OR nodule) AND NOT (calcification OR microcalcification))	July, 2017	25
Springer Link	mammogram AND classification AND (mass OR nodule) AND NOT (calcification AND microcalcification)	July, 2017	46
SPIE Digital Library	mammogram AND (mass OR nodule) AND (classification) NOT (calcification OR microcalcification)	July, 2017	37
Medical Physics	mammogram AND (mass OR nodule) AND (classification) NOT (calcification OR microcalcification) in Article Titles	July, 2017	19

respectively, detect and classify findings in the mammograms. Many pattern recognition and machine learning techniques have been developed, employed and published by academic researchers and industrial companies.

In this paper the objective is to present and discuss the results of a systematic review to identify the state of the art, and possible gaps in the area of classification of masses in mammograms.

Authors of Cheng et al. (2006) and Hadjiiski, Sahiner, and Chan (2006) also performed a literature review with similar goals. The main difference between these studies and this present paper, is that this paper is more focused on mass classification using mammographic images (x-ray), which allows a deeper analysis of a greater number of papers. Here we present a summary of the 129 papers, pointing out which techniques were used in the classification process, which features were used, which methods were employed to select the most discriminative features, which classes were used to classify the masses and the results achieved. Furthermore, we analyzed the methods used in the training and tests phases, which metrics were used, the number of images the researchers used in their experiments, and which and how many databases were used. The cited reviews are more generic, merely touching the areas of mass segmentation, feature extraction and selection, the use of magnetic resonance imaging, and ultrasound images.

Besides this introduction, this paper is divided into the following sections: Section 2 describes the research method used in this systematic review. Section 3 contains a quantitative global analysis of the studied papers. In Section 4 we present a brief overview of the papers analyzed. In Section 5 we discuss our findings and point out some research gaps in this area. Section 6 outlines the final conclusion of this work.

2. Research method

The systematic review conducted in this paper is divided in three phases: 1) planning of a protocol used as research guidelines; 2) searching and selection of the studies of interest according to pre-defined inclusion and exclusion criteria defined in the protocol; and 3) analysis of the selected papers to understand the state of the art in this research area.

In our protocol we state the following research questions:

- What are the techniques used to classify masses in mammograms?
- What are the features used as input to the classifiers?
- What are the results achieved by the several classifiers and features used?

To answer these questions, we searched for papers in two different timelines. The first one was in September 2016, followed by a new search in May and July 2017.

We created strings with the words **mammogra* classi* recognition mass nodule.** We also excluded the word 'microcalcification'. Next, we queried the search engines PubMed[3], Periódicos Capes[4], IEEE Xplorer Digital Library[5], Springer Link[6], Elsevier[7], SPIE Digital Library[8], and Medical Physics[9].

The criterion used to include a paper was:

- papers that undertake the problem of breast mass classification in an automatic or semi-automatic way using mammograms;

The criteria used to exclude a paper were:

- papers that exclusively address the problem of detecting breast mass in mammograms;
- papers that focus on presenting techniques used to segment mammograms;
- papers that address problems of classification in mammograms, but are not related to masses, for example, microcalcifications;
- papers that used other medical image modalities such as ultrasound or magnetic resonance;

It should be noted that papers that reported using a new approach to segment mammograms were not excluded if they also performed mass classification after the segmentation technique was applied.

After removing the duplicated papers and applying the inclusion/exclusion criteria defined above we included 129 papers to this systematic review. Table 1 shows the strings used in each digital library. The process used to select the papers is summarized in Fig. 1.

3. Global analysis

3.1. Published papers per year. Fig. 2 shows that this research area has presented an increasing number of published papers over the last two decades. Also, the majority of the papers we analyzed were published in the last decade (98 papers - 76%). This increase of published papers can be the result of more computational power, more sophisticated algorithms to extract the features and perform the classification, and the increased number of available images to be used in the studies. As we performed our search in the middle of 2017, we are not considering the entire year of 2017, but we can notice that the number of published papers on this

[3] PubMed: https://www.ncbi.nlm.nih.gov/pubmed.
[4] Periódicos Capes: https://www.periodicos.capes.gov.br.
[5] IEEE Xplorer Digital Libray: http://ieeexplore.ieee.org/Xplore/home.jsp.
[6] Springer Link: https://link.springer.com.
[7] Elsevier: http://www.sciencedirect.com.
[8] SPIE Digital Libray: https://www.spiedigitallibrary.org/?SSO=1.
[9] Medical Physics: http://aapm.onlinelibrary.wiley.com/hub/journal/10.1002/(ISSN)2473-4209/.

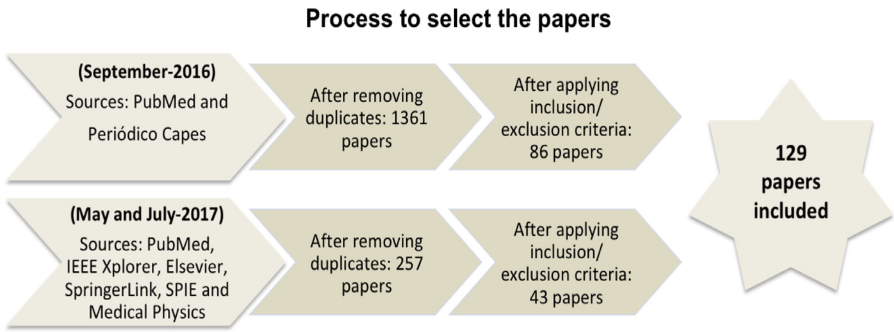

Fig. 1. The process to select the papers that were included in this systematic review.

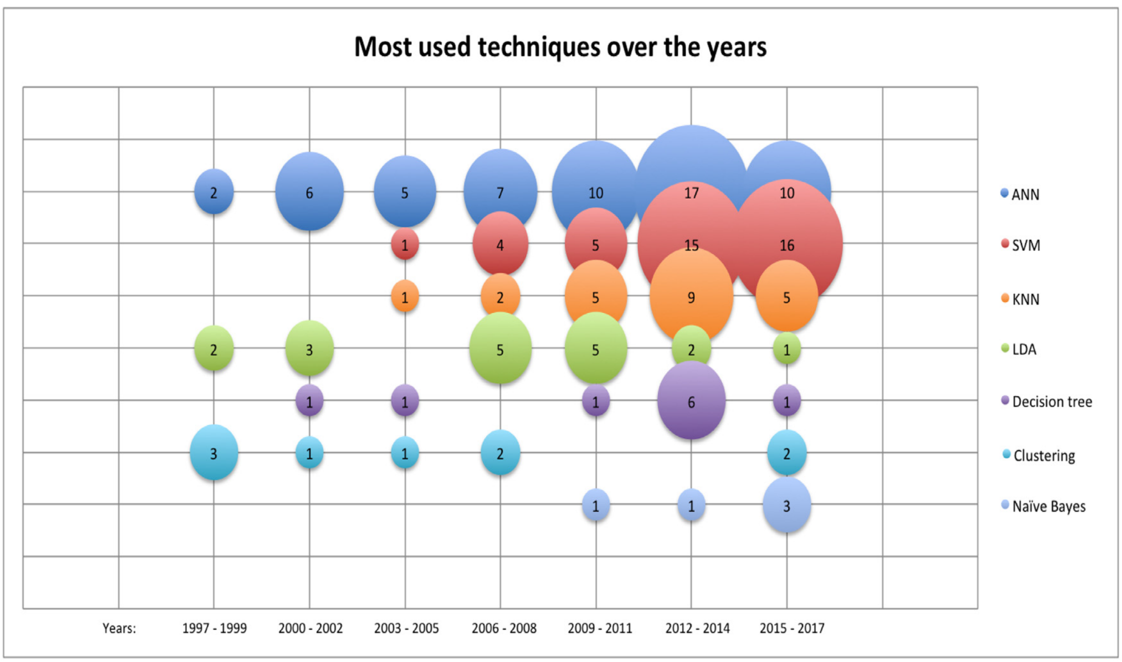

Fig. 2. Most used techniques used to discriminate masses over the years.

topic has been decreasing since 2015. Yet, Fig. 2 shows that this research area is dominated by the use of artificial neural networks, support vector machine and K-nearest neighbors, specially in the last decade.

3.2. Databases. Fig. 3 shows the most common databases used in the studies. The two most used were Digital Database for Screening Mammography (DDSM) cited in 49 papers (38%) and the Mammographic Image Analysis Society (MIAS) database that appeared in 39 studies (30%). Yet, 15 studies (12%) did not mention where the analyzed images came from. DDSM is a public database provided by the University of South Florida that contains more than 2,500 studies (each study includes two images of each breast) while MIAS database is also a public database that contains 322 images (161 pairs). These two databases are the most common as they provided a great number of images and are free to use provided the license agreements are respected. On the other hand, some studies used private databases, such as the database provided by Alberta Program for the Early Detection of Breast Cancer and the database provided by University of Chicago. Private databases tend to appear less frequently in the studies because it is more difficult to have access to them.

It was observed that the majority of the papers used just one database (88 papers - 77%), and only three papers (3%) of the analyzed studies used images from three or more sources. It is important to mention that 15 papers (12%) did not specify how many databases they used to obtain the images.

Fig. 4 shows the number of images used in the studies. We can see that 28 papers (22%) used 100 or fewer images to train the classifiers and perform the tests. Also, 30 papers (23%) used between 101 and 200 images and, 34 papers (26%) used between 200 and 500 images. Furthermore, six papers (5%) did not specify the number of used images. It was observed that 45% of the papers use 200 or less images. On the one hand, it shows that many images are not really necessary in order to train the classifiers with relation to this problem. For instance, McLeod and Verma (2013) used 200 images and achieved the lowest and highest accuracies of 93% and 98%, respectively. On the other hand the built classifiers may not be generic enough to deal with a variety of other images that were not used in the study.

3.3. Features selection. Several techniques were employed to select the most powerful features to discriminate masses. The genetic algorithm was the most commonly used (6% of the papers), for instance (Chaieb, Bacha, Kalti, & Lamine, 2014; Rouhi, Jafari, Kasaei, & Keshavarzian, 2015), followed by selection procedures that used Wilks' lambda criterion that appeared in 5% of the papers - (Li et al., 2008; Shi et al., 2008), and Principal Component Analysis

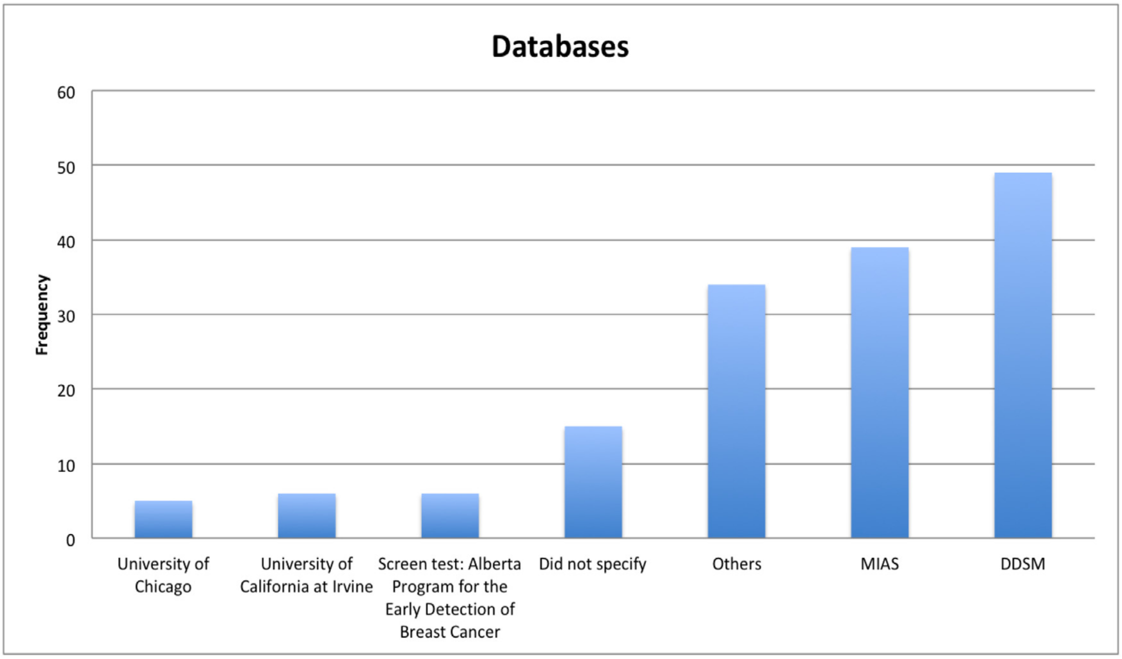

Fig. 3. Most frequent databases used in the analyzed studies.

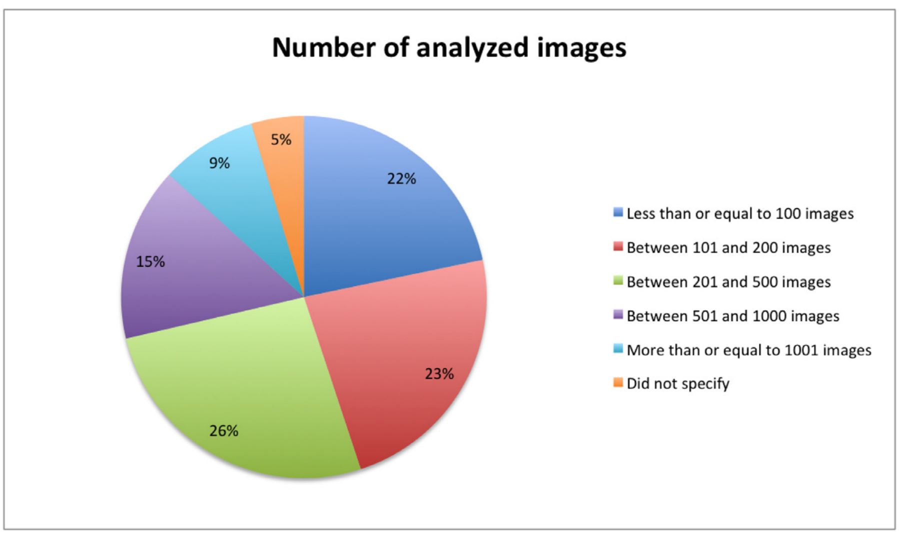

Fig. 4. Number of analyzed images per paper.

(PCA) also employed in 5% of the studies - (Chaieb et al., 2014; Muramatsu, Hara, Endo, & Fujita, 2016).

3.4. Classification techniques. During our analysis, we reviewed the use of many pattern recognition and machine learning techniques to classify the masses in mammograms. Fig. 2 shows the most common techniques used in the analyzed papers. Artificial Neural Network (ANN) with its variations (57 papers - 44%) and Support Vector Machine (SVM) with its variations (41 papers - 34%) are the two most commonly employed techniques. K-Nearest neighbors, the third most frequent technique appeared in 22 papers (17%), followed by linear discriminant analysis (18 papers - 14%) and decision trees (10 papers - 8%). Moreover, several other techniques appeared in the studies, but not as frequently. An analysis of the papers is displayed in Section 4.

3.5. Validation and test techniques. Holdout approach was the most used validation and test technique cited in 29% of the papers. Next, both k-fold cross validation and leave-one-out appeared in 27% of the studies. Fifteen of the analyzed papers (12%) did not mention how the dataset was divided for training and tests. Holdout was the most employed technique probably because it is the fastest method to use. It consists of splitting the dataset into two parts, one for training and the other for testing. Typically, the training set is bigger than the test set. Fig. 5 shows that when there is a small number of images (less than 100 images) the researchers prefer to use leave-one-out technique, but as the number of images increases they tend to use holdout or k-fold cross validation. Bootstrapping and resubstitution were barely used.

3.6. Evaluation metrics. Accuracy, the most common metric used to validate a study, appeared in 80 papers (62%). The following most common metrics were the area under the ROC curve (AUC), sensitivity and specificity, which appeared in 74 papers (57%), 38 papers (29%), and 37 papers (29%), respectively. There were others

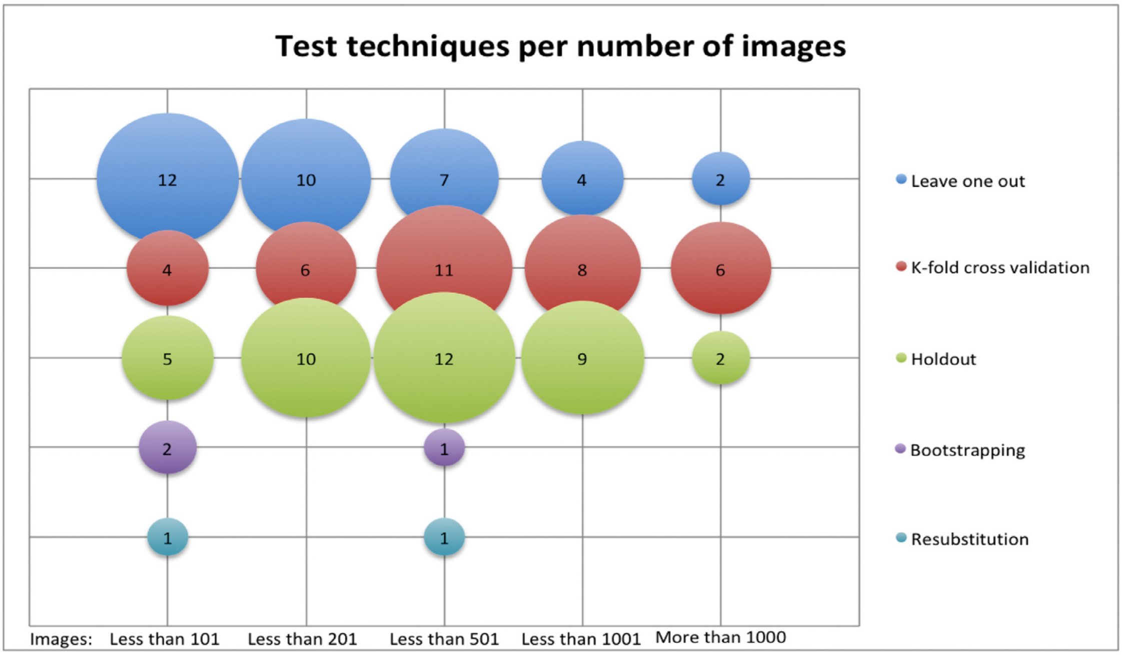

Fig. 5. Techniques used to test and validate the proposed approaches.

Table 2
Definition of metrics used to validate the studies.

Metrics	Definition
Accuracy (or classification rate)	$\frac{TP+TN}{P+N}$
Sensitivity (or true positive rate)	$\frac{TP}{P}$
Specificity (or true negative rate)	$\frac{TN}{N}$
Precision (or positive predictive value)	$\frac{TP}{TP+FP}$
Negative predictive value	$\frac{TN}{TN+FN}$
Fall-out (or false positive rate)	$\frac{FP}{N}$
False negative rate	$\frac{FN}{P}$
Matthews correlation coefficient	$\frac{TP.TN-FP.FN}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}}$
Area under the ROC curve (AUC)	A ROC curve is plotted considering the sensitivity in function of the 1 - specificity. Every point on the ROC curve represents a pair (sensitivity/specificity) corresponding to a particular decision. The AUC is used to measure how well a parameter is able to distinguish between two classes (benign vs. malignant, diseased vs. normal, and so on)

metrics in the analyzed papers, but they did not appear frequently. In Table 2 we can see a definition of each metric, where P is the number of positive examples, N is the number of negative examples, TP is the true positive results, TN is the true negative results, FP is the false positive results and FN is the false negative results.

3.7. Classes. One hundred papers discriminate masses considering only the benign and malignant classes (78%). Also, there were papers that consider some of the following classes: normal, round, lobular, oval, irregular, stellate, and BI-RADS classes.

4. Pattern recognition techniques

In this section we present an analysis of the papers included in this systematic review. We grouped the studies considering the technique used to discriminate masses, i.e., artificial neural network, support vector machine, decision trees and so on. From each paper, we extracted the techniques they used, the features that were employed, the methods to select the most powerful features, the classes used in the classification, and the results they achieved.

4.1. Artificial neural networks. From the 129 papers analyzed in this systematic review, 29% (38 papers) used ANN as the unique classifier to discriminate mammographic masses.

We can observe that 29% (11 from 38 papers) employed some technique to select the most discriminant features while 71% did not (Table S2 in supplementary material). The most common techniques to select the most important features were genetic algorithms (Bhattacharya, Sharma, Goyal, Bhatia, & Das, 2011; Suganthi & Madheswaran, 2009; 2012; Tan, Pu, & Zheng, 2014a) and correlation based on feature selection (Delogu, Fantacci, Kasae, & Retico, 2007; Nugroho, Faisal, Soesanti, & Choridah, 2014a; 2014b).

When analyzing the features employed, it can be seen that there are usually three types: shape, texture and information provided by radiologists. Examples of shape features are area, perimeter, circularity, Zernike moments and so forth. As for texture features, there are features extracted from gray level co-occurrence matrices, density, contrast, local binary patterns, wavelets and so on. The information provided by radiologists is patient age and BI-RADS. However there are several papers that have used all three types of features together or at least a combination of two types. For instance, in Laroussi, Ayed, Masmoudi, and Masmoudi (2013) Zernike moments and local binary patterns were used to classify masses as benign or malignant and they achieved an AUC of 0.96. Yet, considering the papers in Table S2 we noticed that Jiao, Gao, Wang, and Li (2017) did not use some of these well known features, instead they used features learned through convolutional neural networks and achieved an accuracy of 97.4%.

The most common classes used in the classification process are benign and malignant (79% - 30 papers). However, some papers also use the class normal and perform a classification using three classes (benign, malignant and normal), for example Edwards (2004); Nugroho, Faisal, Soesanti, and Choridah (2014b). The authors of Chokri and Farida (2017) not only used benign and malignant classes, but also used BI-RADS classes (2, 3, 4 and 5) to perform the classification. Yet, in Chokri and Farida (2017), when the classification is performed using more classes the accuracy of the classifier decreases.

In terms of results, it is quite difficult to compare all the papers described in Table S2, since they used many different databases, different images, different methods to perform the tests, and they also presented the results in different ways. The most common ways of presenting the results are in the form of accuracy and the area of the ROC curve. In order to explore this point, we selected some papers that are quite similar in order to allow a comparison as fair as possible. Section 5 provides this comparison.

In terms of classifier, many different types/architectures of artificial neural networks were used. For example, single layer and multi-layer perceptron was used in Tralic, Bozek, and Grgic (2011) and, in this case, the single layer perceptron achieved the best results in the classification problem. The other types of ANN and the papers that used them are the following: 1) convolution neural networks (Jiao et al., 2017); 2) probabilistic ANN (Patil, Udupi, & Bhogale, 2013; Yang et al., 2005); 3) Bayesian ANN (Edwards, 2004; Edwards, Lan, Metz, Giger, & Nishikawa, 2003; Li et al., 2008); 4) resilient ANN (Serifovic-Trbalic, Trbalic, Demirovic, Prljaca, & Cattin, 2014); 5) an ensemble of ANN incorporating k-means (Leod & Verma, 2012); 6) fuzzy ANN (Azevedo et al., 2015a; Keleş et al., 2013; Rathi & Aggarwal, 2014); and 7) an adaptive neuro-fuzzy inference system (Bhattacharya et al., 2011; Mousa, Munib, & Moussa, 2005).

Fig. 2 shows the number of published papers that used ANN as classifier over the years. Between 2012 and 2014, this number achieved its peak, and then declined during 2015 to 2017. It should be noted that this fact does not mean the researchers have stopped using ANN, but rather that they are trying new approaches to solve the problem of discriminating masses in mammography.

4.2. Support vector machines. SVM was employed in 16% (21 papers) of the analyzed papers as the unique classifier and it was the second most used technique to classify masses. Table S3 in the supplementary material summarizes all these papers.

From these 21 papers, 8 (38%) used some feature selection techniques. While genetic algorithm is the most employed technique in papers that used only ANN, for SVM the most used techniques to select the most appropriate features are PCA and LDA (Hussain, Khan, Muhammad, & Bebis, 2012; Khan, Hussain, Aboalsamh, & Bebis, 2017). Genetic algorithm appeared only in Azizi, Zemmal, Sellami, and Farah (2014).

Gabor features appeared in three papers of the same research group (Hussain et al., 2012; Khan et al., 2017; Khan et al., 2016). These papers perform comparison between different Gabor features and different ways of extracting them. The results can be seen in Table S3. The other paper that used Gabor features was Abdel-Nasser, Melendez, Moreno, and Puig (2016), but it combined these features with histogram of oriented gradient, local binary patterns, local directory number, and Haralick's features, achieving an AUC of 0.78.

In general terms, the features are pretty much the same ones used in papers that employed ANN. However, some papers reported using different features such as measures extracted from Ripley's K function in Oliveira Martins, da Silva, Silva, de Paiva, and Gattass (2007), Moran's index and Geary's coefficient in Junior, de Paiva, Silva, and de Oliveira (2009) and central and Hu moments in Azizi et al. (2014). In these three papers the best accuracy was of 94.94% achieved by Oliveira Martins et al. (2007). Additionally, Jamieson, Drukker, and Giger (2012) used features learned by using adaptive deconvolutional networks and achieved an AUC of 0.71, while Ovalle, Gonzlez, Ramos-Polln, Oliveira, and Guevara-Lpez (2016) employed convolutional neural networks to learn the features obtaining an accuracy of 94%.

The majority of papers (18 papers - 85.7%) performed the classification using benign and malignant as the possible classes. The exceptions were Cheikhrouhou, Djemal, Sellami, Maaref, and Derbel (2008), which used not only benign and malignant but also four BI-RADS classes; Khan et al. (2016) which used mass, non-mass, benign and malignant classes; and Kanadam and Chereddy (2016) that used as classes calcifications, circumscribed, spiculated, ill-defined, architectural distortions, asymmetry(s) and normal.

4.3. K-nearest neighbors. KNN as a unique classifier to distinguish masses in mammograms appeared in 6% (8 papers) which was the third most commonly used technique. An overview of these research projects is shown in Table S4 in the supplementary material.

Just one of these papers (12%) focus on the problem of selecting the most suitable features, but deal with it using the classifier error with a sequential forward procedure, achieving an AUC of 0.79 when worked with a single image view and 0.84 in a case based view (Varela, Muller, & Karssemeijer, 2003).

Studies regarding the use of texton histograms as features were presented in Li, Chen, Rohde, Yao, and Cheng (2015); Li, Chen, Wei, Peng, and Cheng (2016) to classify masses as benign and malignant and the AUC obtained for these research projects were 0.92 and 0.91, respectively. The other papers used more common features such as shape features, wavelets, features extracted from gray level co-occurrence matrix, and other features.

Eltoukhy, Faye, and Samir (2010) was the only study that considered more classes than benign and malignant. The authors classified the region of interest (ROI) as benign, malignant, normal and diverse abnormalities. Regarding only normal, benign and malignant classes, the highest classification rate was 94.07% using curvlets as features.

As seen in Table S4, the number of published papers that used only KNN (8 papers) as a classifier is small when compared with the number of papers that used only ANN (38 papers) or only SVM (21 papers). However, in Fig. 2 it can be seen that KNN is still being used, especially after the year of 2009 (19 papers of studies that used only KNN or used KNN together with other techniques).

4.4. Clustering. Clustering was applied in seven of the analyzed papers (5%) to classify masses. A summary of each paper can be seen in Table S5 in the supplementary material.

From these studies, Bruce and Adhami (1999); Bruce and Kallergi (1999); Bruce, Kallergi, and Mendoza (1999) used Euclidean distance to separate the masses among the classes. All these papers belong to the same research group, which separated the masses in the following classes: round, lobular, irregular. The highest classification rate in these papers was 80% obtained in Bruce and Adhami (1999).

Mahalanobis distance was used in Azevedo et al. (2015b); Mudigonda, Rangayyan, and Desautels (2000); Santaella, Schiabel, Patrocinio, Nunes, and Romero (2003). In

Azevedo et al. (2015b) the authors used Mahalanobis distance with morphological extreme learning machines and disclosed the results in terms of kappa index (0.66) considering benign, malignant and normal classes. The work in Mudigonda et al. (2000) achieved an accuracy of 85% but considered only benign and malignant classes, while Santaella et al. (2003) achieved the best AUC of 0.97 separating the masses as spiculated and circulate.

In Meriem, Merouani, and Lakhdar (2015) fuzzy c-means were used to separate masses as benign, malignant and normal. The features were based on Zipf curves and the accuracy obtained was 87%.

Analyzing Table S5 we see that few papers were published throughout the years that solved the problem of mass classification using clustering. All papers that used Euclidean distance belong to the same research group and were published in 1999. In 2015 we found two papers published, one used Mahalanobis distance with morphological extreme learning machines and the other used fuzzy c-means.

4.5. Linear discriminant analysis. The fifth most used technique was linear discriminant analysis in seven papers (5%). Table S6 in the supplementary material shows a summary of each one of these papers.

Three of these papers belong to the same research group (Hadjiiski et al., 2001; Sahiner, Chan, Petrick, Helvie, & Goodsitt, 1998; Shi et al., 2008). In Sahiner et al. (1998) the authors used a rubber band straightening transform and texture analysis to classify masses as benign and malignant. In Hadjiiski et al. (2001) an analysis was performed on temporal changes of mammographic features and in Shi et al. (2008) the authors performed a classification based on level set segmentation and patient information. The highest AUC achieved was 0.94 in Sahiner et al. (1998), that is the oldest paper that used only LDA.

From these seven papers, the study (Rangayyan, Mudigonda, & Desautels, 2000) used different classes than just benign and malignant. In fact, the chosen classes were benign, malignant, circumscribed and spiculated. The accuracy for benign vs. malignant classification was of 81.5% while this metric was of 90.7% for circumscribed vs. spiculated.

The works in Bojar and Nieniewski (2008); Mudigonda, Rangayyan, and Desautels (1999) used Fisher's linear discriminant as a classifier instead of its generalization (LDA). The best results achieved in terms of AUC by Bojar and Nieniewski (2008) was 0.724 using only one feature they proposed, while Mudigonda et al. (1999) achieved an accuracy of 81% in the classification process.

The newest paper we found in this systematic review, that used only LDA, was published in 2008 as shown in Table S6. This means that the researchers are no longer using LDA as the only classifier (at least it is not as popular as other techniques) to solve the problem of mass classification.

4.6. Naïve Bayes. In Table S7 in the supplementary material there is a summary of the four papers (3%) that used only Naïve Bayes as classifier to discriminate masses.

From the same research group, the studies (Benndorf, Burnside, Herda, Langer, & Kotter, 2015a; Benndorf et al., 2015b) used BI-RADS descriptors and patient age as features. The work in Benndorf et al. (2015a) achieved the best AUC of 0.90 classifying masses in benign and malignant categories, while Benndorf et al. (2015b) achieved an AUC of 0.935 using BI-RADS category as feature and 0.876 without this feature classifying masses as BI-RADS categories (0 - incomplete, 2 -benign, 3 - probably benign, 4 - suspicious abnormality, 5 - highly suspicious of malignancy).

The study in Mencattini, Salmeri, Rabottino, and Salicone (2010) used shape and texture features to classify masses as benign or malignant, achieving an average AUC of 0.88 selecting the most appropriate features based on classifier error. On the other hand, in Wu et al. (2013) the authors employed mutual information using Shannon's entropy measure to select the most adequate features composed of shape features, density and patient information such as age, hormone therapy, history of breast cancer in family and so on. The highest AUC achieved was 0.807.

Table S7 shows that the four papers we found were published after 2010. Actually, it was one paper in 2010, another one in 2013 and two papers from the same research group in 2015, showing that Naïve Bayes is not often used in the task of discriminating masses. If we consider the papers that did not use BI-RADS as features, we see that the best result was an AUC of 0.88, which although is not a bad result, it is also not the state of art.

4.7. Other techniques. In Table S8 in the supplementary material we have the papers that used other techniques to build the classifiers, such as decision trees, rule-based systems, mixture models, squared discriminant analysis, discriminant function and Bayesian classifier, genetic programming and artificial immune systems.

Decision trees were used in only three (2%) of the analyzed papers as the only classifier to discriminate masses. An overview of all these papers can be seen in Table .

Shape features were used in Todd and Naghdy (2004); Vadivel and Surendiran (2013). The study (Todd & Naghdy, 2004) discriminated the masses as benign and malignant and provided the results in terms of false-negative rate (0%) and false-positive rate (9.3%), while the study in Vadivel and Surendiran (2013) classified the masses as round, oval, lobular and irregular, achieving an accuracy of 87.76% considering all classes. In Mohanty et al. (2013) the authors used texture features to discriminate masses as benign and malignant and as result they achieved an AUC of 0.995.

Decision trees are quite rare considering all the techniques used to build classifiers to perform classification of masses in mammograms. From the three papers we analyzed one was published in 2004 and the others in 2013.

Only two (1%) of the analyzed papers built a rule-based system to classify masses. Table S8 summarizes these papers. The authors of Javadi and Faez (2012) developed a system that makes use of fuzzy rules to classify masses as benign and malignant. The input for the classifier is wavelet coefficients selected using a particle swarm optimization procedure. The accuracy of the system was 93.41%. The other paper created a rule-based system where roundness of a mass was the only feature used to discriminate the masses as benign, probably benign and possibly malignant, probably malignant and possibly benign, and malignant (Al-Najdawi, Biltawi, & Tedmori, 2015). The specificity and sensitivity of the system were 96.2% and 94.4%, respectively.

Although rule-based systems are not so popular, these papers are quite new, especially in Al-Najdawi et al. (2015) which shows that some researchers are trying different approaches to solve the problem of mass classification.

Mixture models were used by just two papers (1%). An overview of these papers can be seen in Table S8. The work in Elguebaly and Bouguila (2013) used Dirichlet mixture model together with features based on local binary pattern and Haralick's features to discriminate masses as benign, malignant and normal. The model worked better with local binary pattern features producing an accuracy of 84.21%. A Gaussian mixture model was developed in Mishra and Ranganathan (2014) to classify masses as benign and malignant. Energy and entropy served as input to the classifier that improved the diagnostic of breast cancer of over 95%.

Similar to what happened with the papers that proposed ruled base systems, the mixture models are not really popular, but as they were published in 2014 and 2015, they are relatively new. This fact can be used to confirm that the researchers are trying new techniques (or at least not so explored techniques) in this area.

The study (Leichter et al., 2000) used squared discriminant analysis to build a CAD system that uses features based on the degree of spiculation of masses to help radiologists discriminate masses as benign and malignant. Without the system the AUC obtained by the radiologists was 0.66, but with the CAD they achieved an AUC of 0.81. An overview of this paper can be seen in Table S8. In Rangayyan and Nguyen (2007) shape features were used to serve as input to the classifier built using discriminant function together with a Bayesian classifier to classify masses as benign and malignant. The tests were performed using MIAS and Alberta databases. The AUC for the combined database was 0.93.

A genetic programming was used to create a classifier in Nandi, Nandi, Rangayyan, and Scutt (2006). The authors used a genetic algorithm to select the most powerful features in a set that contains shape and texture features and classified the masses as benign and malignant classes. The range of the classification rate was from 90.1% to 100% depending on the group of selected images. Table shows an overview of this paper.

Artificial immune systems appeared in just one of the analyzed papers. The study in Dehache and Souici-Meslati (2015) used this technique applying a KNN approach together with BI-RADS features and patient age to separate masses as benign and malignant. The results can be seen in Table S8.

Table S8 shows an overview of the only paper that used radial basis function network as a technique to classify masses. In Zhang, Wang, Shin, Hruska, and Son (2015), the authors used four type of features based on Fourier index/descriptors, compactness and fractal dimension, achieving an AUC of 0.99.

4.8. More than one type of classifier. We analyzed 32 papers (25%) that used more than one technique to classify masses. Some of these papers combined different techniques to create a classifier, while others used different classifiers to test the effectiveness of a new feature or a segmentation method or to test feature selection approaches. For instance, the authors of Huo et al. (1998); Huo, Maryellen Giger, Vyborny, Wolverton, and Metz (2000) created a hybrid classifier using ANN and rule-based systems, while in Rouhi et al. (2015) ANN, SVM, Naïve Bayes, KNN and random forest were used to classify masses segmented using region growing and cellular neural network techniques. In Chaieb et al. (2014) several feature selection techniques were tested such as genetic algorithm, tabu search, ReliefF algorithm, PCA and sequential forward/backward selection. An overview of each paper that used more than one technique is shown in Table S9 in the supplementary material.

The most common classifiers were SVM (20 papers - 62%), ANN (19 papers - 59%), KNN (14 papers - 44%) and LDA (11 papers - 34%). These four classifiers are almost the same most used classifiers in papers that used only one technique to classify masses. However, there is a subtle difference when considering papers that used only one technique, once ANN was the most employed technique, followed by SVM, KNN and Clustering/LDA.

In Table S9 we see there is no prevalence of any method to select the most powerful features. The techniques used were fractal analysis using MANOVA that appeared in Georgiou, Mavroforakis, Dimitropoulos, Cavouras, and Theodoridis (2007); Mavroforakis, Georgiou, Dimitropoulos, Cavouras, and Theodoridis (2006); logistic regression was used in Huang, Hung, Lee, Li, and Wang (2012); Rabidas, Chakraborty, and Midya (2017); genetic algorithm was shown in Chaieb et al. (2014); Rouhi et al. (2015); SVM with extreme learning machine was presented in Xie, Li, and Ma (2016) and SVM using mutual information feature selection filter appeared in Liu and Tang (2014); in Muramatsu et al. (2016) PCA was used; a measure of purity based on entropy was employed in Zhang, Tomuro, Furst, and Raicu (2012); and the classifier error was used to select the most discriminant features in Hapfelmeier and Horsch (2011); Khademi, Sahba, Venetsanopoulos, and Krishnan (2009).

The features used are pretty much the same ones found in the previous analyzed papers, i.e. shape features, wavelets coefficients, Zernike moments, texture features, local binary pattern features, features extracted from Fourier domain, BI-RADS, Hu moments, statistical features, features extracted from Ripplet-II coefficient matrix and so on.

5. Discussion

5.1. Techniques. Although several techniques have been used to discriminate masses in mammograms, Fig. 2 illustrates that three techniques have dominated this area: artificial neural network, support vector machine, and K-nearest neighbors. Despite the good results achieved with these techniques, for example, ANN was used in McLeod and Verma (2013) with an accuracy of 98%; SVM was employed in Khan et al. (2016) and obtained a AUC of 0.948; KNN appeared in Aroquiaraj and Thangavel (2014) with a AUC of 0.973, other techniques were also able to achieve similar results. Naïve Bayes was explored in only five studies but in Benndorf et al. (2015b) good results were found in terms of AUC (0.935 and 0.876 depending on the features used). Another example is the usage of rule-based systems that appeared only in Al-Najdawi et al. (2015); Javadi and Faez (2012) achieving sensibility and specificity higher than 90%, which shows that there is opportunity for this technique to be further explored. Differently from Fig. 2, in Table 3 we can see an abundance of techniques found in the studies, but that received little attention in this context and which can be further explored.

The papers reviewed in this survey are very diverse once they use different databases, define different classes for classification and show different performance metrics. In order to perform a comparison among some of them, as fair as possible, Table 4 shows studies published since 2014 that used only DDSM and MIAS databases, and classified the masses into two classes (benign and malignant). These articles used the metrics AUC or accuracy[10] to show the results.

Among the four studies that used DDSM database and showed their results in terms of accuracy, two of them (Azizi et al., 2014; Zemmal, Azizi, & Sellami, 2015) used SVM to classify the masses, while the others used ANN (Chokri & Farida, 2017; Jiao et al., 2017). The highest accuracy was 97.4% achieved in Jiao et al. (2017), which considers features learned using convolutional neural networks.

Considering the papers that used DDSM database and showed their results in terms of AUC, the best result was achieved using radial basis function network as classifier (AUC of 0.99) with shape-based features (Zhang et al., 2015), followed by Wang, Li, and Gao (2014) that used SVM and latent spatial features, as well as statistical marginal characteristics as input to the classifier (AUC of 0.965). The third and the fourth best results were also reached in studies of the same research group that used SVM and Gabor features (Khan et al., 2017; Khan et al., 2016). In Khan et al. (2017) an AUC of 0.95 was achieved while in Khan et al. (2016) the AUC was of 0.948.

From the five studies using the MIAS database, two showed their results in terms of accuracy. In Jothilakshmi and

[10] The metric accuracy was used only when the AUC was not available for a specific study.

Table 3
Least used techniques to classify mammographic masses.

Technique	Papers
Radial basis network/function	Georgiou et al. (2007); Jaleel, Salim, and S (2014); Zhang et al. (2015)
Mixture model	Elguebaly and Bouguila (2013); Mishra and Ranganathan (2014)
Logistic regression	Huang et al. (2012); Rabidas et al. (2017)
Artificial immune systems	Dehache and Souici-Meslati (2015)
Discriminant function and Bayesian classifier	Rangayyan and Nguyen (2007)
Conditional inference trees	Hapfelmeier and Horsch (2011)
Genetic programming	Nandi et al. (2006)
Squared discriminant analysis	Leichter et al. (2000)

Table 4
Comparison of papers published since 2014 that used only one classifier.

Paper	Technique	Database	Features	Result
Zemmal et al. (2015)	SVM	DDSM	Texture features; central and Hu moments	Accuracy = 93.1%
Azizi et al. (2014)	SVM	DDSM	texture features; central and Hu moments	Accuracy = 93%
Chokri and Farida (2017)	ANN	DDSM	Texture features; shape features; margin features; patient age	Accuracy = 88.02%
Jiao et al. (2017)	ANN	DDSM	Features learned using convolutional neural networks	Accuracy = 97.4%
Zhang et al. (2015)	Radial basis function network	DDSM	Fourier irregularity index, compactness index, fractal dimension, Fourier-descriptor-based shape factor	AUC = 0.99
Li et al. (2016)	KNN	DDSM	Texton histograms	AUC = 0.91
Wang et al. (2014)	SVM	DDSM	Latent spatial features; statistical marginal characteristics	AUC = 0.965
Khan et al. (2017)	SVM	DDSM	Gabor features	AUC = 0.95
Benndorf et al. (2015a)	Naïve Bayes	DDSM	BI-RADS descriptors and patient age	AUC = 0.90
Li et al. (2015)	KNN	DDSM	Texton histograms	AUC = 0.92
Khan et al. (2016)	SVM	DDSM	Gabor Features	AUC = 0.948
Jothilakshmi and Raaza (2017)	SVM	MIAS	Texture features	Accuracy = 94%
Jiao et al. (2017)	ANN	MIAS	Features learned using convolutional neural networks	Accuracy = 96.7%
Serifovic-Trbalic et al. (2014)	ANN	MIAS	Zernike moments	AUC = 0.8920
Aroquiaraj and Thangavel (2014)	KNN	MIAS	Statistical and texture features	AUC = 0.973
Abdel-Nasser et al. (2016)	SVM	MIAS	Local binary patterns; local directory number; histogram of oriented gradient; texture features, Gabor features	AUC = 0.78

Raaza (2017) the authors used SVM and texture features, achieving an accuracy of 94% whereas in Jiao et al. (2017) the authors used ANN with features learned using convolutional neural networks accomplishing an accuracy of 96.7%.

The other studies (three papers) used MIAS database and showed their results in terms of AUC. The best AUC was 0.973 accomplished using KNN as classifier, together with statistical and texture features (Aroquiaraj & Thangavel, 2014). In Serifovic-Trbalic et al. (2014) the authors used Zernike moments as input to an ANN and obtained AUC of 0.8920, while in Abdel-Nasser et al. (2016) the authors employed SVM with a combination of different features (local binary patterns, local directory number, histogram oriented gradient, texture and Gabor features) achieving an AUC of 0.78.

Jiao et al. (2017) appears twice in Table 4 for different databases: DDSM database with accuracy of 97.4% and MIAS database with accuracy of 96.7%. These results show that the overall accuracy of a classifier can vary when it is applied on different databases.

Table 5 summarizes some of the most striking studies reviewed in this survey. Although several works used similar approaches, papers in Table 5 were selected due to their contribution to this research area and because they presented details on the techniques employed. Furthermore, all these papers presented quantitative results regarding the experiments performed.

Despite the fact that many techniques were employed to classify the masses, we did not find any paper that used grammars or syntactic approaches during the systematic review. Especially in last decade, grammars have been applied in pattern recognition of images to recognize and build objects, in layout recognition and image segmentation (Pedro, Nunes, & Machado-Lima, 2013). Grammars were already applied in medical images to deal with 3D visualizations of coronary vessels (Trzupek, Ogiela, & Tadeusiewicz, 2011) and to perform leg bone fracture analysis (Ogiela, Tadeusiewicz, & Ogiela, 2008). However, none of the studies presented in this review used syntactic approaches to address the problem of masses classification. Performing an exploratory review we found only one paper that used grammars together with ANN to discriminate tumors (Tahmasbi, Saki, & Shokouhi, 2011), achieving an accuracy of 91.38% and AUC of 0.858. This paper did not appear in our systematic review because when searching in the IEEE Xplorer database we limited our query to search only in Document Title (as seen in Table 1) due to the high number of studies found when this constraint is not imposed.

5.2. Databases. We also suspect that the vast majority of the published papers are database-dependent, since the majority of the studies used only one database (77%). One of the difficulties of working with different databases is that the images are acquired with different devices and stored with different spatial and pixel resolutions and format. Because of that, each dataset demands a different image preprocessing in order to standardize all images, which by itself it can be a research area. Some papers have presented studies regarding these areas, for example, Rangayyan, Nguyen, Ayres, and Nandi (2010) presented a study on the effect of pixel resolution in the classification process and Huo et al. (2000) presented a study regarding the robustness of a method when dealing with images acquired from different devices. Nevertheless, we believe that this area should be further explored, this way the pattern recognition/machine learning techniques could be used with different data sources and present consistent results.

In general, the studies try to use a balanced number of images of each class, so they can create a less biased classifier. However, this is not possible in all situations. To overcome the problem of unbalanced number of images (Lima, da Silva Filho, & dos Santos, 2016) used linear combinations with random weights to generate synthetic instances of benign and malignant cases to balance their

Table 5
Synthesis of some of the main papers analyzed in this survey.

Papers	Contribution
Meriem et al. (2015)	The use of Zipf and inverse Zipf power laws for mammograms analysis in the field of segmentation and classification of masses.
Li et al. (2015)	Technique to classify masses without a previous segmentation of each mass. The proposed approach combines texton analysis with subsampling strategies.
Kanadam and Chereddy (2016)	A new representation of a ROI using a sparse-ROI, leading to a reduction in size of the ROI (number of pixels), computational time and feature space.
Junior et al. (2009)	The use of Morans index and Gearys coefficient measures, extracted from mammograms, as input to a SVM classifier.
Xie et al. (2016)	The use of SVM combined with Extreme Learning Machine to select the most powerful features and to classify masses.
Jamieson et al. (2012); Ovalle et al. (2016)	The use of deep learning technique or Adaptive Deconvolutional Networks for learning features avoiding the use of handcrafted features.
Eltoukhy et al. (2010)	A comparative study between curvelet and wavelet transform for masses discrimination.
Hadjiiski et al. (2001); Timp, Varela, and Karssemeijer (2007)	The employment of temporal changes in mammographic masses, using interval change information in the processes of masses classification.
Mudigonda et al. (2000)	Representation of the mass using a polygonal modeling for the extraction of pixels across mass margins.
Huo, Giger, and Vyborny (2001)	The use of craniocaudal, mediolateral-oblique and special view (spot compression or spot compression magnification) to classify masses.
Nandi et al. (2006)	The use of genetic programming to select the most powerful features and to classify masses.
Muramatsu et al. (2016)	The use of radial local ternary patterns for classification of benign and malignant masses.
Tan et al. (2014a)	Investigation of a new approach to improve feature selection process and classifier optimization.
Dhungel, Carneiro, and Bradley (2017)	The use of deep learning methods for segmentation and classification of masses using handcrafted features.
Lima et al. (2016)	The use of Zernike moments extracted from decomposed image using multi-resolution wavelets to detect and classify lesions.
Zhang et al. (2015)	Creation of a Fourier Irregularity Index that has better performance than Compactness Index, Fractal Dimension and Fourier-descriptor-based shape Factor for mass classification.
Abdel-Nasser et al. (2016); Bruce and Kallergi (1999); Rangayyan et al. (2010)	Investigation of the effect of pixel resolution, preprocessing and feature normalization on the performance of methods for mass classification.

database. Another example of a study that used synthetic images is Tralic et al. (2011), as they used images manually drawn by specialists.

5.3. Features. Many researchers used wavelet features as input to the classifiers and achieved good results. Some of the papers that used wavelet features were Bruce et al. (1999, 1999); Görgel, Sertbas, and Uçan (2013); Wagner, Elter, Schulz-Wendtland, and Wittenberg (2011). Despite the good results, the authors of Bruce et al. (1999) stated that a relatively large computational cost was involved when using this type of feature. In Eltoukhy et al. (2010) a comparison between wavelet and curvlet features was presented, and curvlet presented a better classification rate. It is important to note that Eltoukhy et al. (2010) was the only paper that used curvlet features and, as it presented a good classification rate of 94.07%, we believe that this type of feature can be further explored.

Features based on fractal analysis were used in some papers, for example, Beheshti, AhmadiNoubari, Fatemizadeh, and Khalili (2014); Yang et al. (2005). Beheshti et al. (2014) stated that fractal features are useful to classify malignant masses in early stages, which can help radiologists provide breast cancer diagnosis sooner. Due to this special ability, fractal features could also be further explored in future researches.

Texture and shape features were the two most common types of features used to discriminate mammographic masses. Texture features can be extracted using many different techniques such as gray-level co-occurrence matrix, gray-level run-length matrix, gray-level aura matrix, run-length statistics matrix and so on. Shape features are features that try to describe the shape of masses, for example, area, perimeter, circularity, concavity or convexity indexes, spiculation index and so on. Malignant masses tend to be more irregular and spiculated, while benign ones tend to be more round and oval. Because of this fact, shape features are widely used in the process of mass classification. Nevertheless, these types of features require using a good segmentation technique that is able to separate the mass from the background of a ROI.

Automatic segmentation of masses can be a challenging task, mainly in high breast density. For this reason, some masses need to be manually segmented by experienced radiologists. To overcome this problem, many papers used only texture features, for example in Kanadam and Chereddy (2016); Mishra and Ranganathan (2014); Mohanty et al. (2013). The fact is that when it is decided to use only shape features, you do need a good method of mass segmentation and, even when this method is available, it will be difficult to deal with scenarios in which malignant masses have round or oval shapes and when benignant masses have an irregular or spiculated structure. On the other hand, using only texture features does not require such a precise mass segmentation method, but some important information is lost, once the vast majority of benign masses are round and oval and malignant ones are irregular and spiculated. In order to not lose important information, some authors combined texture and shape features to discriminate masses, for example in Delogu et al. (2007); Dong et al. (2015); Hadjiiski et al. (2001).

Other authors tried a different approach and did not used features extracted directly from the images, for example McLeod and Verma (2013); Wu et al. (2013). Instead, they used information provided by the radiologist who analyzed the images. This information is provided as BI-RADS descriptors of masses or patient information, such as patient age, patient/family cancer history, patient treatment and so on. However, some papers presented the combination of shape, texture and textual information to classify masses, for example Georgiou et al. (2007); Velthuizen and Gangadharan (2000); Verma (2008). It is important to mention that studies that used only information provided by radiologists are extremely dependent on this type of professional to provide the features. On the other hand, the ones that did not use these features can rely on techniques that are able to extract the features directly from the image automatically.

Although the use of BI-RADS descriptors as features can improve the accuracy of classification, the authors of Panchal and Verma (2006) discourage a classification system based only on these features or a combination with other patient information in order to avoid a system that is too human dependent.

5.4. Features selection. As the number of features increases (wavelet, curvlet, fractal, shape, texture, textual), we fall into a problem known as "the curse of dimensionality". Thus, selecting the most useful features is important in cases where there are features that are meaningless to a specific problem. In general, when more features are used as input to train the classifier, it takes more time to complete this task, but it does not mean the classifier will be more accurate. Yet, a set of features can be redundant, that is, they can have high correlation, thus only one feature of the set needs to be included. Furthermore, there are two aspects that are important: 1) meaningless features are not only useless, but can also include some noise decreasing the classifier accuracy; ii) due to the curse of dimensionality, for a fixed size of examples (instances), after a specific time the classifier accuracy decreases and the estimation error increases, due to the fact that the number or examples is not big enough to estimate all the parameters the classifier needs taking into consideration all the combinations of features. On the other hand, there are times when there are not enough different features to use and the alternative is to train the classifier using all the features available. In addition to that, when some features are removed, some valuable information may be lost, leading to a more inaccurate classification. As it is a well-known problem in the area of machine learning and pattern recognition, many papers presented some techniques to select the most useful features for masses discrimination, for example in Hapfelmeier and Horsch (2011); Khan et al. (2017); Khan et al. (2016); Tan, Pu, and Zheng (2014b). Overall, the papers that presented their results with and without feature selection approaches tend to show better results after applying feature selection techniques.

In Table S1 in the supplementary material we can see all the techniques used to select the most discriminative features. In Tan et al. (2014a), the authors implemented a sequential forward floating selection (SFFS) that had the highest computational time efficiency (3% - 5%) when compared to genetic algorithm. The authors suggested that although genetic algorithm is a powerful tool to be used in CAD systems, it is a very computationally intensive method and using SFFS can improve the feature selection efficiency.

5.5. Gaps and challenges. As noticed during this review, most of the studies used only one database. A possible problem when only one database is used is that the classifier can be good to deal with this database, but it may perform poorly with images from different sources. Thus, one possible gap in this research area has to do with creating more robust systems that can handle a wider variety of databases, which could increase the possibility of clinicians using the systems on a daily basis.

Another gap has to do with the creation of systems that can learn and improve in on-line mode. The classifiers shown in most of the papers can learn only in off-line mode or batch. The creation of systems able to learn new patterns as new images are presented in on-line mode could be very helpful for this area.

No paper was found where the authors built a classifier using grammars, syntactic approaches or graphs. Many researchers are using these methods to deal with the problem of image classification, but it seems that these techniques are not being employed in mass classification. We believe that this gap can be further explored in future works.

At the same time that some gaps were found, we also noticed some recurring problems in some of the analyzed papers. The first one is the lack of standardizing with relation of the metrics used to evaluate the works. While accuracy and the AUC are the most used metrics, we also found many other metrics, for example, sensibility, specificity, precision, false positive rate, false negative rate and so on. The problem identified here is the difficulty of comparing a study that presents only the sensibility and specificity with another that presents only the AUC. This difficulty also makes it very hard to find the state of the art in this research area.

Another common problem we point out is that in general the researchers do not compare the results obtained by the classifier with the results that are obtained by clinicians. For example, an accuracy of 90% obtained by a classifier can be good or bad depending on the accuracy that could be obtained by a doctor analyzing the same images, otherwise the system cannot be used on a daily basis by clinicians. Here, we can question why this comparison is not performed. Maybe the doctors are not interested in this type of research; maybe the researchers think the doctor's evaluation is not important; or maybe it is just the lack of available clinicians.

The next problem we found was the fact that in many papers the method used during the tests is not clear or was not present. In some cases, the authors made it quite clear they were using a k-fold cross validation, a leave one out method, a hold out method and so on. But it was quite common to find papers where the authors simply did not mention which technique was used. Furthermore, there were cases where it was hard to figure out if the same images were used during the training and the test phases. It is important to bear in mind that this information is very useful, mainly because when the tests are performed with the same images used during the training phase the results tend to be better.

The last problem addresses the fact that some research groups tend to publish more than one paper with the same subject, but only with some changes made in the model. Sometimes the changes made are just a parameter, a feature or a different method to select the most powerful features. We believe this is because the researchers tend to prioritize the number of published papers instead of their quality. The vast amount of incomplete papers increases the difficulties of finding complete ones, so we believe that this is in fact a problem in this research area.

During this systematic review we were able to verify the many techniques used to classify tumors, the use of numerous different databases, the many different features selected, several methods applied to reduce feature dimension, as well as different test techniques and metrics used to test and evaluate the results. Because of the various methods, approaches and techniques, it is very difficult to quantitatively compare the results of one paper to the others, since they generally vary in many different aspects, for example, database, segmentation techniques, features employed, machine learning techniques and so on. To perform a fair comparison a common database would be necessary and the classifiers should be built using exactly the same images. For instance, when using k-fold cross validation, the k folds should have the same images.

6. Conclusion

The papers reviewed here allowed an extensive overview regarding the use of pattern recognition and machine learning techniques to classify masses in mammograms. We believe there is much interest in this research area, since every year studies using a variety of different techniques and features are published. Moreover, many of the analyzed articles provided a recognition rate higher than 90%, which shows this area may be reaching its prime.

Considering pattern recognition and machine learning techniques, we showed that artificial neural network, support vector machine and k-nearest neighbors dominate this area. Some traditional techniques, such as Naïve Bayes, logistic regression, random forest were used in some studies, but they appeared at a very low frequency.

We noticed that syntactic methods were not used in any of the analyzed papers and it can be considered a gap in this research area. Another gap is the fact that the majority of the studies used only one database. Using different image sources can lead to the creation of more robust classifiers. Lastly, the lack of classifiers that can learn in an on-line mode can be considered another possible gap in this research area.

In addition, we identified some challenges to be overcome. We can cite the lack of standardizing the metrics used to evaluated the studies, the absence of a benchmarking database or system to facilitate the comparison of classifiers, the majority of researchers do not compare their results with clinicians, quite often the method used to perform the training and tests are not clear, and the number of incomplete published studies is very high.

Finally, we would like to highlight how this research area is important to society as a whole. Breast cancer is one of the main health problems of our time and deserves attention from private companies and public research institutes. With sooner and more precise diagnosis, more women will survive this ordeal.

7. Competing interests

The authors declare that they have no competing interests.

8. Declarations of interest

None.

9. Funding

Brazilian National Council of Scientific and Technological Development, (CNPq) (Process 157535/2017-7), and the National Institute of Science and Technology – Medicine Assisted by Scientific Computing (INCT-MACC).

NAP eScience-PRP-USP and São Paulo Research Foundation (FAPESP): grant #2005/556288.

10. Author's contributions

Nunes, Ftima L. S. and Machado-Lima, A. designed and coordinated this research. Pedro, R.W.D, analyzed the papers and drafted the manuscript. The authors reviewed this document and approved the final manuscript.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eswa.2018.10.032.

References

Abdel-Nasser, M., Melendez, J., Moreno, A., & Puig, D. (2016). The impact of pixel resolution, integration scale, preprocessing, and feature normalization on texture analysis for mass classification in mammograms. *International Journal of Optics, 2016*. doi:10.1155/2016/1370259.

Al-Najdawi, N., Biltawi, M., & Tedmori, S. (2015). Mammogram image visual enhancement, mass segmentation and classification. *Applied Soft Compututing, 35*, 175–185.

Aroquiaraj, I. L., & Thangavel, K. (2014). Mass classification method in mammogram using fuzzy k-nearest neighbour equality. *CoRR, abs/1406.4770*.

Azevedo, W. W., Lima, S. M. L., Fernandes, I. M. M., Rocha, A. D. D., Cordeiro, F. R., da Silva-Filho, A. G., & dos Santos, W. P. (2015a). Fuzzy morphological extreme learning machines to detect and classify masses in mammograms. In *2015 ieee international conference on fuzzy systems (fuzz-ieee)* (pp. 1–8). doi:10.1109/FUZZ-IEEE.2015.7337975.

Azevedo, W. W., Lima, S. M. L., Fernandes, I. M. M., Rocha, A. D. D., Cordeiro, F. R., da Silva-Filho, A. G., & dos Santos, W. P. (2015b). Morphological extreme learning machines applied to detect and classify masses in mammograms. In *2015 international joint conference on neural networks (ijcnn)* (pp. 1–8). doi:10.1109/IJCNN.2015.7280774.

Azizi, N., Zemmal, N., Sellami, M., & Farah, N. (2014). A new hybrid method combining genetic algorithm and support vector machine classifier: Application to cad system for mammogram images. In *2014 international conference on multimedia computing and systems (icmcs)* (pp. 415–420). doi:10.1109/ICMCS.2014.6911285.

Beheshti, S. M. A., AhmadiNoubari, H., Fatemizadeh, E., & Khalili, M. (2014). An efficient fractal method for detection and diagnosis of breast masses in mammograms. *Journal of Digital Imaging, 27*(5), 661–669.

Benndorf, M., Burnside, E. S., Herda, C., Langer, M., & Kotter, E. (2015a). External validation of a publicly available computer assisted diagnostic tool for mammographic mass lesions with two high prevalence research datasets. *Medical Physics, 42*(8), 49874996.

Benndorf, M., Kotter, E., Langer, M., Herda, C., Wu, Y., & Burnside, E. S. (2015b). Development of an online, publicly accessible naive bayesian decision support tool for mammographic mass lesions based on the american college of radiology (acr) bi-rads lexicon. *European Radiology, 25*(6), 1768–1775. doi:10.1007/s00330-014-3570-6.

Bhattacharya, M., Sharma, N., Goyal, V., Bhatia, S., & Das, A. (2011). A study on genetic algorithm based hybrid softcomputing model for benignancy/malignancy detection of masses using digital mammogram. *International Journal of Computational Intelligence and Applications, 10*(02), 141–165. doi:10.1142/S1469026811003033.

Bojar, K., & Nieniewski, M. (2008). New features for classification of cancerous masses in mammograms based on morphological dilation. In *2008 5th international conference on visual information engineering (vie 2008)* (pp. 111–116). doi:10.1049/cp:20080293.

Bruce, L. M., & Adhami, R. R. (1999). Classifying mammographic mass shapes using the wavelet transform modulus-maxima method. *IEEE Transactions on Medical Imaging, 18*(12), 1170–1177. doi:10.1109/42.819326.

Bruce, L. M., & Kallergi, M. (1999). Effects of image resolution and segmentation method on automated mammographic mass shape classification. 10.1117/12.348654

Bruce, L. M., Kallergi, M., & Mendoza, A. (1999). Wavelet scalar-energy features for recognition of mammographic mass shapes. In H. H. Szu (Ed.), *Wavelet applications vi: 3723* (pp. 156–162). doi:10.1117/12.342923.

Chaieb, R., Bacha, A., Kalti, K., & Lamine, F. B. (2014). Image features extraction for masses classification in mammograms. In *2014 6th international conference of soft computing and pattern recognition (socpar)* (pp. 203–208). doi:10.1109/SOCPAR.2014.7008006.

Cheikhrouhou, I., Djemal, K., Sellami, D., Maaref, H., & Derbel, N. (2008). New mass description in mammographies. In *2008 first workshops on image processing theory, tools and applications* (pp. 1–5). doi:10.1109/IPTA.2008.4743751.

Cheng, H. D., Shi, X. J., Min, R., Hu, L. M., Cai, X. P., & Du, H. N. (2006). Approaches for automated detection and classification of masses in mammograms. *Pattern Recogn, 39*(4), 646–668. doi:10.1016/j.patcog.2005.07.006.

Chokri, F., & Farida, M. H. (2017). Mammographic mass classification according to bi-rads lexicon. *IET Computer Vision, 11*(3), 189–198. doi:10.1049/iet-cvi.2016.0244.

Dehache, I., & Souici-Meslati, L. (2015). Artificial immune recognition system for mammographic mass classification. In *2015 third world conference on complex systems (wccs)* (pp. 1–5). doi:10.1109/ICoCS.2015.7483253.

Delogu, P., Fantacci, M. E., Kasae, P., & Retico, A. (2007). Characterization of mammographic masses using a gradient-based segmentation algorithm and a neural classifier. *Computers in Biology and Medicine, 37*(10), 1479–1491.

Dhungel, N., Carneiro, G., & Bradley, A. P. (2017). A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Medical Image Analysis, 37*, 114–128.

Dong, M., Lu, X., Ma, Y., Guo, Y., Ma, Y., & Wang, K. (2015). An efficient approach for automated mass segmentation and classification in mammograms. *Journal of Digital Imaging, 28*(5), 613–625.

Edwards, D. C. (2004). Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions. *Medical Physics, 31*, 81. doi:10.1118/1.1631912.

Edwards, D. C., Lan, L., Metz, C. E., Giger, M. L., & Nishikawa, R. M. (2003). Bayesian ann estimates of three-class ideal observer decision variables for classification of mammographic masses. 10.1117/12.480343

Elguebaly, T., & Bouguila, N. (2013). A bayesian approach for the classification of mammographic masses. In *2013 sixth international conference on developments in esystems engineering* (pp. 99–104). doi:10.1109/DeSE.2013.26.

Eltoukhy, M. M., Faye, I., & Samir, B. B. (2010). A comparison of wavelet and curvelet for breast cancer diagnosis in digital mammogram. *Comp. in Bio. and Med., 40*(4), 384–391.

Georgiou, H. V., Mavroforakis, M. E., Dimitropoulos, N., Cavouras, D., & Theodoridis, S. (2007). Multi-scaled morphological features for the characterization of mammographic masses using statistical classification schemes. *Artificial Intelligence in Medicine, 41*(1), 39–55.

Görgel, P., Sertbas, A., & Uçan, O. N. (2013). Mammographical mass detection and classification using local seed region growing-spherical wavelet transform (lsrg-swt) hybrid scheme. *Computers in Biology and Medicine, 43*(6), 765–774.

Hadjiiski, L., Sahiner, B., & Chan, H. (2006). Advances in computer-aided diagnosis for breast cancer. *Current Opinion in Obstetrics and Gynecology, 18*(1), 64–70. doi:10.1097/01.gco.0000192965.29449.da.

Hadjiiski, L., Sahiner, B., Chan, H., Petrick, N., Helvie, M., & Gurcan, M. (2001). Analysis of temporal changes of mammographic features: Computer-aided classification of malignant and benign breast masses. *Medical Physics, 28*(11), 2309–2317. doi:10.1118/1.1412242.

Hapfelmeier, A., & Horsch, A. (2011). Image feature evaluation in two new mammography cad prototypes. *International Journal of Computer Assisted Radiology and Surgery, 6*(6), 721–735.

Huang, M.-L., Hung, Y.-H., Lee, W.-M., Li, R. K., & Wang, T.-H. (2012). Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification techniques in breast cancer dataset classification diagnosis. *Journal of Medical Systems, 36*(2), 407–414.

Huo, Z., Giger, M. L., & Vyborny, C. J. (2001). Computerized analysis of multiple-mammographic views: potential usefulness of special view mammograms in computer-aided diagnosis.. *IEEE Transactions on Medical Imaging, 20*(12), 1285–1292.

Huo, Z., Giger, M. L., Vyborny, C. J., Wolverton, D. E., Schmidt, R. A., & Doi, K. (1998). Automated computerized classification of malignant and benign masses on digitized mammograms. *Academic Radiology, 5*(3), 155–168. doi:10.1016/S1076-6332(98)80278-X.

Huo, Z., Maryellen Giger, L., Vyborny, C. J., Wolverton, D. E., & Metz, C. E. (2000). Computerized classification of benign and malignant masses on digitized mammograms: A study of robustness. *Academic Radiology, 7*(12), 1077–1084. doi:10.1016/S1076-6332(00)80060-4.

Hussain, M., Khan, S., Muhammad, G., & Bebis, G. (2012). A comparison of different gabor features for mass classification in mammography. In *2012 eighth international conference on signal image technology and internet based systems* (pp. 142–148). doi:10.1109/SITIS.2012.31.

Jaleel, J. A., Salim, S., & S, A. (2014). Mammogram mass classification based on discrete wavelet transform textural features. In *2014 international conference on advances in computing, communications and informatics (icacci)* (pp. 718–722). doi:10.1109/ICACCI.2014.6968479.

Jamieson, A. R., Drukker, K., & Giger, M. L. (2012). Breast image feature learning with adaptive deconvolutional networks. 10.1117/12.910710

Javadi, S. M. T., & Faez, K. (2012). Finding suspicious masses of breast cancer in mammography images using particle swarm algorithm and its classification using fuzzy methods. In *2012 international conference on computer communication and informatics* (pp. 1–5). doi:10.1109/ICCCI.2012.6158915.

Jiao, Z., Gao, X., Wang, Y., & Li, J. (2017). A parasitic metric learning net for breast mass classification based on mammography. *Pattern Recognition*. doi:10.1016/j.patcog.2017.07.008.

Jothilakshmi, G. R., & Raaza, A. (2017). Effective detection of mass abnormalities and its classification using multi-svm classifier with digital mammogram images. In *2017 international conference on computer, communication and signal processing (icccsp)* (pp. 1–6). doi:10.1109/ICCCSP.2017.7944090.

Junior, G. B., de Paiva, A. C., Silva, A. C., & de Oliveira, A. C. M. (2009). Classification of breast tissues using moran's index and geary's coefficient as texture signatures and svm. *Computers in Biology and Medicine, 39*(12), 1063–1072.

Kanadam, K. P., & Chereddy, S. R. (2016). Mammogram classification using sparse-roi: A novel representation to arbitrary shaped masses. *Expert Systems Applications, 57*, 204–213.

Keleş, A., Keleş, A., & Yavuz, U. (2013). Extracting fuzzy rules for the diagnosis of breast cancer. *Turkish Journal of Electrical Engineering and Computer Sciences, 21*(5), 1495–1503.

Khademi, A., Sahba, F., Venetsanopoulos, A. N., & Krishnan, S. (2009). Region, lesion and border-based multiresolution analysis of mammogram lesions. In M. S. Kamel, & A. C. Campilho (Eds.), *Iciar*. In *Lecture Notes in Computer Science: 5627* (pp. 802–813). Springer.

Khan, S., Hussain, M., Aboalsamh, H., & Bebis, G. (2017). A comparison of different gabor feature extraction approaches for mass classification in mammography. *Multimedia Tools Applications, 76*(1), 33–57.

Khan, S., Hussain, M., Aboalsamh, H., Mathkour, H., Bebis, G., & Zakariah, M. (2016). Optimized gabor features for mass classification in mammography. *Applied Soft Computing, 44*, 267–280.

Laroussi, M. G., Ayed, N. G. B., Masmoudi, A. D., & Masmoudi, D. S. (2013). Diagnosis of masses in mammographic images based on zernike moments and local binary attributes. In *2013 world congress on computer and information technology (wccit)* (pp. 1–6). doi:10.1109/WCCIT.2013.6618683.

Leichter, I., Fields, S., Nirel, R., Bamberger, P., Novak, B., Lederman, R., & Buchbinder, S. (2000). Improved mammographic interpretation of masses using computer-aided diagnosis. *European Radiology, 10*(2), 377–383. doi:10.1007/s003300050059.

Leod, P. M., & Verma, B. (2012). Clustered ensemble neural network for breast mass classification in digital mammography. In *The 2012 international joint conference on neural networks (ijcnn)* (pp. 1–6). doi:10.1109/IJCNN.2012.6252539.

Li, H., Giger, M. L., Yuan, Y., Chen, W., Horsch, K., Lan, L., et al. (2008). Evaluation of computer-aided diagnosis on a large clinical full-field digital mammographic dataset. *Academic Radiology, 15*(11), 1437–1445. doi:10.1016/j.acra.2008.05.004.

Li, H., Meng, X., Wang, T., Tang, Y., & Yin, Y. (2017). Breast masses in mammography classification with local contour features. *BioMedical Engineering OnLine, 16*, 44. doi:10.1186/s12938-017-0332-0.

Li, Y., Chen, H., Rohde, G. K., Yao, C., & Cheng, L. (2015). Texton analysis for mass classification in mammograms. *Pattern Recognition Letters, 52*(C), 87–93. doi:10.1016/j.patrec.2014.10.008.

Li, Y., Chen, H., Wei, X.-Y., Peng, Y., & Cheng, L. (2016). Mass classification in mammograms based on two-concentric masks and discriminating texton. *Pattern Recognition, 60*, 648–656.

Lima, S. M. L., da Silva Filho, A. G., & dos Santos, W. P. (2016). Detection and classification of masses in mammographic images in a multi-kernel approach. *Computer Methods and Programs in Biomedicine, 134*, 11–29.

Liu, X., & Tang, J. (2014). Mass classification in mammograms using selected geometry and texture features, and a new svm-based feature selection method. *IEEE Systems Journal, 8*(3), 910–920.

Mavroforakis, M. E., Georgiou, H. V., Dimitropoulos, N., Cavouras, D., & Theodoridis, S. (2006). Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers. *Artificial Intelligence in Medicine, 37*(2), 145–162.

McLeod, P., & Verma, B. (2013). Variable hidden neuron ensemble for mass classification in digital mammograms [application notes]. *IEEE Comp. Int. Mag., 8*(1), 68–76.

Mencattini, A., Salmeri, M., Rabottino, G., & Salicone, S. (2010). Metrological characterization of a cadx system for the classification of breast masses in mammograms. *IEEE Transactions on Instrumentation and Measurement, 59*(11), 2792–2799. doi:10.1109/TIM.2010.2060751.

Meriem, H., Merouani, H. F., & Lakhdar, L. (2015). The power laws: Zipf and inverse zipf for automated segmentation and classification of masses within mammograms. *Evolving Systems, 6*(3), 209–227.

Mishra, S., & Ranganathan, H. (2014). Multi layer architecture for breast cancer diagnosis. *Indian Journal of Computer Science and Engineering, 5*(1), 18–25.

Mohanty, A. K., Senapati, M. R., Beberta, S., & Lenka, S. K. (2013). Texture-based features for classification of mammograms using decision tree. *Neural Computing and Applications, 23*(3), 1011–1017. doi:10.1007/s00521-012-1025-z.

Mousa, R., Munib, Q., & Moussa, A. (2005). Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural. *Expert Systems Applications, 28*(4), 713–723.

Mudigonda, N. R., Rangayyan, R. M., & Desautels, J. E. L. (1999). Concavity and convexity analysis of mammographic masses via an iterative boundary segmentation algorithm. In *Engineering solutions for the next millennium. 1999 ieee canadian conference on electrical and computer engineering (cat. no.99th8411): 3* (pp. 1489–1494 vol.3). doi:10.1109/CCECE.1999.804929.

Mudigonda, N. R., Rangayyan, R. M., & Desautels, J. E. L. (2000). Gradient and texture analysis for the classification of mammography masses. *IEEE Transactions on Medical Imaging, 19*(10), 1032–1043.

Muramatsu, C., Hara, T., Endo, T., & Fujita, H. (2016). Breast mass classification on mammograms using radial local ternary patterns. *Computers in Biology and Medicine, 72*, 43–53.

Nandi, R. J., Nandi, A. K., Rangayyan, R. M., & Scutt, D. (2006). Classification of breast masses in mammograms using genetic programming and feature selection. *Medical Biological Engineering Computing, 44*(8), 683–694.

Nugroho, H. A., Faisal, N., Soesanti, I., & Choridah, L. (2014a). Analysis of digital mammograms for detection of breast cancer. In *2014 international conference on computer, control, informatics and its applications (ic3ina)* (pp. 25–29). doi:10.1109/IC3INA.2014.7042595.

Nugroho, H. A., Faisal, N., Soesanti, I., & Choridah, L. (2014b). Identification of malignant masses on digital mammogram images based on texture feature and correlation based feature selection. In *2014 6th international conference on information technology and electrical engineering (icitee)* (pp. 1–6). doi:10.1109/ICITEED.2014.7007907.

Ogiela, L., Tadeusiewicz, R., & Ogiela, M. (2008). Cognitive modeling in medical pattern semantic understanding. In *Multimedia and ubiquitous engineering, 2008. mue 2008. international conference on* (pp. 15–18). Busan, South Korea. doi:10.1109/MUE.2008.47.

Oliveira Martins, L., da Silva, E. C., Silva, A. C., de Paiva, A. C., & Gattass, M. (2007). Classification of breast masses in mammogram images using ripley's k function and support vector machine. In P. Perner (Ed.), *Mldm*. In *Lecture Notes in Computer Science: 4571* (pp. 784–794). Springer.

Ovalle, J. E. A., Gonzlez, F. A., Ramos-Polln, R., Oliveira, J. L., & Guevara-Lpez, M. Ã. (2016). Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer Methods and Programs in Biomedicine, 127*, 248–257.

Panchal, R., & Verma, B. (2006). Neural classification of mass abnormalities with different types of features in digital mammography. *International Journal of Computational Intelligence and Applications, 6*(1), 61–75.

Patil, S., Udupi, V. R., & Bhogale, P. (2013). An efficient cad system for detection and classification of tumors in mammographic images using variety features and probabilistic neural network. In *Confluence 2013: The next generation information technology summit (4th international conference)* (pp. 115–119). doi:10.1049/cp.2013.2303.

Pedro, R. W. D., Nunes, F. L. S., & Machado-Lima, A. (2013). Using grammars for pattern recognition in images: A systematic review. *ACM Computing Surveys, 46*(2), 26:1–26:34. doi:10.1145/2543581.2543593.

Rabidas, R., Chakraborty, J., & Midya, A. (2017). Analysis of 2d singularities for mammographic mass classification. *IET Computer Vision, 11*(1), 22–32. doi:10.1049/iet-cvi.2016.0163.

Rangayyan, R. M., Mudigonda, N. R., & Desautels, J. E. L. (2000). Boundary modelling and shape analysis methods for classification of mammographic masses. *Medical and Biological Engineering and Computing, 38*(5), 487–496. doi:10.1007/BF02345742.

Rangayyan, R. M., & Nguyen, T. M. (2007). Fractal analysis of contours of breast masses in mammograms. *Journal of Digital Imaging, 20*(3), 223–237.

Rangayyan, R. M., Nguyen, T. M., Ayres, F. J., & Nandi, A. K. (2010). Effect of pixel resolution on texture features of breast masses in mammograms. *Journal of Digital Imaging, 23*(5), 547–553.

Rathi, V., & Aggarwal, S. (2014). Comparing the performance of ann with fnn on mammography mass data set. In *2014 ieee international advance computing conference (iacc)* (pp. 1307–1314). doi:10.1109/IAdCC.2014.6779516.

Rouhi, R., Jafari, M., Kasaei, S., & Keshavarzian, P. (2015). Benign and malignant breast tumors classification based on region growing and cnn segmentation. *Expert Systems Applications, 42*(3), 990–1002.

Sahiner, B., Chan, H., Petrick, N., Helvie, M., & Goodsitt, M. (1998). Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis. *Medical Physics, 25*(4), 516–526. doi:10.1118/1.598228.

Santaella, C. H. M., Schiabel, H., Patrocinio, A. C., Nunes, F. d. L. d. S., & Romero, R. A. F. (2003). Classification of nodules in mammograms image by using wavelet transform. 10.1117/12.483552

Serifovic-Trbalic, A., Trbalic, A., Demirovic, D., Prljaca, N., & Cattin, P. C. (2014). Classification of benign and malignant masses in breast mammograms. In *Mipro* (pp. 228–233). IEEE.

Shi, J., Sahiner, B., Chan, H.-P., Ge, J., Hadjiiski, L., Helvie, M. A., et al. (2008). Characterization of mammographic masses based on level set segmentation with new image features and patient information. *Medical Physics, 35*(1), 280–290. doi:10.1118/1.2820630.

Suganthi, M., & Madheswaran, M. (2009). Mammogram tumor classification using multimodal features and genetic algorithm. In *2009 international conference on control, automation, communication and energy conservation* (pp. 1–6).

Suganthi, M., & Madheswaran, M. (2012). An improved medical decision support system to identify the breast cancer using mammogram. *Journal of Medical Systems, 36*(1), 79–91.

Tahmasbi, A., Saki, F., & Shokouhi, S. B. (2011). Cwla: A novel cognitive classifier for breast mass diagnosis. In *2011 18th iranian conference of biomedical engineering (icbme)* (pp. 255–259). doi:10.1109/ICBME.2011.6168567.

Tan, M., Pu, J., & Zheng, B. (2014a). A new and fast image feature selection method for developing an optimal mammographic mass detection scheme. *Medical Physics, 41*(8). doi:10.1118/1.4890080.

Tan, M., Pu, J., & Zheng, B. (2014b). Optimization of breast mass classification using sequential forward floating selection (sffs) and a support vector machine (svm) model. *International Journal of Computer Assisted Radiology and Surgery, 9*(6), 1005–1020.

Timp, S., Varela, C., & Karssemeijer, N. (2007). Temporal change analysis for characterization of mass lesions in mammography. *IEEE Transaction of Medical Imaging, 26*(7), 945–953.

Todd, C. A., & Naghdy, G. (2004). Method for breast cancer classification based solely on morphological descriptors. 10.1117/12.533938

Tralic, D., Bozek, J., & Grgic, S. (2011). Shape analysis and classification of masses in mammographic images using neural networks. In *2011 18th international conference on systems, signals and image processing* (pp. 1–5).

Trzupek, M., Ogiela, M., & Tadeusiewicz, R. (2011). Intelligent image content description and analysis for 3d visualizations of coronary vessels. In N. Nguyen, C.-G. Kim, & A. Janiak (Eds.), *Intelligent information and database systems*. In *Lecture Notes in Computer Science: 6592* (pp. 193–202). Springer Berlin Heidelberg. doi:10.1007/978-3-642-20042-7_20.

Vadivel, A., & Surendiran, B. (2013). A fuzzy rule-based approach for characterization of mammogram masses into bi-rads shape categories. *Comp. in Bio. and Med., 43*(4), 259–267.

Varela, C., Muller, J. M., & Karssemeijer, N. (2003). Mammographic mass characterization using sharpness and lobulation measures. 10.1117/12.480161

Velthuizen, R. P., & Gangadharan, D. (2000). Mammographic mass classification: initial results. 10.1117/12.387703

Verma, B. (2008). Novel network architecture and learning algorithm for the classification of mass abnormalities in digitized mammograms. *Artificial Intelligence in Medicine, 42*(1), 67–79.

Wagner, F., Elter, M., Schulz-Wendtland, R., & Wittenberg, T. (2011). Classification of mammographic masses: use and influence of a bilateral-filter-based flat-texture approach. 10.1117/12.878004

Wang, Y., Li, J., & Gao, X. (2014). Latent feature mining of spatial and marginal characteristics for mammographic mass classification. *Neurocomputing, 144*, 107–118.

Wu, Y., Alagöz, O., Ayvaci, M., del Rio, A. M., Vanness, D. J., Woods, R. S., & Burnside, E. S. (2013). A comprehensive methodology for determining the most informative mammographic features. *Journal of Digital Imaging, 26*(5), 941–947.

Xie, W., Li, Y., & Ma, Y. (2016). Breast mass classification in digital mammography based on extreme learning machine. *Neurocomputing, 173*, 930–941.

Yang, S.-C., Wang, C., Chung, Y., Hsu, G., Lee, S., Chung, P., & Chang, C. (2005). A computer-aided system for mass detection and classification in digitized mammograms. *Biomedical Engineering: Applications, Basis and Communications, 17*(05), 215–228. doi:10.4015/S1016237205000330.

Zemmal, N., Azizi, N., & Sellami, M. (2015). Cad system for classification of mammographic abnormalities using transductive semi supervised learning algorithm and heterogeneous features. In *2015 12th international symposium on programming and systems (isps)* (pp. 1–9). doi:10.1109/ISPS.2015.7244993.

Zhang, G., Wang, W., Shin, S. Y., Hruska, C. B., & Son, S.-H. (2015). Fourier irregularity index: A new approach to measure tumor mass irregularity in breast mammogram images. *Multimedia Tools and Applications, 74*(11), 3783–3798.

Zhang, Y., Tomuro, N., Furst, J. D., & Raicu, D. S. (2012). Building an ensemble system for diagnosing masses in mammograms. *International Journal of Computer Assisted Radiology and Surgery, 7*(2), 323–329.