A Case-Based Reasoning Methodology for the Classification of Microcalcification Clusters and Masses in Breast Cancer Diagnosis

Edén A. Alanís-Reyes[a,*], Sigfrido Iglesias-González[a], Hugo Terashima-Marín[a], Santiago E. Conant-Pablos[a], Benjamin Bustos[b]

[a] *Evolutionary Optimization Research Chair. Tecnológico de Monterrey. Monterrey, NL 64849, Mexico*
[b] *PRISMA Research Group. Department of Computer Science. Universidad de Chile. Santiago, Chile*

Abstract

Breast cancer diagnosis is a clinical field in which CAD systems have been introduced with the objective of assisting radiologists during the detection process, considering that an accurate prognosis in early stages can result in patient survival. In this paper we propose a Case-Based Reasoning approach to discriminate between benign and malignant breast cancer lesions (masses and micro calcification clusters, specifically), analyzing digital mammography, as a tool to support and enhance their diagnosis. We designed a k-nearest neighbour similarity-search combined with a binary genetic algorithm, as a feature selection mechanism, to retrieve similar cases from our database. We present an empirical study in which we analyze the fitness of six different similarity metrics applied in a k-nearest neighbors similarity-search on our datasets, for case retrieval. Pairwise statistical tests determine that linear correlation is the best similarity metric for our datasets. Finally, we explore the performance of our proposed CBR approach in assessing the malignancy of detected lesions by using four different classifiers: k-NN, neural network, SVM and Linear Discriminant Analysis.

Keywords: Case-Based Reasoning, Supervised learning, Classification, Computer Aided Diagnosis, Breast cancer detection

1. Introduction

The use of Computer-Aided Diagnosis (CAD) tools in medical imaging is nowadays a common practice that has proved to be useful in enhancing the detection process of several diseases. Trained physicians take CAD systems' output as a second opinion for them to perform a more accurate prognosis of the encountered pathologies.

Breast cancer diagnosis is a clinical field in which CAD systems have been introduced with the objective of assisting radiologists during the detection process. An accurate diagnosis in the early stages of this disease is of major importance, since there exist medical treatment can be followed to enhance patient survival [1]. There are several tools for breast cancer screening, but *mammography* is the most common technology in *early* stages, since it can reveal lesions that include calcifications, masses, architectural distortions and asymmetric densities in breast tissue, even before the become palpable [2].

The use of computational systems represents one powerful option for analyzing biomedical images since they provide the physician with information that might be important but not easily observed. Regarding breast cancer screening, studies show that the use of CAD tools for cancer diagnosis improves the sensitivity of conventional reading [3].

Therefore, several research efforts have been conducted toward the implementation of efficient CAD systems, often based on Artificial Intelligence (AI) techniques, in which the objective is to provide the physician with an automatic

[*] Corresponding author

Email addresses: eden.alanis@itesm.mx (Edén A. Alanís-Reyes), sigfrido@itesm.mx (Sigfrido Iglesias-González), terashima@itesm.mx (Hugo Terashima-Marín), sconant@itesm.mx (Santiago E. Conant-Pablos), bebustos@dcc.uchile.cl (Benjamin Bustos)

detection of suspicious regions of a mammogram that may contain a lesion and its malignancy assessment, based on its visual features, with the final objective of improving the overall accuracy of the test.

Regarding research on breast cancer computerized prognosis, a broad variety of methods have been considered [4, 5, 6, 7, 8], going from image processing techniques that automatically detect suspicious regions, to the classification of cases using AI-based algorithms that assess the malignancy of a given lesion.

In this research work we implemented a CAD system for the diagnosis of breast cancer using digital mammography, based on a Case-Based Reasoning (CBR) approach in which a query image is used to *retrieve* from a database of historical cases those similar to it, in order to *re-use* them for training classification algorithms that will discriminate between benign and malignant cases. We used a kd-tree to index the database upon which we implemented a knn-similarity-search for the retrieval of similar cases; six different dissimilarity metrics were evaluated to select the one that fits best within the retrieval mechanism.

We use a binary genetic algorithm, within a wrapper approach, to explore the feature space and find an optimal subset that provides the highest discriminant power for a given classifier. We integrate the similarity-search mechanism into this feature selection process and compute the similarity score on the resulting lower-dimensional feature vectors.

Our CAD focuses in the detection and diagnosis of microcalcification clusters (MCCs) and masses, which are known to be the two major indicators of malignancy [6]. Using these two datasets, the classification of cases into benign or malignant is performed using four classification methods: knn-classifier (k-NN), artificial neural networks (NN), support vector machines (SVM) and linear discriminant analysis (LDA).

The system's output provides the physician with a malignancy assessment of the lesions found in the query case as well as the set of similar cases that were retrieved from the database.

The primary contributions of the paper are as follows:

1. The paper tackles the problem of classifying breast cancer lesions from a Case-Based Reasoning perspective, in which there exists a database of historical cases that were previously revised by an expert and serves as the ground truth of the system. Query images are segmented and every encountered lesion is considered a new case. The system performs a similarity search on the case base and the *k* most similar ones are re-used to train classification algorithms to assess the malignancy of query cases.
2. The paper presents the results of an empirical study to select, by way of pairwise statistical tests, the most suitable similarity metric to accurately perform the retrieval of historical cases on each of our datasets. We used the classification performance obtained with a majority vote k-NN classifier to evaluate six different metrics that were considered to conduct the similarity-search on the feature space of each dataset.
3. Using the most accurate similarity metric, we conducted a second empirical study that integrates the four aforementioned classifiers into our CBR-based classification scheme. In this process we re-use the retrieved *k* similar cases to train the classifiers and evaluate their performance with a leave-one-out cross-validation scheme. We also determined the optimal amount of neighbours for each of the classifiers, based on the observed performance results.

The paper is organized as follows. First, we present a description of related research efforts. Section 3 presents the methods and materials that were used in our work to design the CBR processes related to *retrieving* and *re-using* similar cases for breast cancer diagnosis. In section 4 the experimentation phase is presented, showing the evaluation of the dissimilarity metrics that were considered as well as the performance of the aforementioned classifiers of interest, when diagnosing query lesions using a set of similar cases. Finally, we present conclusions and suggest future work in section 5.

2. Related Work

In breast cancer screening, an accurate diagnosis is difficult to achieve, because there is a wide range of features that can indicate malignancy, but not all the changes that can be observed in breast tissue are necessarily malignant. As a result, 65% to 90% of the biopsies of suspicious lesions turn out to be benign [9].

Therefore, an important consideration during the design of a CAD system is to implement processes that aim to determine the set of visual features that are more relevant for assessing malignancy. This *feature selection* phase also serves to decrease the dimensionality of the feature set and has been conducted in several ways and using different techniques.

Particle Swarm Optimization (PSO) is used by Zyout et al. [8] to select parameters in an heuristic way and implements a k-NN approach for the classification of lesions. Sequential Forward Selection (SFS), Sequential Backward Selection (SBS) and statistical tests were performed by Nandi et al. [10] to determine the best features to conduct Genetic-Programming-based classification.

A genetic algorithm (GA) implemented by Verma and Zhang [11] to determine the best combination of features by using the performance of a NN as fitness function for individuals. Conant-Pablos et al. [12] also explored a GA to find the most relevant features extracted from both individual microcalcifications and microcalcification clusters (MCC), and compared their results against the methods used by Cantú-Paz [13] and Cantú-Paz et al. [14], where features are ordered according to their *class separability*.

Once the subset of features that provide the highest discriminant power has been selected, a *classification* phase is conducted to determine the actual malignancy of lesions. Hernández-Cisneros and Terashima-Marín [15] proposed a procedure for the classification of MCCs in mammograms using three Evolutionary Artificial Neural Networks (EANNs) that were built using a GA to optimize different parameters, such as connection weights and topology. Zhang et al. [7] proposed a SVM-based approach to distinguish microcalcifications from other ROIs, and Hadjiiski et al. [16] presented a classification scheme for mammographic masses based on interval change information. An extensive study of classifier algorithms is presented in [5] by Kumar et al, including SVM, ANN, Bayesian and kNN techniques.

Furthermore, the use of Case-Based Reasoning [17] has also been explored in different medical and industrial applications [18]. CBR is a methodology that solves new problems by taking advantage of previous experience, based on the principle that similar cases have similar solutions. It is a problem solving approach in which problems, called cases, are solved by re-using and, if necessary, adapting previous solutions that were conveniently stored in a database and that are similar to the new case.

It is different from other AI-based techniques in the sense that it works only with a specific set of cases from its knowledge base and uses their information to solve new cases. In this way, CBR focus in extracting knowledge only from the most relevant instances contained in their training set and, if the case base is broad enough, the accuracy that can be achieved by implementing a CBR-based approach can be high [19].

The CBR cycle is composed of four basic tasks: retrieve, reuse, revise and retain. Typically, when a new case is provided to query the system, it is pre-processed in order to construct an appropriate representation of it, which is then interpreted in the reasoning cycle, considering that the abstraction of the original problem should match the knowledge representation of the case base. A description of the CBR stages is provided in the following [20]:

1. **Retrieve:** A similarity-search is conducted over the database for the set of cases which are most similar to the query case. In this stage the system applies a similarity measure between the new case and those stored in the database. One of the most common similarity-search methods is the *nearest neighbors*, in which typically the similarity between the query case and those stored in the database is determined by a metric such as cosine similarity, manhattan distance, correlation factor, among others, which are calculated upon the vector of visual features that were extracted from the images. In medical-imaging applications, the retrieval of similar cases is performed by implementing Content-Based Image Retrieval (CBIR) on a database that contains, both, images and the metadata that are used to compute similarity-search [21].
2. **Reuse:** Once the set of similar cases has been retrieved, the system reuses the information of one or more of them, by way of a interpretation or possible adaptation, in order to provide a solution to the new problem which is the main objective of this stage; the approach that will be used to reuse the information of retrieved cases depends on the nature of the problem that is going to be solved, being adaptation the most frequently used in diagnosis tasks. However, not all CBR systems provide a suggested solution, but focus only in retrieving the most relevant cases, as a form case-based retrieval systems.
3. **Revise:** This step evaluates the solution suggested by the system in a real world scenario or against the revision of an expert which will reject, correct or confirm it. In CBR both correct and incorrect solutions are equally important, since they represent the experience that is drawn from by the system from the its knowledge base, but the latter have to be identified and repaired to prevent failing again in future.
4. **Retain:** The revised solution is stored as a new case in the database for future problem solving. This is why it is important to validate the fitness of the proposed solutions; otherwise, reusing historical information would not be useful to compute a solution and the accuracy of the system would be compromised. With this process the

ground truth of the system increases which results in the fact that subsequent queries are solved over a broader experience.

CBIR systems have been applied in different clinical applications [22], including breast cancer diagnosis by mammography [23], in which the fundamental process is to retrieve historical images containing lesions that are similar to the ones detected in a query image provided by a physician and, also, reuse them for classifying lesions [24, 25].

Medical diagnosis systems based on a CBR methodology have been developed for different domains. Armengol [26], proposed a method for classifying melanomas *in situ* using clustering techniques under the CBR philosophy. In this work, the clustering algorithm is not used to organize the case base to enable an efficient retrieval mechanism; rather, they proposed a method in which CBR is used for clustering in which a lazy learning method produces *explanations* that are used as clusters' descriptors. This way, clusters can show the expert a picture of some parts of the domain of the problem, with its main characteristics.

A biofeedback training method proposed by Ahmed et al. [27] uses modified distance function, similarity matrix and fuzzy similarity for case-retrieval. The system receives a time series signal related to finger temperature and uses a CBR approach to support classification of patients, parameter estimation and biofeedback training in stress medicine and can be queried by less experienced physicians who can also use this system as a decision support tool which provides a second opinion in diagnosis tasks. In every module of this system, the CBR technique retrieves the most similar cases from the knowledge base by comparing a new finger temperature reading with previously solved measurements.

A cost-sensitive learning approach was presented by Park et al. [28], incorporating unequal penalizations for misclassifying positive or negative cases, all within a CBR model. This research work describes a method based on a GA to find the optimal cut-off point to discriminate between malignant and benign cases, as well as the cut-off point for determining the optimal number of neighbors that should be retrieved from the case base, considering their similarity scores. This cost-sensitive method allows to take into account that false-negatives have more severe implications than false-negatives and that, therefore, they should receive different penalization costs.

Furthermore, databases are typically indexed to enable an efficient search for case retrieval. Common indexing techniques include kd-tree [29] and Locality Sensitive Hashing (LSH) [30, 31, 32]. The former creates a partition of the feature space using a tree structure; the latter uses hash functions to compute the distance between the query and a subset of reference points.

3. Methods and Materials

In this research effort we present a classification methodology that implements key processes of the Case-Based Reasoning (CBR) paradigm, applied to a dataset of features extracted from breast cancer lesions found in a set of mammographic images that are contained in a database.

Microcalcification clusters and masses are automatically detected and segmented in digital mammographies in the way described in a previous work reported by Alanís-Reyes et al. [33], which used the Mammographic Image Analysis Society (MIAS) [34] database. With those methods, we detected 38 microcalcification clusters, out of which 9 are positive cases and 29 negative; on the other hand, we detected 19 positive cases of tumoral masses and 33 negative ones, adding up to a dataset of 52 cases.

The set of features listed by Conant-Pablos et al. [12] is computed upon those datasets, to characterize the visual content of the lesions. The methods presented in this document are performed using this feature set.

Figure 1 depicts the proposed CBR model that we have designed to support medical diagnosis of breast cancer lesions. A key component of this model is the database which contains information of historical cases with previously diagnosed pathologies. Considering the aforementioned feature set, our system **retrieves** from the database a subset of k lesions that are similar to the ones found in a query image, by performing a k-NN-based similarity search. Finally, the retrieved cases are **reused** to compute the diagnosis of the encountered lesions, based on four classifiers of interest, namely: the k-nearest neighbors (k-NN) [35], a neural network (NN) [36], support vector machine (SVM) classifier [37] and the linear discriminant analysis (LDA) method [35, 38].

A further description of all inner processes within the model presented in this paper is provided in the following sections.

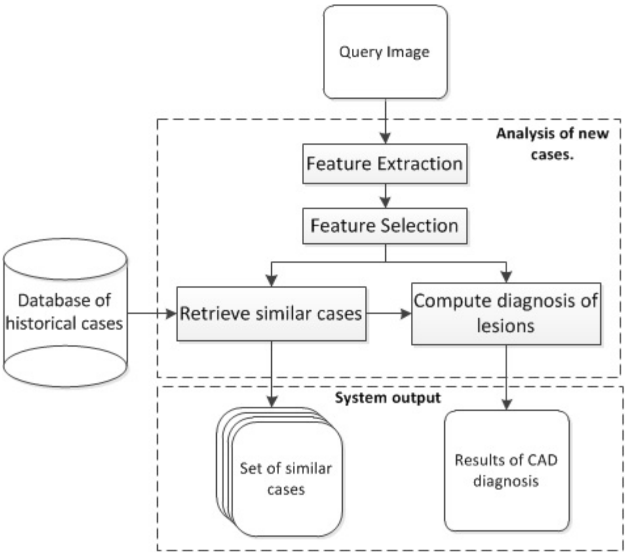


Figure 1: Overall model

3.1. Feature Extraction

After segmenting the query mammogram to extract microcalcification clusters and masses as regions of interest (ROIs) the way described in [33], a feature extraction process is computed on them in order to obtain a vector that describes their visual content. This feature vector is then used for retrieving similar historic cases and classifying lesions into benign or malignant, as we will describe further in the next sections.

Features from detected microcalcification clusters and masses are computed separately. The set of 31 features listed in Table 1 are computed for microcalcification clusters. It can be observed that 14 of them are related to the shape of the cluster, 6 describe the area of individual microcalcifications and 11 define the absolute contrast between them and their background.

	Computed Features
Cluster shape (14 features)	Number of calcifications, convex perimeter, convex area, compactness, microcalcification density, total radius, maximum radius, minimum radius, mean radius, standard deviation of radii, maximum diameter, minimum diameter, mean of the distances between microcalcifications, standard deviation of the distances between microcalcifications.
Area of Microcalcifications (6 features)	Total area of microcalcifications, mean area of microcalcifications, standard deviation of the area of microcalcifications, maximum area of the microcalcifications, minimum area of the microcalcifications, relative area.
Microcalcification Contrast (11 features)	Total gray mean level of microcalcifications, mean of the mean gray levels of microcalcifications, standard deviation of the mean gray levels of microcalcifications, maximum mean gray level of microcalcifications, minimum mean gray level of microcalcifications, median of the man tray level of microcalcifications, total absolute contrast, mean absolute contrast, standard deviation of the absolute contrast, maximum absolute contrast, minimum absolute contrast.

Table 1: Features extracted from microcalcification Clusters

Furthermore, the set of 50 features presented in Table 2 are extracted from ROIs containing a mass. In this set, 7 features describe the signal contrast, 7 characterize the background contrast, 3 of them define the relative contrast and 20 the shape of the mass; additionally, there are 6 features about the sequence moments ant 7 regarding the first invariant moments.

	Computed Features
Signal contrast (7 features)	Maximum gray level, Minimum gray level, Median gray level, Mean gray level, Standard deviation of the gray level, Gray level asymmetry (skewness), Kurtosis of gray level.
Background contrast (7 features)	Maximum gray level, Minimum gray level, Median gray level, Mean gray level, Standard deviation of the gray level, Gray level asymmetry (skewness), Kurtosis of gray level.
Relative Contrast (3 features)	Absolute contrast, Relative contrast, Portional contrast.
Shape (20 features)	Area, convex area, background area, filled area, perimeter, maximum diameter, minimum diameter, orientation, eccentricity, Euler number, circular diameter equivalent, solidity, Extent, shape factor, roundness, aspect ratio, elongation, compactness 1, compactness 2, compactness 3.
Contour sequence moment (6 features)	Contour sequence moment 1, contour sequence moment 2, contour sequence moment 3, contour sequence moment 4, mean radii, standard deviation of radii.
First invariant moments (7 features)	Invariant moment 1, invariant moment 2, invariant moment 3, invariant moment 4, invariant moment 5, invariant moment 6, invariant moment 7.

Table 2: Features extracted from Masses

3.2. Feature Selection

The next step of the analysis consists in taking the set of features extracted for each ROI into a feature selection process. The main purpose of this phase is to find the subset of the whole set of features that most contribute to the performance of a given classifier and, also, has a reduced dimensionality.

This process was carried out using a wrapper approach based on a GA and four different classifiers, namely: the k-nearest neighbors (k-NN) [35], a feed-forward back-propagation neural network (NN) [36], support vector machine (SVM) classifier [37] and the linear discriminant analysis (LDA) method [35, 38].

The implemented wrapper used a GA as search algorithm to explore through the space of possible feature subsets, taking advantage from its ability to exploit accumulating information about an initially unknown search space in order to bias subsequent search into promising subspaces [39]. It works with a population of candidate solutions, or individuals, and through the generations looks for the fittest individual.

In our feature selection process each individual of the population represent a subset of features. The chromosomes of the individuals in the GA contain an amount of bits equal to the total number of features, i.e. one bit for each extracted feature. Consequently, the chromosomes of the individuals in this study had lengths of 30 for clusters of microcalcifications and 50 for masses, corresponding to the amount of features extracted for each kind of lesion.

The value of each bit within the chromosome determines whether feature will be selected or not, and therefore if it will be considered in the subset of features provided as input to the classifier. Figure 2 shows an individual that determines that features 1, 3 and *n* were selected to be part of the classifier input, while feature 2 was not.

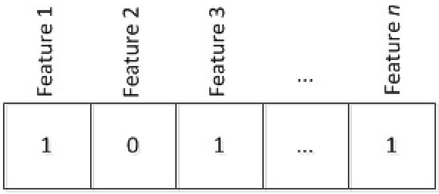


Figure 2: A sample individual of the GA which determines that Feature 1, Feature 3 and Feature *n* are selected.

The evaluation process by which the GA computes the fitness of each individual is depicted in Figure 3. For any given query case, the subset of features determined by each individual is taken to perform a k-nearest neighbors similarity search through database of historical cases. The result of that search is a set of *k* cases that are similar to the query case and which are used to construct and train a classifier algorithm. This process is performed individually for the four classification methods considered in this study.

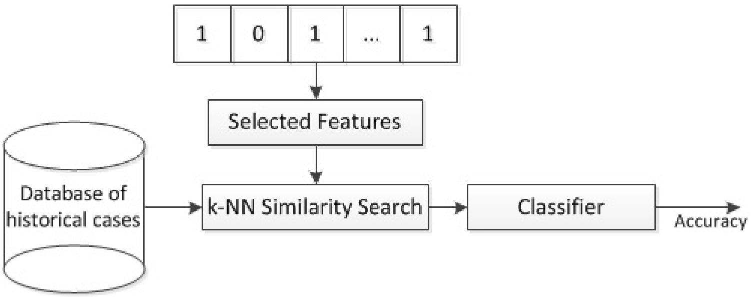


Figure 3: Computing the fitness of individuals in GA-based Feature Selection

Afterwards, a leave-one-out cross-validation through the whole database of historical cases is performed to compute the individual's fitness, which is determined by the AUC that a given classifier achieves if it uses the subset of features determined by the individual, which can be approximated as follows:

$$\text{AUC} = \frac{1 + \text{Sensitivity} - \text{FP}_{rate}}{2} \tag{1}$$

where the Sensitivity, or true positive rate, is defined as:

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{2}$$

and the false positive rate, FP_{rate}, is computed as:

$$\text{FP}_{rate} = \frac{\text{false positives}}{\text{true negatives} + \text{false positives}} \tag{3}$$

The GA can stop due to two reasons: either the generations' limit or the maximum number of evaluations of the fitness function has been reached. The best individual of this evolutionary process determines the subset of features that have the highest discriminant power with respect to a given classifier.

3.3. Retrieving Similar Cases and Reusing them to Compute Diagnosis of lesions

Retrieving and reusing similar cases are key processes within the CBR paradigm, since they enable solving new problems by looking at relevant information that can be drawn from the solutions of similar cases that were revised and validated by an expert sometime in the past. In this section we will explain how these processes are conducted within our model.

Our similarity search is computed upon the feature-vector that contains only the most discriminant ones, which were selected by the GA-based feature selection method. Therefore, it should be noted that in this work, *feature selection* is not only of major importance for optimizing the classification accuracy of the model, but it also enables a more efficient similarity search in the database, due to the associated dimensionality reduction.

The database of our model contains a set of mammographic images and the feature vector of the lesion that was found in each one. It is indexed by a kd-tree upon which the retrieval of cases is performed in a *k-Nearest Neighbors Similarity Search* paradigm, as depicted in Figure 4, with the objective of optimizing the retrieval process. It can also be observed that the activities involved in this stage are grouped in two different procedures, corresponding to the retrieval of similar cases and the computing the diagnosis of the new case, by training classifiers with the retrieved instances.

As we depict in Algorithm 1, in order to retrieve similar cases, the *dissimilarity score* between the new case and those stored in the database is computed by applying a distance metric to the feature vector of the query case **x** and any feature vector stored in the database. In this study, we explore the performance of six different dissimilarity metrics, namely: euclidean distance, manhattan distance, linear correlation, cosine similarity, Mahalanobis distance and the

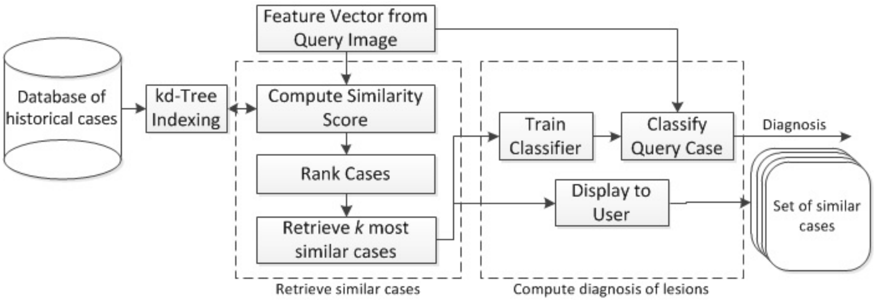


Figure 4: Retrieving similar cases and reusing them for classification of lesions

Spearman distance. We measured the AUC obtained by a majority-vote k-NN classifier that uses the considered dissimilarity metrics and conducted pairwise statistical tests between them to determine the one that is more suitable for the retrieval of cases.

Algorithm 1 Proposed CBR classification methodology

```
Require: $x$, the feature vector of the query case, $s$, the similarity metric, $k$, the number of similar cases to be considered, $Idx$ the kd-tree index of the database of historical cases and $c$, the classifier to be used.
procedure RetrieveSimilarCases($x$, $s$, $k$, $Idx$)
    for all $case \in Idx$ do
        $case.similarity \leftarrow computeSimilarity(x, s)$
    end for
    $rankedCases \leftarrow rankBySimilarity(Idx)$
    $kMostSimilar \leftarrow selectTopK(rankedCases, k)$
    $ReuseSimilarCases(kMostSimilar, x, c)$
end procedure

Require: $T$, the set of $k$ most similar cases retrieved from the data base, $x$, the feature vector of the query case and $c$, the classifier to be used.
procedure ReuseSimilarCases($T$, $x$, c)
    $classifier \leftarrow train(T, c)$
    $Dx \leftarrow test(classifier, x)$
    return $Dx$
end procedure
```

Then, our methodology ranks the set of stored instances in descending order, based on the the similarity score that was previously computed between the query case and the historical ones and, finally, the *k* most similar ones are reused in computing the diagnosis of detected breast cancer lesions.

This assessment of malignancy is performed by taking the *k* retrieved instances as the training set of a given classifier algorithm. In this way, the classifier is going to be fed a set of relevant historical cases that contain data from lesions that have similar features to those encountered in the query case, focusing the learning process in a subset of training samples within a neighbourhood of similarity instead of the whole dataset. This training process is executed for every new query case, providing the classifiers with the ability to adapt to the new problem, by re-using each time the most relevant instances to solve it.

To compute the diagnosis of lesions we are considering four classifiers: k-NN, NN, SVM and LDA, as we previously mentioned. Each of them are tested separately within our CBR framework with the aforementioned feature selection and training strategy. The performance measures that are considered for evaluating classifiers are the AUC (see equation 1), sensitivity (see equation 2), and overall accuracy and specificity computed as follows:

$$\text{Accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{total number of samples}} \tag{4}$$

$$\text{Specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \tag{5}$$

Moreover, it is worth mentioning that our CBR model is not dependant of specific classification algorithms, but rather describes a series of processes that are aimed to the classification of breast cancer cases in which learning algorithms are integrated as a tool to discriminate between benign and malignant lesions.

In this paper we have integrated four different types of classifier algorithms, so as to explore the performance that can be achieved within the proposed CBR framework, from different learning perspectives. The NN is a robust, *non-linear* classifier that is highly adaptable and has been tested in several classification domains; the SVM represents a *kernel-based* classifier that has strong theoretical foundations and a good generalization capability and has recently become very popular as a high-performance classifier in several domains, as well. A *linear* approach is explored by integrating the LDA and a simplistic *majority-vote* strategy is represented by the k-NN algorithm.

4. Results and Discussion

As we mentioned previously, the database of digital mammographies used in this research work was provided by the Mammographic Image Analysis Society (MIAS) [34], out of which we detected a set of 38 microcalcification clusters and 52 masses. All experiments were performed in MATLAB Version 7.8.0.347 (R2009a), under a 2.8 GHz Intel Xeon processor with 3.48 GB of memory.

The aforementioned wrapper-based feature selection process was performed separately for MCCs and masses. In the case of microcalcification clusters, individuals within the GA had chromosomes of 31 bits of length, while, for tumoral masses, individual had a length of 50 bits. In both cases, we used a simple GA with a population of 200 individuals, 100 generations, binary tournament selection and two-point crossover. The initial population was initialized uniformly at random.

The experiments have three objectives: (1) to determine the most suitable dissimilarity metric for the retrieval task, (2) to determine the optimal amount of k cases to be retrieved/re-used from the database and (3) to find the most relevant subset of features, in order to obtain the highest performance of each of the classification algorithms considered in this study. Therefore, we first explore the performance of six different dissimilarity metrics. Then, we perform feature selection and classification accuracy assessment for $k = 3, 5, 7, 9, 11, 13, 15, 17, 19, 21$. We then analyse which k value provides the highest performance and present the related subset of selected features.

The whole process was performed four times for every k, one for each of the classifiers of interest (k-NN, NN, SVM, LDA). The k-NN performed a majority-rule classification with respect to the k retrieved cases. The NN had one hidden layer with $2n + 1$ neurones, where n is the number of input units, according Kolmogorov's theorem [40], as well a single neuron in the output layer; all neurons considered the sigmoid hyperbolic tangent as transfer function. Regarding the SVM, we used a scaling factor of $\sigma = 2.0$ in a Gaussian Radial Basis Function kernel and the *Quadratic Programming* technique to find the separating hyperplane. We also used the LDA method as a fourth classifier.

4.1. Evaluating dissimilarity metrics

In order to determine the most suitable dissimilarity metric to be used for the retrieval of similar cases from the database, we evaluated six different metrics in terms of the AUC that is obtained by performing a k-NN classification with different amounts k of retrieved instances.

Considering a query row vector $\mathbf{x}$ and any row vector $\mathbf{y}$ stored in the database, the dissimilarity metrics evaluated are:

- Euclidean distance:

$$d = \sqrt{(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})'} \tag{6}$$

Table 3: Evaluation of dissimilarity metrics considering AUC obtained by k-NN classification of MCCs.

Classifier	***k***									
	3	**5**	**7**	**9**	**11**	**13**	**15**	**17**	**19**	**21**
Euclidean	0.8103	0.8103	0.7931	$\underline{0.7759}$	0.7759	0.7069	0.6724	0.6207	0.5517	0.5345
Manhattan	0.8448	0.7931	0.7931	$\underline{0.7759}$	0.7414	0.7069	0.6724	0.6379	0.5690	0.5172
Correlation	$\underline{0.8621}$	$\underline{0.8621}$	$\underline{0.8103}$	$\underline{0.7759}$	$\underline{0.7931}$	0.7414	$\underline{0.7759}$	$\underline{0.7375}$	$\underline{0.6341}$	$\underline{0.5690}$
Cosine	0.8238	0.7720	0.7414	$\underline{0.7586}$	0.7241	0.6897	0.6897	0.6724	0.5517	0.5345
Mahalanobis	0.8103	0.7931	0.7241	0.7241	0.6379	0.5862	0.5517	0.5172	0.5172	0.5172
Spearman	0.8027	0.8276	0.7893	$\underline{0.7759}$	0.7893	$\underline{0.8276}$	0.7203	0.6686	0.5862	$\underline{0.5690}$

- Manhattan distance:

$$d = \sum_{i=1}^{n} |x_i - y_i| \tag{7}$$

- Linear correlation:

$$d = 1 - \frac{(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{y} - \bar{\mathbf{y}})'}{\sqrt{(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})'}\,\sqrt{(\mathbf{y} - \bar{\mathbf{y}})(\mathbf{y} - \bar{\mathbf{y}})'}} \tag{8}$$

- Cosine similarity:

$$d = 1 - \frac{\mathbf{x}\mathbf{y}'}{\sqrt{(\mathbf{x}\mathbf{x}')(\mathbf{y}\mathbf{y}')}} \tag{9}$$

- Mahalanobis distance:

$$d = \sqrt{(\mathbf{x} - \mathbf{y})\,\mathbf{C}^{-1}\,(\mathbf{x} - \mathbf{y})'} \tag{10}$$

  where **C** is the covariance matrix between **x** and **y**.

- Spearman distance:

$$d = \frac{\sum_i (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{11}$$

  where scores $\mathbf{x_i}$ and $\mathbf{y_i}$ are converted to ranks x_i, y_i.

Table 3 shows the results for this process, regarding the classification of MCCs. Each column shows the results of the AUC that was obtained by applying each of the considered metrics for k-NN classification and the highest AUC is underlined for each k.

We can see that the *correlation* metric consistently provided the highest accuracy, regardless of the amount of retrieved cases. It was only outperformed by the *spearman* metric ($k = 13$). They both achieved the same AUC for $k = 21$. For $k = 9$, all metrics achieved the same performance, except for *Mahalanobis* metric.

Table 4 shows the performance for the k-NN classification of masses for all the dissimilarity metrics considered. Once again, the correlation metric provided the highest AUC in almost all the experiments, except for $k = 9$, which *cosine* metric outperformed it. Moreover, for $k = 5, 13, 15$ both metrics achieved the same performance.

The dispersion of the AUC for MCCs and masses is depicted in Figure 5 and Figure 6, respectively. It can be observed that each metric has a different median performance and in both datasets correlation metric which presents the highest median.

We used Friedman's rank sum test [41] to determine if these performance differences are significant. Table 5 shows the mean ranks of AUC for each metric and the p-values of the test for both MCCs and masses. Both p-values (4.6e-09 and 4.49e-06) show significance at a 5% level.

Table 4: Evaluation of dissimilarity metrics considering AUC obtained by k-NN classification of Masses.

Classifier	***k***									
	3	**5**	**7**	**9**	**11**	**13**	**15**	**17**	**19**	**21**
Euclidean	0.7624	0.7209	0.7289	0.6986	0.6834	0.6946	0.7057	0.7209	0.6723	0.6723
Manhattan	0.7775	0.7440	0.7137	0.7209	0.7026	0.7209	0.6946	0.6946	0.6499	0.6986
Correlation	0.8381	0.7967	0.8038	0.7329	0.7624	0.7624	0.7624	0.7624	0.7663	0.7624
Cosine	0.8118	0.7967	0.7887	0.7361	0.7472	0.7624	0.7624	0.7472	0.7472	0.7361
Mahalanobis	0.6938	0.6252	0.5798	0.5606	0.5303	0.5152	0.5000	0.5000	0.5000	0.5000
Spearman	0.6611	0.6388	0.6459	0.6459	0.6388	0.6459	0.6459	0.5598	0.5303	0.5152

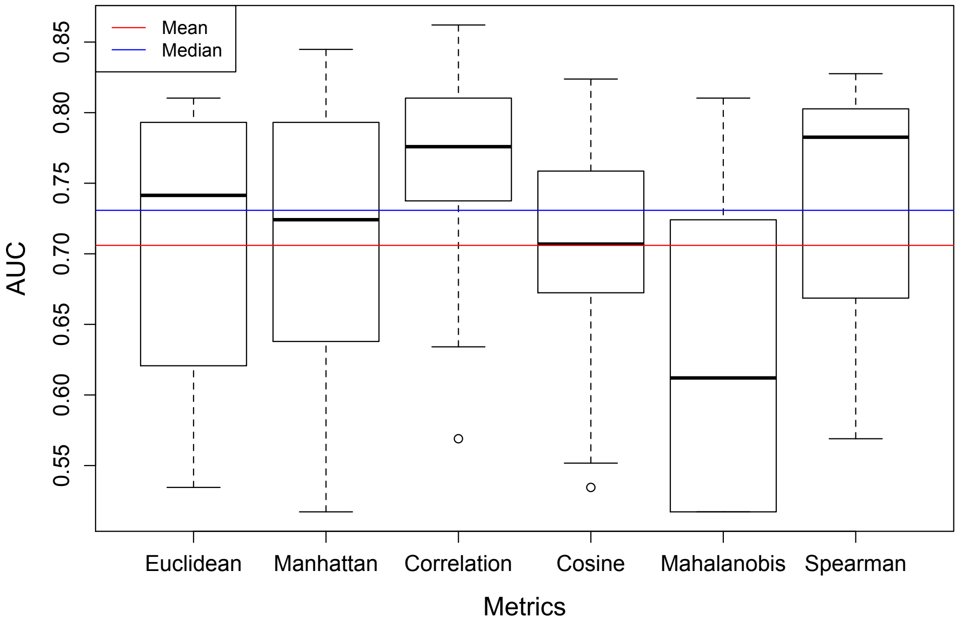


Figure 5: Boxplots of AUC performance for MCCs.

Table 5: Results of Friedman Test.

Dissimilarity metric	**Mean rank for MCCs**	**Mean rank for masses**
Euclidean	3.35	3.40
Manhattan	3.40	3.60
Correlation	5.70	5.75
Cosine	2.80	5.25
Mahalanobis	1.35	1.10
Spearman	4.40	1.90
	p-value = 4.6e-09	p-value = 4.49e-06

Multiple comparisons tests as described in [42] were performed to verify the overall mean rank outperformance of the correlation metric. We compare the absolute mean rank difference between the correlation metric and the rest, $\left|\bar{R}_c - \bar{R}_i\right|$, against $z_{\alpha/(2k)}\sqrt{\frac{t(t+1)}{6n}} = 1.821$, where $z_{\alpha/(2k)} = 2.576$ is a standard normal quantile, with $\alpha = .05$, $k = 5$

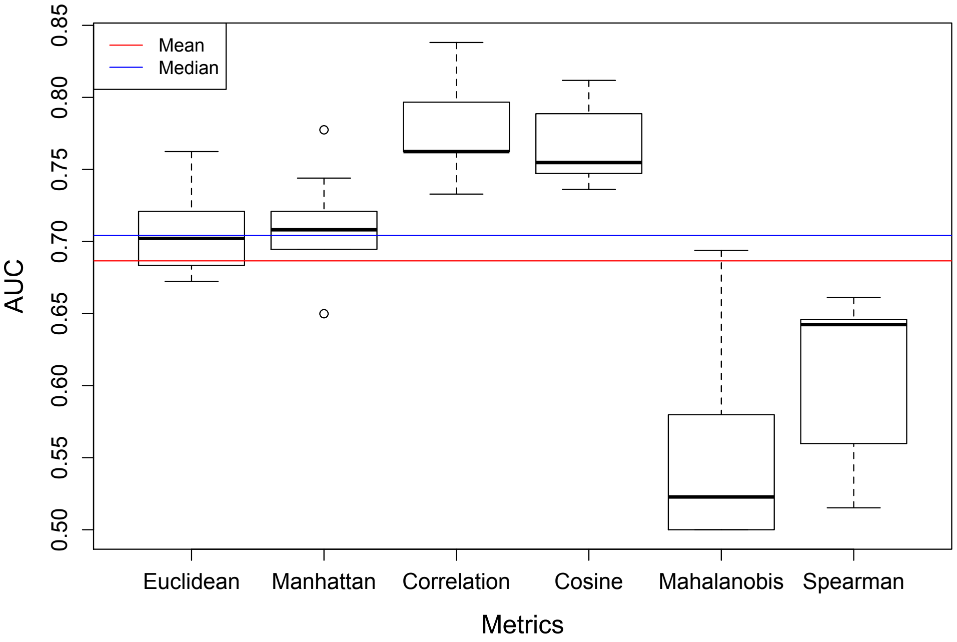


Figure 6: Boxplots of AUC performance for Masses.

Table 6: Pairwise comparison of metrics' mean ranks, with MCCs dataset.

Comparison	$\left|\bar{R}_c - \bar{R}_i\right|$
Correlation - Euclidean	2.35*
Correlation - Manhattan	2.30*
Correlation - Cosine	2.90*
Correlation - Mahalanobis	4.35*
Correlation - Spearman	1.30

* Significant at a familywise type I error rate of 5%.

comparisons, $t = 6$ number of treatments and $n = 10$ number of groups, since we are doing 5 comparisons among six metrics on ten blocks.

As shown in Table 6 and Table 7, the difference of performance between the correlation metric and the rest was significant in all cases, except for the cases in which it is compared to spearman (MMCs) and cosine metric (masses). All differences were positive. We conclude that the correlation metric has an overall tendency that outperforms the other metrics.

Based on these results, we determined to use the *correlation* metric in the similarity search process within the aforementioned CBR model, in which we train different classification algorithms with the set of similar cases that are retrieved from the database of historical cases.

4.2. Results for microcalcification clusters

We implemented a leave-one-out validation in which all the classifier algorithms were trained using the set of k similar cases retrieved from the database using the correlation metric. Table 8 shows the cross-validation results of ten different tests in which the four classifiers of interest were applied to assess the malignancy of microcalcification

Table 7: Pairwise comparison of metrics' mean ranks, with masses dataset.

Comparison	$\left\|\bar{R}_c - \bar{R}_i\right\|$
Correlation - Euclidean	2.35*
Correlation - Manhattan	2.15*
Correlation - Cosine	0.50
Correlation - Mahalanobis	4.65*
Correlation - Spearman	3.85*

* Significant at a familywise type I error rate of 5%.

Table 8: Classification performance for Microcalcification Clusters

Classifier	**Performance**	***k***									
		3	**5**	**7**	**9**	**11**	**13**	**15**	**17**	**19**	**21**
k-NN	**AUC**	0.8621	0.8621	0.8103	0.7759	0.7931	0.7414	0.7759	0.7375	0.6341	0.5690
	Accuracy	0.7895	0.7895	0.7632	0.7632	0.8158	0.7368	0.7632	0.7895	0.6842	0.5789
	Sensitivity	1.0000	1.0000	0.8889	0.7778	0.7778	0.7778	0.7778	0.6667	0.5556	0.5556
	Specificity	0.7241	0.7241	0.7241	0.7586	0.8276	0.7241	0.7586	0.8276	0.7241	0.5862
NN	**AUC**	0.8161	0.8544	0.8927	0.9272	0.9100	0.9100	0.8889	0.8889	0.9100	0.9655
	Accuracy	0.8947	0.8947	0.8947	0.9474	0.9211	0.9211	0.9474	0.9474	0.9211	0.9474
	Sensitivity	0.6667	0.7778	0.8889	0.8889	0.8889	0.8889	0.7778	0.7778	0.8889	1.0000
	Specificity	0.9656	0.9310	0.8966	0.9655	0.9310	0.9310	1.0000	1.0000	0.9310	0.9310
SVM	**AUC**	0.8544	0.8333	0.8161	0.8544	0.8717	0.8717	0.8544	0.8372	0.8717	0.8544
	Accuracy	0.8947	0.9211	0.8947	0.8947	0.9211	0.9211	0.8948	0.8684	0.9211	0.8947
	Sensitivity	0.7778	0.6667	0.6667	0.7778	0.7778	0.7778	0.7778	0.7778	0.7778	0.7778
	Specificity	0.9310	1.0000	0.9656	0.9310	0.9656	0.9656	0.9310	0.8966	0.9656	0.9310
LDA	**AUC**	0.7989	0.8372	0.8717	0.8717	0.9310	0.8966	0.8793	0.8793	0.8582	0.8582
	Accuracy	0.8684	0.8684	0.9211	0.9211	0.8948	0.8421	0.8158	0.8158	0.8421	0.8421
	Sensitivity	0.6667	0.7778	0.7778	0.7778	1.0000	1.0000	1.0000	1.0000	0.8889	0.8889
	Specificity	0.9310	0.8966	0.9656	0.9656	0.8621	0.7931	0.7586	0.7586	0.8276	0.8276

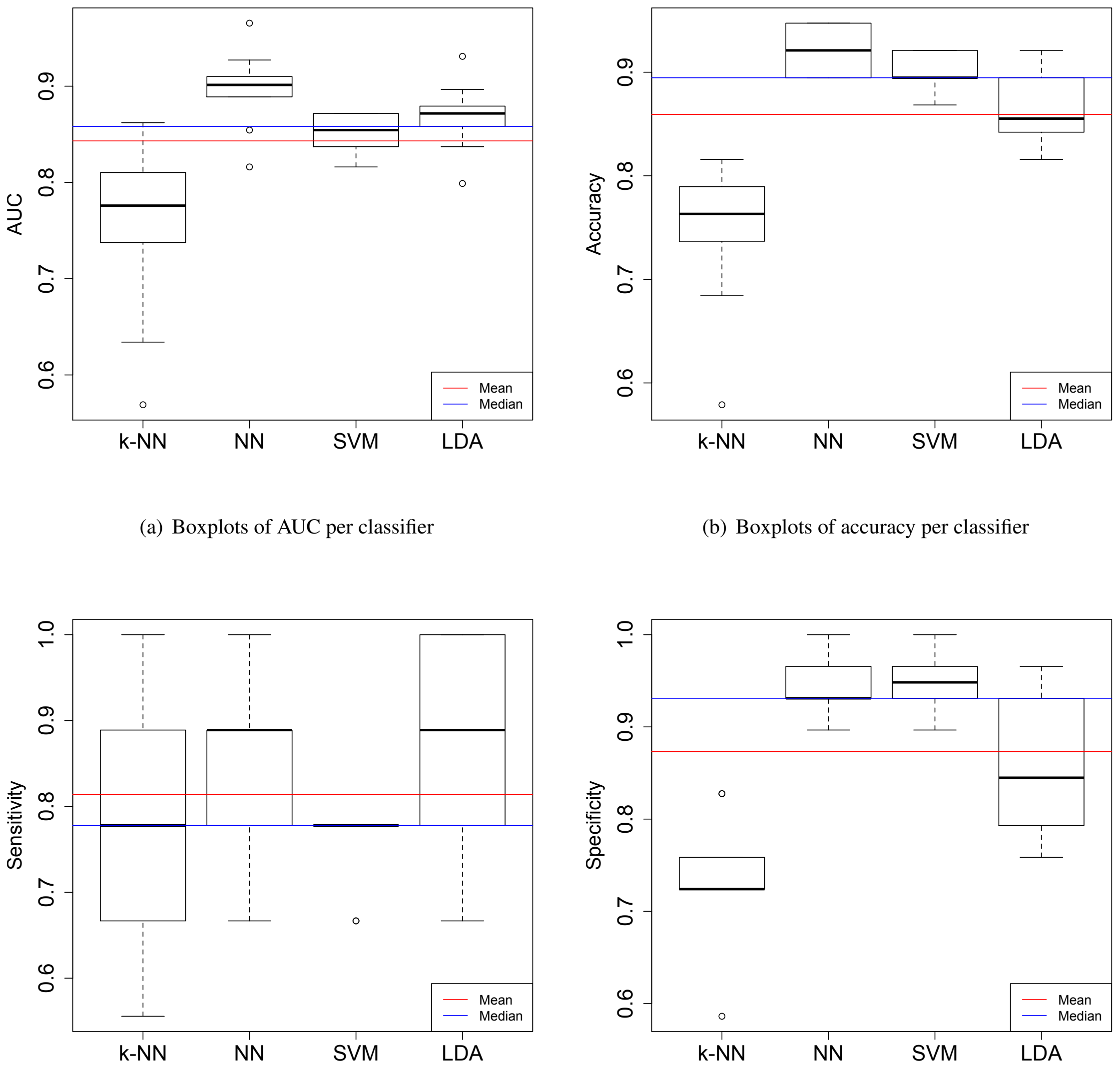
clusters, training with different amounts of k retrieved samples, ranging from $k = 3$ to $k = 21$. We measured the AUC, overall accuracy, sensitivity and specificity, and the best results for each classifier, in terms of AUC, are underlined.

Figure 7 shows the dispersion of all four performance parameters for every classifier. We can see that the performance of the CBR model varies considerably among and within classificators. Every performance measure exhibits the presence of oultiers, specially the AUC. The NN and SVM classifiers have a median performance at or above the overall median on every measure considered, with one expection in the SWM's AUC, which is just slightly below. Also, these two classifiers tend to have a smaller dispersion. As a general conclusion, we may say that the performance of the proposed CBR model is sensitive to the classifier employed.

We can observe that the k-NN presented its best performance with $k = 3, 5$ with an $AUC = 0.8621$, which consistently decreased with higher amounts of retrieved cases k, as well as the sensitivity. Furthermore, we can see in Figure 7(a) that the sensitivity of the classifier seems to be highly unstable, since it presented the most variability. The performance of the rest of parameters was more stable. Is it worth mentioning the presence of outliers at $k = 21, 11, 17$, the first one with high influence on all tests and the latter on the specificity. Therefore, regarding this classifier, it can be observed that a more accurate and stable performance can be achieved considering a lower number of training samples, which is to be expected, given the simplistic approach of a majority-vote k-NN used in this experiment.

As for the NN, the best performance was obtained with $k = 21$, providing an $AUC = 0.9655$, an overall accuracy of 0.9474, 1.00 sensitivity and 0.9310 specificity. High results were obtained in terms of specificity and overall accuracy, with a low variability, as can be observed in Figure 7(b). On the other hand, the sensitivity of this classifier presented a high variability, indicating once again that the number of retrieved samples used for training the NN affects this parameter. Moreover, it achieved higher results in terms of AUC but presented outliers for $k = 3, 5, 21$.

The SVM achieved its highest AUC, 0.8717%, with $k = 11, 13, 19$, resulting in an overall accuracy of 0.9211. Throughout the experiments the sensitivity achieved by this classifier was 0.7778 in all cases except for $k = 5, 7$, in which it decreased to 0.6667; however, these results are outliers, as can be observed in Figure 7(c). Moreover, it can be seen that the performance of this classifier was more stable across the different amounts of training samples k that



(a) Boxplots of AUC per classifier

(b) Boxplots of accuracy per classifier

(c) Boxplots of sensitivity per classifier

(d) Boxplots of specificity per classifier

Figure 7: Boxplots of performance measures for MCCs classification.

were considered in the experiments.

As for LDA, the highest AUC was 0.9310, with 0.8948 overall accuracy, 1.000 sensitivity and 0.8621 specificity for $k = 11$. Moreover, the highest sensitivity was achieved for $k = 11, 13, 15, 17$, while the highest specificity, 0.9656, was obtained with $k = 7, 9$, in which also the highest overall accuracy of 0.9211 was achieved. We can observe in Figure 7(d) that this classifier's AUC presented outliers for $k = 3, 11$ with low variability, while both sensitivity and specificity were more disperse.

Table 9 presents, both, the subset of features and the size of k retrieved training samples with which we obtained the best classification performance regarding microcalcification clusters, based on the previous results. For k-NN we determined to use $k = 3$ which provides an AUC of 0.8621, since no enhancements were found with different values and, also, considering the retrieval and re-use of only 3 cases from the database represents less computational demand.

Table 9: Features selected from Microcalcification Clusters

Classifier	k	Features Selected
k-NN	3	**Cluster shape:** convex perimeter, total radius, maximum radius, mean radius, standard deviation of radii, maximum diameter, mean of the distances between microcalcifications, standard deviation of the distances between microcalcifications. **Area of microcalfications:** total area, standard deviation of the area, relative area. **Microcalcification contrast:** total gray mean level of microcalcifications, standard deviation of the mean gray level of microcalcifications, mean absolute contrast, maximum absolute contrast.
NN	21	**Cluster shape:** convex area, microcalcification density, maximum radius, mean radius, minimum diameter.
SVM	11	**Cluster shape:** convex perimeter, convex area, minimum radius, minimum diameter.
LDA	11	**Cluster shape:** microcalcification density, total radius, minimum radius, minimum diameter.

Similarly, for the NN classifier we determined to use $k = 21$ with an AUC of 0.9655, 1.000 sensitivity and 0.9310 specificity, resulting in 0.9474 overall accuracy and for SVM and LDA we selected the results obtained with $k = 11$, with an AUC of 0.8717 and 0.9310, respectively.

4.3. Results for masses

The same GA-based feature selection process was applied for the diagnosis of masses. We performed the same leave-one-out cross-validation scheme that we described in the previous section. Table 10 presents the classification performance of the four classifiers considered, with respect to the diagnosis of tumoral masses; for each classifier, the best results in terms of AUC are underlined.

Boxplots in figure 8 show the dispersion of the various measures among classfiers. As with the MCCs, it is apparent the great sensitivity of the model to the classifier used. In contrast to the MCCs, k-NN and LDA classifiers achieved a performance overall the median levels for each measure, except for the specificy achieved by k-NN, which is slightly below. We may note as well that the median performances of each measure differ from the MCCs results. The presence of outliers is also apparent from these plots.

Figure 8(a) shows that the k-NN classifier achieved its highest AUC performance, 0.8381, for $k = 3$, with 0.7368 sensitivity and a specificity of 0.9394 which was also the highest specificity of this classifier. We can also see that the variability of its performance was small, as opposed to the behavior that we observed in classifying microcalcification clusters. This is because the number of cases within this dataset is greater and the positive and negative classes have a low imbalance ratio.

The highest performance of the NN achieved an AUC of 0.7967, achieved with $k = 7$, resulting in 0.6842 sensitivity, 0.9091 specificity and an overall accuracy of 0.8269. In this case, the highest variability of this classifier's performance was observed in terms of sensitivity and preserved a high average specificity through all experiments, as can be seen in Figure 8(b).

As with the MCCs dataset, the SVM's performance remained stable in all runs of experimentation. This can be observed in Figure 8(c), which shows that the variability of this classifiers performance was relatively small in all metrics, except for its specificity. Additionally, the AUC, overall accuracy and specificity presented an outlier in $k = 21$, as well as its sensitivity in $k = 9$. Based on the AUC parameter, the best results of this classifier were obtained at $k = 5$, resulting in and AUC of 0.7815, overall accuracy of 0.8077, with a sensitivity of 0.6842 and a specificity of 0.8788

Finally, the highest performance of the LDA was achieved with $k = 7$, with an AUC of 0.8230, overall accuracy of 0.8462, a sensitivity of 0.7368 and 0.9091 specificity. Figure 8(d) shows that this classifier presented a high variability in terms of sensitivity and specificity, with outliers in overall accuracy and sensitivity, corresponding to $k = 7$ and $k = 11$, respectively.

Table 10: Classification performance for masses

Classifier	**Performance**	***k***									
		3	**5**	**7**	**9**	**11**	**13**	**15**	**17**	**19**	**21**
k-NN	**AUC**	0.8381	0.7967	0.8038	0.7329	0.7624	0.7624	0.7624	0.7624	0.7663	0.7624
	Accuracy	0.8654	0.8269	0.8077	0.7885	0.7692	0.7692	0.7692	0.7692	0.7885	0.7692
	Sensitivity	0.7368	0.6842	0.7895	0.5263	0.7368	0.7368	0.7368	0.7368	0.6842	0.7368
	Specificity	0.9394	0.9091	0.8182	0.9394	0.7879	0.7879	0.7879	0.7879	0.8485	0.7879
NN	**AUC**	0.6874	0.7289	0.7967	0.7855	0.7137	0.7289	0.7026	0.7097	0.6874	0.6834
	Accuracy	0.7308	0.7692	0.8269	0.8269	0.7500	0.7692	0.7500	0.7308	0.7308	0.7115
	Sensitivity	0.5263	0.5789	0.6842	0.6316	0.5789	0.5789	0.5263	0.6316	0.5263	0.5789
	Specificity	0.8485	0.8788	0.9091	0.9394	0.8485	0.8788	0.8788	0.7879	0.8485	0.7879
SVM	**AUC**	0.7624	0.7815	0.7472	0.7137	0.7209	0.7472	0.7281	0.7321	0.7169	0.6675
	Accuracy	0.7692	0.8077	0.7500	0.7500	0.7308	0.7500	0.7115	0.7308	0.7115	0.6346
	Sensitivity	0.7368	0.6842	0.7368	0.5789	0.6842	0.7368	0.7895	0.7368	0.7368	0.7895
	Specificity	0.7879	0.8788	0.7576	0.8485	0.7576	0.7576	0.6667	0.7273	0.6970	0.5455
LDA	**AUC**	0.8038	0.7512	0.8230	0.7815	0.8150	0.7815	0.7775	0.7663	0.7663	0.7624
	Accuracy	0.8077	0.7692	0.8462	0.8077	0.8077	0.8077	0.7885	0.7885	0.7885	0.7692
	Sensitivity	0.7895	0.6842	0.7368	0.6842	0.8421	0.6842	0.7368	0.6842	0.6842	0.7368
	Specificity	0.8182	0.8182	0.9091	0.8788	0.7879	0.8788	0.8182	0.8485	0.8485	0.7879

Table 11: Features selected from Masses

Classifier	***k***	**Features Selected**
k-NN	3	**Signal contrast:** Standard deviation of the gray level, Kurtosis of gray level. **Background contrast:** Maximum gray level, Minimum gray level.
NN	7	**Background contrast:** Median gray level, Mean gray level, Gray level asymmetry (skewness), Kurtosis of gray level. **Relative contrast:** 'Portional contrast'. **Shape:** Area.
SVM	5	**Background contrast:** Minimum gray level, Mean gray level, Gray level asymmetry (skewness), Kurtosis of gray level.
LDA	7	**Background contrast:** Gray level asymmetry (skewness). **Relative contrast:** Absolute contrast, Relative contrast, Portional contrast. **Shape:** Convex area, background area.

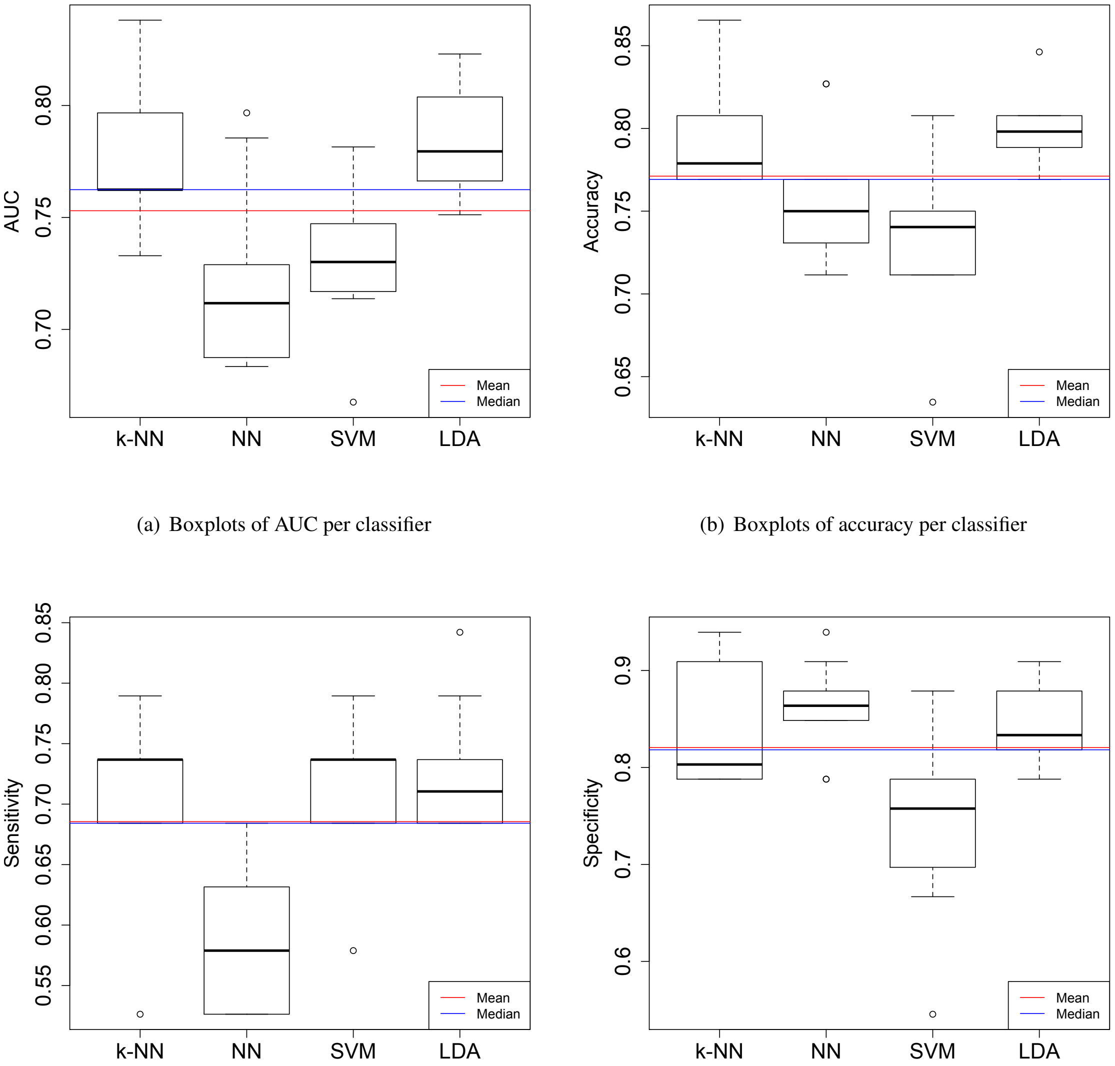
Table 11 shows the set of selected features for the best run of each classifier (i.e. that in which the classifier achieved its best performance considering the AUC), which were selected based on the previous cross-validation results. We determined that the most desirable performance was obtained with $k = 3$ for the k-NN, resulting in a $AUC = 0.8381$, with $k = 7$ for NN and LDA algorithms presenting a 0.7967 and 0.8230 AUC, respectively and, finally, regarding the SVM, the optimal value was $k = 5$ with 0.7815 AUC.

5. Conclusions and Future Work

In this study we successfully implemented two key processes of the CBR methodology: *retrieval* and *re-use* of similar cases that are stored in a database of historical data. We applied them to the classification of breast cancer lesions that include MCCs and masses.

We explored the performance of six different dissimilarity metrics applied to compute the retrieval of similar cases within an indexed *k-nearest-neighbors* similarity search, based on the vectors of visual features extracted from the lesions related to MCCs and masses. We used the AUC of a k-NN classifier to evaluate the precision of dissimilarity metrics and, afterwards, we used the Friedman test to determine if the difference of performance between all of them was significant. We then performed a pairwise comparison of the metrics' mean ranks to test if the overall mean outperformance of the *correlation* metric was statistically significant. Based on these results, we determined to use *correlation* for our similarity search for both datasets.

A cross-validation was carried on four classifiers algorithms (k-NN, NN, SVM and LDA), which were trained by *re-using* a set of k similar cases retrieved from our historical database. We considered several runs for this test in order to explore the performance of the classifiers across training sets of different sizes to finally be able to determine



(a) Boxplots of AUC per classifier

(b) Boxplots of accuracy per classifier

(c) Boxplots of sensitivity per classifier

(d) Boxplots of specificity per classifier

Figure 8: Boxplots of performance measures for mass-classification.

the amount of training samples that provided the best performance for each algorithm. Regarding MCCs, the best performance was obtained with $k = 5$ for k-NN, $k = 21$ for NN and $k = 11$ for SVM and LDA, with an AUC of 0.8621, 0.9655, 0.8717 and 0.9310, respectively. On the other hand, the NN and LDA with $k = 7$ presented the highest mass-classification performance, while the k-NN performed best with $k = 3$ and the SVM with $k = 5$, achieving an AUC of 0.7967, 0.8230, 0.8381 and 0.7815, respectively.

Since in this research work we explored the performance of several classification algorithms applied individually in the computerized diagnosis of breast cancer lesions, future work will conduct an implementation of a classifier ensemble which will combine the outputs of the four algorithms considered in this study in order to issue a unique and more accurate classification of cases.

Assessing the sensitivity of the CBR model to the choice of classifier is also of primary concern and this can be better addressed by controlling the variability of the performance measures. A study will be conducted in order to overcome issues related to the class-imbalanced nature of our datasets, which is the main source of outliers and variation in the results. By modifying pre-processing techniques and including cost-sensitive learning techniques within the retrieval mechanism of the model we aim to generate training sets containing not only similar cases, but also a balanced number of instances from both positive and negative class.

Acknowledgment

This research work was supported by the National Council for Science and Technology (CONACyT) from Mexico. We would like to thank the support of the ITESM Strategic Project PRY-075 and PRISMA Research Group of the University of Chile.

[1] American Cancer Society, The importance of finding breast cancer early, `http://www.cancer.org/Cancer/BreastCancer/MoreInformation/BreastCancerEarlyDetection/breast-cancer-early-detection-importance-of-finding-early`, 2012.

[2] C. Tukington, K. Krag, Encyclopedia of Breast Cancer, Facts on File Library of Health and Living, 2005.

[3] S. Ciatto, M. D. Turco, G. Risso, S. Catarzi, R. Bonardi, V. Viterbo, P. Gnutti, B. Guglielmoni, L. Ponelli, A. Pandiscia, F. Navarra, A. Lauria, R. Palmiero, P. Indovina, Comparison of standard reading and computer aided detection (cad) on a national proficiency test of screening mammography, European Journal of Radiology 45 (2003) 135–138.

[4] S. Deepa, B. A. Devi, A survey on artificial intelligence approaches for medical image classification, volume 4, Indian Journal of Science and Technology, 2011, pp. 1583–1595.

[5] A. K. M.n, H. S. Sheshadri, Article: On the classification of imbalanced datasets, International Journal of Computer Applications 44 (2012) 1–7. Published by Foundation of Computer Science, New York, USA.

[6] M. Rizzi, M. D. Aloia, B. Castagnolo, Health Care CAD Systems for Breast Microcalcification Cluster Detection, volume 32, Journal of Medical and Biological Engineering, 2012, pp. 147–156.

[7] X. Zhang, X. Gao, Y. Wang, MCs Detection with Combined Image Features and Twin Support Vector Machines, volume 4, Journal of Computers, 2009, pp. 215–221.

[8] I. Zyout, I. Abdel-Qader, C. Jacobs, Embedded feature selection using pso-knn: Shape-based diagnosis of microcalcification clusters in mammography, JUSPN (2011) 7–11.

[9] H. D. Cheng, X. J. Shi, R. Min, L. M. Hu, X. P. Cai, H. N. Du, Approaches for automated detection and classification of masses in mammograms, Pattern Recognition 39 (2006) 646–668.

[10] R. Nandi, A. Nandi, R. Rangayyan, D. Scutt, Classification of breast masses in mammograms using genetic programming and feature selection, Medical and Biological Engineering and Computing 44 (2006) 683–694.

[11] B. Verma, P. Zhang, A novel neural-genetic algorithm to find the most significant combination of features in digital mammograms, Applied Soft Computing Journal 7 (2007) 612–625.

[12] S. E. Conant-Pablos, R. R. Hernández-Cisneros, H. Terashima-Marín, Feature Selection for the Classification of Digital Mammograms using Genetic Algorithms, Sequential Search and Class Separability., Genetic and Evolutionary Computation: Medical Applications. S. Smith and S. Cagnoni, Wiley, 2010.

[13] E. Cantú-Paz, Feature subset selections, class separability and genetic algorithms, in: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), 2004, pp. 957–970.

[14] E. Cantú-Paz, S. Newsam, C. Kamath, Feature selection in scientific applications, in: Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 788–793.

[15] R. R. Hernández-Cisneros, H. Terashima-Marín, Evolutionary neural networks applied to the classification of microcalcification clusters in digital mammograms, in: Proceedings of the 2006 IEEE Congress on Evolutionary Computation, Vancouver, BC, Canada, 2006, pp. 2459–2466.

[16] L. Hadjiiski, B. Sahiner, H. Chan, N. Petrick, M. Helvie, M. Gurcan, Analysis of temporal changes of mammographic features: Computer-aided classification of malignant and benign breast masses, Medical Physics 28 (2001) 2309–2317.

[17] J. Kolodner, Case-based reasoning, Morgan Kauffman, San Mateo, 1993.

[18] M. U. Ahmed, S. Begum, E. Olsson, N. Xiong, P. Funk, Successful Case-based Reasoning Applications, Studies in Computational Intelligence, Springer-Verlag, 2010, pp. 7–52.

[19] I. Bichindaritz, S. Montani, Advances in case-based reasoning in the health sciences, Artificial Intelligence in Medicine. Special issue on Advances in Case-Based Reasoning in the Health Sciences 51 (2011) 75–79.

[20] A. Aamodt, E. Plaza, Case-based reasoning: Foundational issues, methodological variations, and system approaches., AI Communications 7 (1994) 39–59.

[21] S. Mitra, T. Acharya, Data Mining. Multimedia, Soft Computing, and Bioinformatics, Wiley, 2003.

[22] A. Depeursinge, B. Fischer, H. Mller, T. M. Deserno, Prototypes for content-based image retrieval in clinical practice, The Open Medical Informatics Journal 5 (2011) 58–72.

[23] B. Zheng, Computer-aided diagnosis in mammography using content-based image retrieval approaches: Current status and future perspectives, Algorithms 2 (2009) 828–849.

[24] H. Jing, Y. Yang, Image retrieval for computer-aided diagnosis of breast cancer, in: Image Analysis Interpretation (SSIAI), 2010 IEEE Southwest Symposium on, 2010, pp. 9 –12. doi:`10.1109/SSIAI.2010.5483930`.

[25] L. Wei, Y. Yang, R. M. Nishikawa, Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis, Pattern Recognition 42 (2009) 1126–1132.

[26] E. Armengol, Classification of melanomas *in situ* using knowledge discovery with explained case-based reasoning, Artificial Intelligence in Medicine. Special issue on Advances in Case-Based Reasoning in the Health Sciences 51 (2011) 93–105.

[27] M. U. Ahmed, S. Begum, P. Funk, N. Xiong, B. von Scheele, A multi-module case-based biofeedback system for stress treatment, Artificial Intelligence in Medicine. Special issue on Advances in Case-Based Reasoning in the Health Sciences 51 (2011) 107–115.

[28] Y.-J. Park, S.-H. Chun, B.-C. Kim, Cost-sensitive case-based reasoning using a genetic algorithm: Application to medical diagnosis, Artificial Intelligence in Medicine. Special issue on Advances in Case-Based Reasoning in the Health Sciences 51 (2011) 133–145.

[29] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, A. Y. Wu, An optimal algorithm for approximate nearest neighbor searching fixed dimensions, Journal of the ACM 45 (1998) 891–923.

[30] A. Gionis, P. Indyk, R. Motwani, Similarity search in high dimensions via hashing, in: International Conference on Very Large Data Bases, 1999, pp. 518–529.

[31] M. Datar, N. Immorlica, P. Indyk, V. Mirrokni, Locality-sensitive hashing scheme based on p-stable distributions, in: Symposium on Computational Geometry, ACM Press, 2004, pp. 253–262.

[32] A. Andoni, P. Indyk, Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions, in: IEEE Symposium on Foundations of Computer Science, volume 51, 2006, pp. 459–468.

[33] E. A. Alanís-Reyes, J. L. Hernández-Cruz, J. S. Cepeda, C. Castro, H. Terashima-Marín, S. E. Conant-Pablos, Analysis of machine learning techniques applied to the classification of masses and microcalcification clusters in breast cancer computer-aided detection, Journal of Cancer Therapy 3 (2012) 1020–1028.

[34] J. Suckling, J. Parker, D. Dance, The mammographic image analysis society digital mammogram database, Exerpta Medica. International Congress Series 1069 (1994) 375–378.

[35] T. Hastie, R. Tibshirani, A. Buja, Flexible discriminant analysis by optimal scoring, Journal of the American Statistical Association 89 (1994) 1255–1270.

[36] S. Haykin, Neural Networks: A comprehensive Foundation, second ed., Macmillan College Publishing Co., New York, 1999.

[37] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, New York, 1998.

[38] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Springer Series in Statistics, second ed., Springer, 2008.

[39] K. Jong, Learning with genetic algorithms: An overview, Machine Learning 3 (1988) 121–138.

[40] V. Kurkova, Kolmogorov's theorem, MIT Press, Cambridge, Massachusetts, 1995.

[41] M. Hollander, D. A. Wolfe, Nonparametric Statistical Methods, John Wiley & Sons, Inc, Hoboken, NJ, 1999.

[42] D. Sheskin, Handbook of parametric and nonparametric statistical procedures, Boca Raton: Chapman & Hall/CRC, 2004.