BREAST CANCER RISK ASSESSMENT THROUGH CONTENT-BASED MAMMOGRAPHIC IMAGE RETRIEVAL

ASTRID CAROLINA PADILLA ARRIETA

UNIVERSIDAD INDUSTRIAL DE SANTANDER FACULTAD DE INGENIERÍA
FISICOMECÁNICAS
ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y DE
TELECOMUNICACIONES
BUCARAMANGA
2020

BREAST CANCER RISK ASSESSMENT THROUGH CONTENT-BASED MAMMOGRAPHIC IMAGE RETRIEVAL

ASTRID CAROLINA PADILLA ARRIETA

Trabajo de grado presentado como requisito parcial para optar al título de ingeniera electrónica

Director:
SAID DAVID PERTUZ ARROYO
PhD

UNIVERSIDAD INDUSTRIAL DE SANTANDER FACULTAD DE INGENIERÍA FISICOMECÁNICAS
ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
BUCARAMANGA
2020

AGRADECIMIENTOS

Agradezco primeramente a Dios por darme todo lo que he necesitado en la vida. A mis padres por su apoyo incondicional en cada una de las metas que me he propuesto a alcanzar. A mi familia, amigos y a cada una de las personas que me han apoyado y animado en el camino.

CONTENTS

pag.

INTRODUCTION 11
1. CONTENT-BASED MAMMOGRAPHIC IMAGE RETRIEVAL FRAMEWORK ... 13
1.1 FEATURE EXTRACTION 14
1.2 FEATURE SELECTION 15
1.3 SIMILARITY MEASURE ………………………………………………………...17

2. BAG OF WORDS RETRIEVAL SYSTEM FRAMEWORK 19

3. BREAST CANCER RISK ASSESSMENT THROUGH IMAGE RETRIEVAL 21
3.1 HIERARCHICAL REGRESSION.. 21
3.2 CONSENSUS SCORING. 22

4. PERFORMANCE MEASURES 24

5. EXPERIMENTS AND RESULTS 26

6. CONCLUSIONS.. 31

BIBLIOGRAPHY 32

LIST OF TABLES

pag.

Table 1. Distance Metrics 17

Table 2 Features 26

Table 3 Hierarchical regression. Performance in terms of area under the ROC curve (AUC) with 95% confidence interval in parenthesis. 26

Table 4 Consensus scoring. Performance in terms of area under the ROC curve (AUC) with 95% confidence interval in parenthesis. 26

LIST OF FIGURES

pag.

Figure 1 Content Based Mammographic Image Retrieval general scheme…….. 13

Figure 2 Bag of Words Content Based Mammographic Image Retrieval general scheme. 18

Figure 3 Hierarchical regression scheme. $X_{tst} \in R^P$, P is the number of final features after feature selection. It 2 and IQ are the test set of images and the query image notation respectively used for BoW CBIR implementation. 21

Figure 4 Consensus scoring. $X_{tri} \in R^P$, $X_{tst} \in$ RP, where P is the number of final features after feature selection. I_{t1}, I_{t2} and I_Q are the notation for BoW CBIR implementation train set, test set and query image, respectively.. 22

Figure 5 Comparison of CBIR systems performances in (a) TAYS I and (b) TAYS I GE datasets. CBIR: Content-Based Image Retrieval with complete set of features. CBIR + FS: Content-Based Image Retrieval with feature selection. CBIR-BoW: Content-Based Image Retrieval with Bag of Words implementation. 27

RESUMEN

TÍTULO: EVALUACIÓN DE RIESGO DE CÁNCER DE SENO A TRAVÉS DE SISTEMAS DE IMAGE RETRIEVAL QUE ANALIZAN IMÁGENES MAMOGRÁFICAS A PARTIR DE SU CONTENIDO *

AUTOR: ASTRID CAROLINA PADILLA ARRIETA **

PALABRAS CLAVE: EVALUACION DE RIESGO, CANCER DE SENO, MAMOGRAFÍA.

DESCRIPCIÓN: En este trabajo, un sistema básico de image retrieval el cual analiza imágenes a partir de su contenido y permite obtener una cantidad de imágenes similares en cuanto a su contenido con respecto a una imagen de entrada o imagen patrón, es adaptado para su uso con imágenes mamográficas. Para la tarea del image retrieval se consideran dos enfoques, el primero de estos es un método que tiene en cuenta las características de textura extraídas de las imágenes y las distancias entre dichos vectores de características; el segundo, es un método basado en el esquema de bolsa de palabras. Las imágenes más similares son usadas para llevar a cabo la tarea de evaluación de riesgo. Se compara la estimación del riesgo usando la técnica de image retrieval en contraste con el análisis clásico del tejido parenquimatoso del seno para la evaluación de riesgo. La implementación del image retrieval basado en las características de textura no mostró mejoras en la tarea de la evaluación del riesgo, mientras que la implementación basada en el esquema de bolsa de palabras mostró una mejora significativa en la precisión de la predicción.

---

*Trabajo de grado

** Facultad de Ingeniería Fisicomecánicas. Escuela de Ingeniería Eléctrica, Electrónica y de Telecomunicaciones. Director PhD. Said David Pertuz Arroyo.

ABSTRACT

TITLE: BREAST CANCER RISK ASSESSMENT THROUGH CONTENT -BASED MAMMOGRAPHIC IMAGE RETRIEVAL [*]

AUTHOR: ASTRID CAROLINA PADILLA ARRIETA [**]

KEY WORDS: MAMMOGRAPHY, BREAST CANCER, RISK ASSESSMENT, CON- TENT -BASED, IMAGE RETRIEVAL.

DESCRIPTION: In this work, a basic content-based image retrieval system which analyzes images from their content and allows a similar amount of images to be obtained with respect to an input image or query image, is adapted for its use with mammographic images, and finally use the resultant images in the task of breast cancer risk assessment. Two approaches are considered for the retrieval image task. The first is a method that takes into account the texture features extracted from the images and the distances between these features vectors and the query features vector to measure similarity; the second is a scheme based on Bag of Words method, the similar images obtained with any of both methods are used then to perform the risk assessment task. We compare risk estimation using the retrieval image technique in contrast to the classic analysis of the parenchymal breast tissue for risk assessment. The implementation of image retrieval based on texture characteristics showed no improvement in the risk assessment task, while the implementation based on the bag of words scheme showed a significant improvement in predictive accuracy.

[*] Bachelor thesis

[**] Faculty of Physical-Mechanical Engineering. School of Electrical, Electronic and Telecommunications Engineering. Director PhD Said David Pertuz Arroyo

INTRODUCTION

Breast is the most common cancer in women worldwide and causes the greatest number of cancer-related deaths. Early detection and prevention can help in reducing mortality rates.[5] There are many documented risk factors for the development of breast cancer, such as family history of breast cancer in relatives, hormonal and reproductive risk factors and genetic risk factors, among others[6]. In the same way, image analysis has been established as a promising tool for the diagnosis and assessment of breast cancer.

Mammography is one of the most adopted screening tools for diagnosis and early detection[7]. With this in mind, risk assessment can be done through the analysis of mammograms, due to the fact that some breast characteristics, such as density and texture/parenchymal patterns have been demonstrated to be strong and independent risk factors for breast cancer[8] [9] [10]. The number of mammograms is increasing daily. A lot of hospitals and screening centers store them in large databases with the aim to provide retrospective reference images for research and training purposes[11]. For instance, these databases can be utilized to compare and search for images that have similarities to the current patient's image. For this purpose, there is an image processing technique called *image retrieval* that helps

---

[5]Li, Xi-Zhao; Williams, Simon; Bottema, Murk J. "Texture and region dependent breast cancer risk assessment from screening mammograms". In: *Pattern Recognition Letters* 36 (2014), pp. 117–124. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2013.10.001. URL: http://www.sciencedirect. com/science/article/pii/S0167865513003747.

[6]Howell, D Gareth R Evansand Anthony. "Breast cancer risk-assessment models". In: *Breast Cancer Research* 9.05 (2007), 1–8*.

[7]Singh, V. P. et al. "An efficient content based image retrieval for normal and abnormal mammograms". In: *2015 IEEE UP Section Conference on Electrical Computer and Electronics (UPCON)*. 2015, pp. 1–6.

[8]Li, Jingmei et al. "High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer". In: *Breast Cancer Research* 14.4* (July 2012). DOI: 10.1186/bcr3238.

[9]Aimilia Gastounioti, Emily F. Conant; Kontos, Despina. "Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment". In: *Breast Cancer Research* 18*.1* (Sept. 2016). DOI: 10.1186/s13058-016-0755-8.

[10]Singh et al., "An efficient content based image retrieval for normal and abnormal mammograms", op. cit.

[11]Ibid.

radiologists to analyze the content of the current image and compare with past cases[12] for supporting their decision making regarding various medical procedures[13]. The most common practices for content-based image retrieval (CBIR) in mammograms are lesion retrieval and density pattern categorization[14] [15] [16]. Most of these systems are used as a pre stage for CAD systems, due to the fact that CBIR systems for themselves do not provide any diagnostic information.

In this work, we propose and implement two methods: a basic scheme for a Content Based Mammographic Image Retrieval (CBIR) based on texture features, and a Bag of Visual Words CBIR which is based in Speeded Up Robust Features (SURF), where both allow to obtain a certain number of similar mammography images to a query one. In view of that different previous works have shown satisfactory results returning relevant images to an input image[17][18], our purpose is to evaluate whether image retrieval has potential as a pre-processing step for parenchymal analysis in breast cancer risk assessment.

---

[12] Ibid.

[13] Liu, J. et al. “Design and implementation of content-based medical image retrieval system on mammograms”. In: *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*. vol. 1. 2011, pp. 237–240.

[14] Singh et al., “An efficient content based image retrieval for normal and abnormal mammograms”, op. cit.

[15] Liu et al., “Design and implementation of content-based medical image retrieval system on mam- mograms”, op. cit.

[16] Jose, S.; Chandy, D. A. “Content based mammogram retrieval using biorthogonal wavelet filters in DDSM database”. In: *2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)*. 2014, pp. 1–6.

[17] Shamna, P.; Govindan, V.K.; Nazeer], K.A. [Abdul. “Content-based medical image retrieval by spatial matching of visual words”. In: *Journal of King Saud University - Computer and Information Sci- ences* (2018). ISSN: 1319-1578. DOI: 10.1016/j.jksuci.2018.10.002. URL: http://www.sciencedirect. com/science/article/pii/S131915781830750X.

[18] Liu, J. et al. “Scalable Mammogram Retrieval Using Composite Anchor Graph Hashing With Iterative Quantization”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 27.11 (2017), pp. 2450–2460.

1. CONTENT-BASED MAMMOGRAPHIC IMAGE RETRIEVAL FRAMEWORK

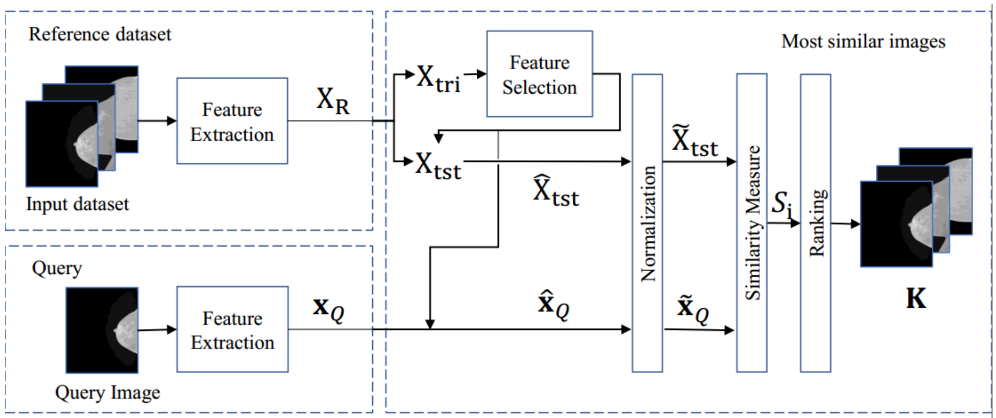


Figure 1: Content Based Mammographic Image Retrieval general scheme

The basic scheme implemented in this work is illustrated in figure 1. First of all, pre- processing and feature extraction step are applied to each image available in the dataset yielding a feature vector set $X_R = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, x_N\}$, where $\mathbf{x}_i \in \mathrm{R}^M$ is the feature vector of the ith reference image, N is the total number of the reference images and M is the number of features for each image.

The same process is performed on the query image, yielding $\mathbf{x}_Q \in \mathrm{R}^M$ as its feature vector. Subsequently, *feature selection* is performed to get rid of redundant and ir- relevant features which can affect the performance of the retrieval system.[1] For this task, the feature vector set X_R is split into train and test sets, X_{tri} and X_{tst} respec- tively. Feature selection is performed with the training data X_{tri}, and the final set of features are used into X_{tst} and $\mathbf{x}_Q$, resulting in a final test set $\hat{X}_{tst} \in \mathrm{R}^P$ an a query feature vector $\hat{\mathbf{x}}_Q \in \mathrm{R}^P$, where P is the number of features after feature selection, being $P < M$.

[1]Silva, Anthony Mihirana De; Leong, Philip H.W. *Grammar-based feature generation for time-series prediction*. Springer Singapore Heidelberg New York Dordrecht London, 2015. DOI: https//doi.org/10.1007/978-981-287-411-5.

In the next step, z-score normalization is applied into $\hat{X}_{tst}$ and $\hat{\mathbf{x}}_Q$ turning out in $\tilde{X}_{tst}$ and $\tilde{\mathbf{x}}_Q$ respectively. The objective of normalization is to give all features the same weight which helps to prevent that if there are features with initially large ranges, they outweigh other features with initially smaller ranges[2]. At the end, features will have a mean equal to 0 and a standard deviation of 1[3].

Finally, a similarity between the query feature vector $\mathbf{x}_Q$ and each feature vector of the test set $\mathbf{x}_i \in \tilde{X}_{test}$ as $S_i = s(\mathbf{x}_Q, \mathbf{x}_i \in \hat{X}_{tst})$ where $s(\cdot, \cdot)$ is a similarity measure. Images are then ranked in ascending order by their similarity level with the query image. The number of k retrieved images is defined experimentally. Algorithm 1 shows a summary of the CBIR for mammography images. A detailed description of each stage is presented below.

Algorithm 1: CBIR

Input: $\mathbf{x_Q}$, X_{tri}, X_{tst} k
Output: $\mathbf{K}$ → Feature vector set of the k most similar images.

1: $\hat{X}_{tst} \leftarrow$ featSelect(X_{tri})
2: $\hat{\mathbf{x}}_Q \leftarrow$ featSelect(X_{tri})
3: $\tilde{\mathbf{x}}_Q = zscore(\hat{\mathbf{x}}_Q)$
4: $\tilde{X}_{tst} = zscore(\hat{X}_{tst})$
5: $S_i = s(\tilde{\mathbf{x}}_Q, \mathbf{x}_i \in \tilde{X}_{tst})$
6: $\mathbf{K} = arg\ max\ S_i$

1.1 FEATURE EXTRACTION

One of the main components of CBIR systems is feature extraction, which assumes that the content of an image can be described by some quantitative measures that encode its attributes[4] . In the context of mammography images,

---

[2]Han, Jiawei; Kamber, Micheline; Pei, Jian. *Data mining: Data mining concepts and techniques*. 3rd. 225 Wyman Street, Waltham, MA 02451, USA: Morgan Kaufmann Publishers, 2012.

[3]Salvador García, Julián Luengo; Herrera, Francisco. *Data Preprocessing in Data Mining*. Vol. 72*. 978-3-319-10246-7. Dirección del editor: Springer Cham Heidelberg New York Dordrecht London, 2015. DOI: 10.1007/978-3-319-10247-4.

[4]Feng, David Dagan. “Biomedical Information Technology”. In: 1st. Elsevier/Academic Press, 2008. Chap. 4, pp. 83–113. DOI: 10.1016/C2017-0-03305-0.

texture and shape features are the most used, since these are the ones that better describe important elements such as fibro glandular tissue and mass contour attributes which can be a discriminative factor to represent different lesion characteristics[5][6]. In addition, these features have shown certain association with breast cancer risk[7], which is the problem of interest addressed in this work.

For visual representation of mammography images, feature extraction was performed by the feature extraction module from OpenBreast. OpenBreast is a public, mammographic image analysis framework developed at Universidad Industrial de Santander for the task of breast cancer risk assessment[8]. This module extracts a total of 33 quantitative texture measures, which are dived into five main groups: statistical features (STA), gray-level co-occurrence features (GLC), gray-level run-length features (GLR), gradient-based features (GRA) and spatial-frequency analysis (SFA).[9]

1.2 FEATURE SELECTION

Because texture features comprehend an important element in the retrieval process, most of the studied feature selection methods that are applied to

[5]Tsochatzidis, Lazaros et al. "Computer-aided diagnosis of mammographic masses based on a supervised content-based image retrieval approach". In: *Pattern Recognition* 71 (2017), pp. 106–117. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2017.05.023.

[6]Wei, Chia-Hung; Chen, Sherry Y.; Liu, Xiaohui. "Mammogram retrieval on similar mass lesions". In: *Computer Methods and Programs in Biomedicine* 106.3 (2012), pp. 234–248. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2010.09.002.

[7]Pertuz, Said et al. "Clinical evaluation of a fully-automated parenchymal analysis software for breast cancer risk assessment: A pilot study in a Finnish sample". In: *European Journal of Radiology* 121 (2019), p. 108710. ISSN: 0720-048X. DOI: 10.1016/j.ejrad.2019.108710. URL: http://www.
sciencedirect.com/science/article/pii/S0720048X19303602.

[8]Pertuz, S. et al. "Open Framework for Mammography-based Breast Cancer Risk Assessment". In: *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*. 2019, pp. 1– -4.

[9]Ibid.

classification problems can be, in the same way, applied to the CBIR system[10 11], with the aim to improve its performance retrieving the most similar images.

For this reason, we consider two feature selection methods: a filter and a wrapper method[12 13]. In filter method, a correlation analysis was performed to identify those features with a high correlation between them, altogether with the respective *p* values, in the same way a statistical analysis was executed to assess how these features are related and to evaluate if there is an important difference in their behavior between cases and controls groups. For this purpose, Mann-Whitney U test was used to assess these differences. The final features of this filter method did not yield good results in image retrieval tasks. For this reason, only wrapper-based methods were considered in the remainder of this work. After preliminary experiments, we selected a wrapper method, Neighborhood Component Analysis (NCA) feature selection, which is applied as a pre-processing step for the CBIR system implementation.

In NCA feature selection, a reference point is selected as the nearest neighbor of a new point, where the probability of a point of being selected is higher if it is closer to the new point, proximity that is defined by a distance function with a weight factor given by the feature weights (**w**). Then, the algorithm considers the average leave- one out probability of correct classification $F\ (w)$ with a regularization term λ in order to maximize $F\ (w)$ with respect to **w**.

[10]da Silva, S. F. et al. "Ranking evaluation functions to improve genetic feature selection in content- based image retrieval of mammograms". In: *2009 22nd IEEE International Symposium on Computer- Based Medical Systems*. 2009, pp. 1–8.

[11]Bugatti, P. H. et al. "Content-Based Retrieval of Medical Images by Continuous Feature Se- lection". In: *2008 21st IEEE International Symposium on Computer-Based Medical Systems*. 2008, pp. 272–277.

12Liu, Huan. "Evolving Feature Selection". In: Trends and Controversies, IEEE Computer Society (Nov. 2005), pp. 64–76.

[13]Guyon, Isabelle; Elisseeff, Andre. "An Introduction to Variable and Feature Selection". In: *Journal of Machine Learning Research* 3 (Apr. 2003), pp. 1157–1182. DOI: 10.1162/153244303322753616.

Tuning λ through cross validation, leads that many of the weights in **w** are driven to 0. At the end, the final weight vector is the one that minimize the classification error[14].

1.3 SIMILARITY MEASURE

Each image in the database is represented by a feature vector obtained in feature extraction step. The same feature extraction is performed in the query image. The similarity between two images is usually found as the distance between their feature vectors[15]. The similarity measure has an important role in image retrieval because of its direct relation with the performance of the system[16]. This measure is used to sort the reference images in ascending order according to their similarity with the query image. The similarity measures considered in this work were: Euclidean, Canberra, Chebyshev, Cosine, Dice, Lorentzian, Manhattan and Sorensen, see table 1; with the aim to evaluate the one that present the best performance in the retrieval task. The number of retrieved images is chosen empirically.

---

[14]Wei Yang, Kuanquan Wang; Zuo, Wangmeng. "Neighborhood component feature selection for high-dimensional data". In: *Journal of Computers* 7.1* (Jan. 2012), 162–168*. DOI: 10.4304/jcp.7.1. 161-168.

[15]Wei; Chen; Liu, "Mammogram retrieval on similar mass lesions", op. cit.

[16]Meharban, M.S.; Priya, Dr.S. "A Review on Image Retrieval Techniques". In: *Bonfring Interna- tional Journal of Advances in Image Processing* 6*.2* (Apr. 2016), 07–10*. DOI: 10.9756/bijaip.8136.

Table 1: Distance Metrics

Distance Name	Definition
Euclidean	$\sqrt{\sum_{i=1}^{n} \lvert x_i - y_i \rvert^2}$
Canberra	$\sum_{i=1}^{n} \frac{\lvert x_i - y_i \rvert}{\lvert x_i \rvert - \lvert y_i \rvert}$
Chebyshev	$max_i \lvert x_i - y_i \rvert$
Cosine	$1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} x_i^2}}$
Dice	$1 - \frac{2 \sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} x_i^2}$
Lorentzian	$\sum_{i=1}^{n} ln(1 + \lvert x_i - y_i \rvert)$
Manhattan	$\sum_{i=1}^{n} \lvert x_i - y_i \rvert$
Sorensen	$\frac{\sum_{i=1}^{n} \lvert x_i - y_i \rvert}{\sum_{i=1}^{n} (x_i + y_i)}$

2. BAG OF WORDS RETRIEVAL SYSTEM FRAMEWORK

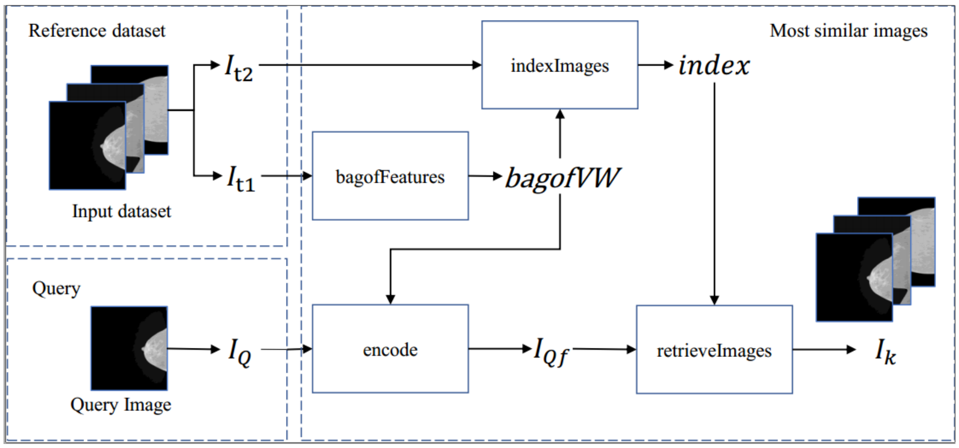


Figure 2: Bag of Words Content Based Mammographic Image Retrieval general scheme.

The proposed CBIR uses low-level features to describe the images in the database. For comparison, in this work we also utilized a CBIR pipeline based on Bag of Visual Words (BoW), as shown in figure 2, with the aim to compare the performance of the proposed system. First of all, given a train set of images I_{t1}, which contain representative images from both classes, a bag of visual words *bagofVW* is trained to create a group of image descriptors which represent the data set of images. These features are referred to as visual words, which are defined by using k-means clustering in the features descriptors extracted from I_{t1}, where each visual word is represented by a cluster center. In this work it was defined a size of 5000 visual words. Feature descriptors are extracted using an upright SURF detector into a pre-defined uniform grid of 15×15 pixels, where intersections of the grid are the locations for feature extraction.

Subsequently these descriptors or visual words are mapped according to their occurrences in the test set of images I_{t2}, creating a search index *index*, which allows to speed up the retrieval task. In the same way, given a query image I_Q, its occurrences are mapped with the visual words such I_{Qf}, to afterwards compare it into the created *index* and provide the most similar images I_k[1]. This process is resumed in Algorithm 2.

The training set for the bag of features, was the same used for the training of the NCA feature selection, then the indexing process is performed with the test set to subsequently search for the most similar images to a query within the test set. Every image in the test set is used as well, as in the aforementioned system, as query image.

Algorithm 2: BoW CBIR

Input: I_{t1} , I_{t2}, I_Q, k
Output: $\boldsymbol{I_k}$

1: $bagofVW$ = bagOfFeatures(I_{t1})
2: $index$ = indexImages(bagofVW,I_{t2})
3: I_{Qf} = encode(bagofVW,I_Q)
4: imageIDs = retrieveImages(I_{Qf} , index)
5: I_k = imageIDs(1:K)

---

[1] The MathWorks, Inc ®. *Image Retrieval with bag of visual words*. Access on 04 may 2020. 2018. URL: https://es.mathworks.com/help/vision/ug/image-retrieval-with-bag-of-visual-words.html.

3. BREAST CANCER RISK ASSESSMENT THORUGH IMAGE RETRIEVAL

Most of the reviewed approaches for breast cancer risk assessment through mammography analysis, are generally based on fibroglandular tissue and density assessment of the breast [1][2][3]. In the literature, we did not find any previous work utilizing CBIR for the task of breast cancer risk assessment. The novelty of this work is the design and implementation of a scheme that uses the retrieved images from the previous CBIR system implemented to assess the risk of suffering breast cancer in the query input mammography.

Two different schemes were implemented, in order to assess which one present the best performance within the context of this work: hierarchical regression and consensus scoring. In this way, each implementation is compared with the the classical parenchymal analysis for breast cancer risk assessment[4].

3.1 HIERARCHICAL REGRESSION

The output of any of both CBIR systems introduced previously, is the set of the K most similar images $\mathbf{K}_{mj}$ of a query $\mathbf{x}_Q$. In this case, as shown in figure 3, the main idea is to choose the set of the K most similar images $\mathbf{K}_{mj}$, and the set of the K least similar images $\mathbf{K}_{mn}$ to conform a train set $\mathbf{x}_{train}$, to train a logistic regression model defined as $model$ to assess the risk in the input image $\mathbf{x}_Q$ as the test set, giving it

[1] Li et al., "High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer", op. cit.

[2] Tan, Maxine et al. "Assessment of a Four-View Mammographic Image Feature Based Fusion Model to Predict Near-Term Breast Cancer Risk". In: *Annals of Biomedical Engineering* 43 (2015), pp. 2416–2428. DOI: 10.1007/s10439-015-1316-5.

[3] Tan, Maxine et al. "Assessment of a Four-View Mammographic Image Feature Based Fusion Model to Predict Near-Term Breast Cancer Risk". In: *Annals of Biomedical Engineering* 43 (2015), pp. 2416–2428. DOI: 10.1007/s10439-015-1316-5.

[4] Cheddad, Abbas et al. "Area and volumetric density estimation in processed full-field digital mammograms for risk assessment of breast cancer". In: *PLoS One* 9.10 (Oct. 2014). DOI: 10.1371/journal. pone.0110690.

a risk scored S_Q. At the end, this approach tries to reduce the set of images used for training the risk model compared with the parenchymal analysis which uses all the images in the database for this task[55]. For validation purposes, the procedure is repeated for every image in the database, using leave-one-out cross validation. This process is resumed into Algorithm 3.

Algorithm 3: Hierarchical regression

Input: $\mathbf{x}_Q$,$\mathbf{K}_{mj}$, $\mathbf{K}_{mn}$
Output: S_Q

1: $\mathbf{x}_{\mathbf{train}} = [\mathbf{K}_{mj}; \mathbf{K}_{mn}]$
2: model = logReg($\mathbf{x}_{\mathbf{train}}$)
3: S_Q = predict(model, $\mathbf{x}_Q$)
return S_Q

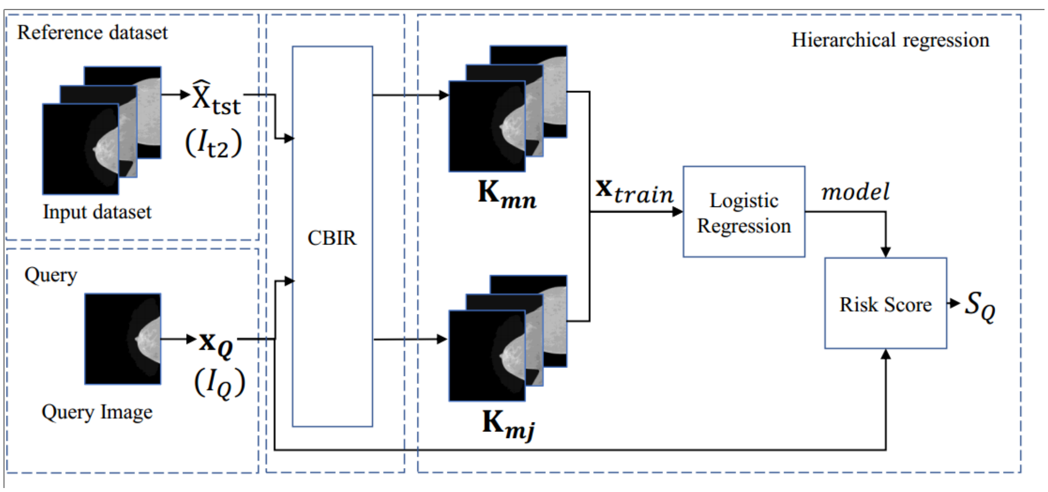


Figure 3: Hierarchical regression scheme. $\hat{X}_{tst} \in \mathrm{R}^P$, P is the number of final features after feature selection. I_{t2} and I_Q are the test set of images and the query image notation respectively used for BoW CBIR implementation.

3.2 CONSENSUS SCORING

In a previous step, the database is divided into train set X_{tri} and a test set X_{tst} , for the respective feature selection. In this approach as illustrated in figure 4, a

[5] Pertuz et al., “Open Framework for Mammography-based Breast Cancer Risk Assessment”, op.cit.

logistic regression model is trained with the same training set X_{tri} and the final features that showed the best performance in the retrieval process. Simultaneously, CBIR is performed into the test set X_{tst}, retrieving the set of the K images that present the most similarity $\mathbf{K}_{mj}$ with the respective query $\mathbf{x}_Q$ being tested. For each retrieved image, the trained model is employed to give a risk score $\mathbf{S}_k$. Keeping in mind that the main objective of the retrieval system is to return the most similar images, in the best scenario, most of the images would belong to the true class of the query image. The risk score for the query image is set as the mean value of the scores of the retrieved images S_Q, as seen in algorithm 4.

Algorithm 4: Consensus scoring

Input: $\mathbf{x_Q}$,$\mathbf{K}_{mj}$, $\mathbf{X_{tri}}$
Output: S_Q

1: model = logReg($\mathbf{X_{tri}}$)
2: $\mathbf{S_k}$ = predict(model, $\mathbf{K}_{mj}$)
3: $S_Q = \frac{1}{k} \sum_{i=1}^{i}$
return S_Q

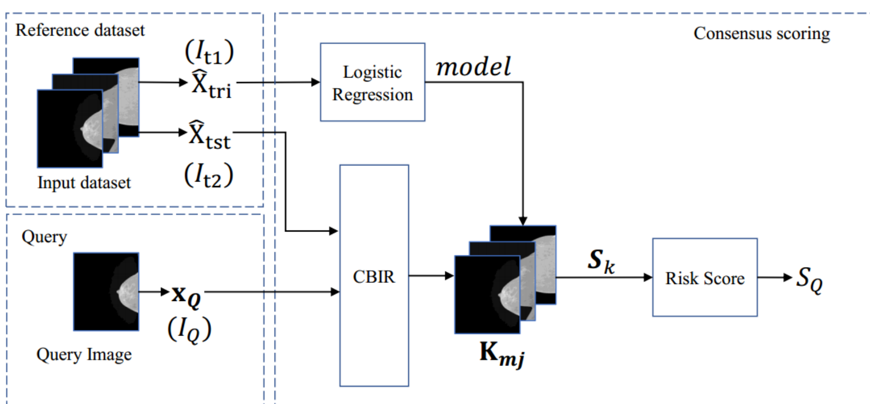


Figure 4: Consensus scoring. $\hat{X}_{tri} \in \mathrm{R}^P$, $\hat{X}_{tst} \in \mathrm{R}^P$, where P is the number of final features after feature selection. I_{t1}, I_{t2} and I_Q are the notation for BoW CBIR implementation train set, test set and query image, respectively.

4. PERFORMANCE MEAURES

Performance is measured in this work for two main aspects: the first one is for the retrieval image system precision which is measured with Precision and Recall curves; and the second one is for the performance in risk assessment, which is measured by the area under the receiver operating characteristic curve (AUC).

In order to assess the performance of image retrieval, we define a *relevant image* as the one with the same class as the query image, contrarily to an *irrelevant* result which has a different class. The retrieval system returns a ranked list, where the K most similar images are in the beginning of the list. Precision (P) is the proportion of the K images that are relevant [1]. It can be also understood as the probability that an image in the retrieved set of images is relevant[2]. Likewise, recall (R) measures the ability of the retrieval system to locate relevant material in its index[3]. Other definition is that it is the probability that an image that has been retrieved to be relevant.

- n : number of relevant retrieved images
- N_R : number of retrieved images
- N_r : number of relevant images in database

$$P = \frac{n}{N_R} \tag{4.1}$$

$$R = \frac{n}{N_r} \tag{4.2}$$

[1] Ling Liu, M. Tamer Özsu. *Encyclopedia of Database Systems*. 1st. 233 Spring Street, New York, NY 10013, U.S.A: Springer Science+Business Media, LLC 2009 (USA), 2009.

[2] Losee, Robert M. *Text Retrieval and Filtering: analytic models of performance*. 1st*. Springer Science+Business Media Dordrecht, 1998. DOI: 10.1007/978-1-4615-5705-0.

[3] Ling Liu, *Encyclopedia of Database Systems*, op. cit.

The precision and recall values reported here in figure 5a and 5b, correspond to the average of the values obtained, due to the retrieval process is repeated using each image on the dataset as query image.

To measure the performance of the risk prediction accuracy of both implementations proposed in section 3, it was used the AUC. Each image in the test set is used as query, and thus the schemes gives a risk score to the query; these scores are stored and finally used to evaluate the performance.

Parenchymal analysis for breast cancer risk assessment was implemented in order to compare the proposed schemes performance for risk assessment through CBIR. With this in mind, it was employed the same split data used before. A logistic regression model with stepwise feature selection was trained into the training set and perform the risk assessment into the test set, then AUC values were computed. Differences in AUCs were assessed using DeLong's test.[4]

---

[4] Elizabeth R. DeLong, David M. DeLong; Clarke, Daniel L. "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach". In: *Biometrics* 44*.3* (Sept. 1998), 837–845*. DOI: 10.2307/2531595. URL: https://www.jstor.org/stable/2531595.

5. EXPERIMENTS AND RESULTS

Experiments of each introduced system in this work, are performed into two databases:

- TAYS I : this is a database of mammograms divided into two groups: cases, with 114 mammography images (contralateral CC view) of patients with screening- detected asymptomatic cancers; and controls, with 114 (right breast CC view) matched consecutive healthy women. In this database, the mammography images were acquired with either MicroDose SI (Philips Healthcare, the Netherlands) or a Senographe Essential (General Electric Medical Systems, USA) systems.[11]

- TAYS I GE: According to previous works [2] when images are from different mammographic systems, it can affect the performance of computerized image analysis. For this reason, a subset of mammography images were selected from the database described above, where only images acquired with General Electric Medical System are taken into account. There are a total of 118 images in this dataset, 59 cases and 59 controls.

Image retrieval with different similarity measures, showed a similar behavior for all the measures used in this work, however, the similarity measure which showed the best performance was Chevychev distance for TAYS I and Sorensen distance for TAYS I GE. For comparison purposes in risk assessment, we report the results with the Chevychev distance measure.

Man Whitney U-test was performed to evaluate the features differences between cases and controls groups in both databases. It was found that some features present differences in their medians values among both groups with a significance level $p < 0.05$, the features that show these difference are listed in table 2.

---

[1] Pertuz et al., "Clinical evaluation of a fully-automated parenchymal analysis software for breast cancer risk assessment: A pilot study in a Finnish sample", op. cit.

[2] Pertuz, S. et al. "Do Mammographic Systems Affect the Performance of Computerized Parenchymal Analysis?" In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2019, pp. 4863–4866.

Table 2: Features

Dataset	Features with different medians	Final set of features
TAYS I	imin, iske, iran, rrln, rlgr, rhgr, stev, fdim	imin, iske, iran, rhgr,fdim
TAYS I - GE	imin,iba2, iske,ikur, iran,cene,ccon chom, cent,rsre, rlre, rrpe,rrln,rhgr sgra, stev, fdim	stev, fdim

For NCA feature selection, the database is split into train ($\mathbf{X}_{tri}$) and test set ($\mathbf{X}_{tst}$). In order to find parameter λ of NCA feature selection, we perform 5-fold cross validation on the training set. The optimal λ value is then utilized on the test set for validation. The output of the model applied to a set, is a weight feature vector, where it can be identified the irrelevant features as the ones with a factor nearly close to zero, and thus, considering useful those whose weight factor exceed a threshold value[3].

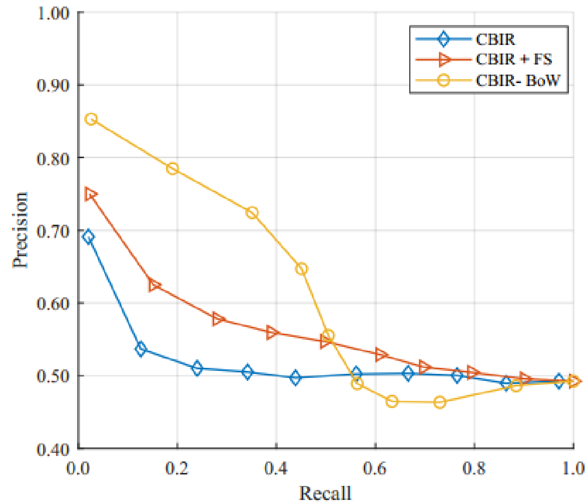
Table 3: Hierarchical regression. Performance in terms of area under the ROC curve (AUC) with 95% confidence interval in parenthesis.

Database	K	CBIR	BoW	Parenchymal analysis
TAYS I	5	0.862 (0.772-0.951)	1.000 (0.980 - 1.000)	0.862 (0.772 - 0.951)
TAYS I GE	15	0.782 (0.625 -0.939)	0.941 (0.857 - 1.000)	0.761 (0.598 - 0.924)

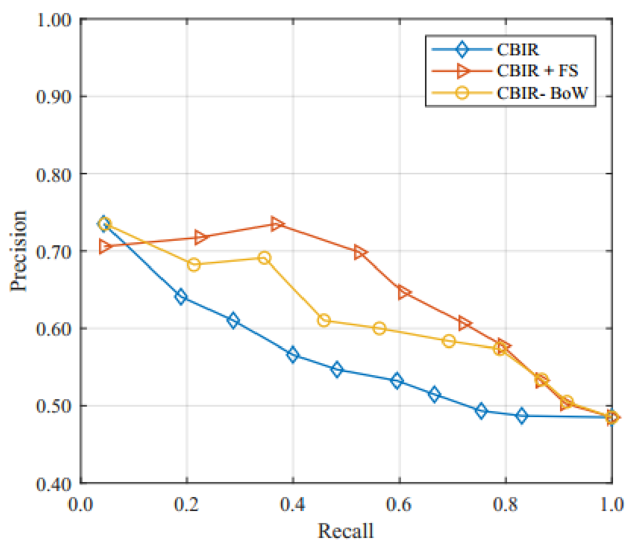
Table 4: Consensus scoring. Performance in terms of area under the ROC curve (AUC) with 95% confidence interval in parenthesis.

Database	K	CBIR	BoW	Parenchymal analysis
TAYS I	5	0.811 (0.707 - 0.914)	0.380 (0.246 - 0.513)	0.862 (0.772 - 0.951)
TAYS I GE	15	0.875 (0.754 - 0.997)	0.882 (0.764 - 1.000)	0.761 (0.598 - 0.924)

[3]Wei Yang; Zuo, "Neighborhood component feature selection for high-dimensional data", op. cit.



(a)



(b)

Figure 5: Comparison of CBIR systems performances in (a) TAYS I and (b) TAYS I GE datasets. CBIR: Content-Based Image Retrieval with complete set of features. CBIR + FS: Content-Based Image Retrieval with feature selection. CBIR-BoW: Content-Based Image Retrieval with Bag of Words implementation.

Reducing the set of features by feature selection, resulted in an improvement on the retrieval system performance with respect to the results obtained when CBIR basic system using the full feature set, as shown in figures 5a and 5b. Different threshold values for the feature's weights were chosen and tested in the NCA feature selection module; final sets of features which reflected the best performance in each database are the ones used for comparison in risk assessment. These features are listed in table 2.

Among the two databases employed in this work, the best performance in terms of retrieving the most relevant images, is obtained in the second database (TAYS I GE). After feature selection, the precision improved in a more wide range of recall values (K images retrieved) than in the other database, where precision decays rapidly as recall value increase.

Additionally, there was deployed a retrieval system with the scheme of BoW into the same data used to evaluate the CBIR based on parenchymal features. This implementation shows, in terms of retrieval performance, higher precision values than the basic CBIR when less than the half of relevant images are retrieved, in the case of TAYS I database as shown in figure 5a. Otherwise, with TAYS I GE, the CBIR, reaches higher precision values than BoW implementation as seen in figure 5b.

To evaluate the performance of the two risk assessment schemes proposed in this work, different values of *K* (retrieved images) were tested for both of them, and the ones with the best performance in risk assessment are listed in tables 3 and 4, where the specified *K* value is shown as well.

For hierarchical regression in TAYS I dataset, with feature-based CBIR, we did not find statistically significant improvements in terms of the AUC. In dataset TAYS I GE, however, a minimum improvement is obtained as as shown at table 3 although, this was not statistically significant ($p = 0.774$).

When risk assessment is performed with the BoW CBIR system, the AUC improves compared with the classical parenchymal analysis, showing a statistical significance of $p = 0.005$ for the first database. In the database with only images from the GE system, even though the improvement in the accuracy, it is not statistically significant ($p = 0.114$).

Meanwhile, consensus scoring, where the risk score is given by the mean scores values of the K most similar images to the query, the CBIR and the BoW CBIR applied into TAYS I database did not show any improvement in the assessment task. With TAYS I GE, a better performance is reflected as can be seen in table 4, however these results are not statistically significant with p = 0.171 and p = 0.331 respectively.

6. CONCLUSIONS

Feature selection plays an important role in the retrieval system, as it allows to increase the number of relevant images that the system returns. Additionally, final set of features are found to be inside those features that present relevant differences between both study groups (cases and controls), when Mann-Whitney U test was applied.

In this work, similarity between feature vectors is measured with simple distance metrics, and due to complexity that features spaces can represent, it is an aspect to improve, to look into new similarity measures to evaluate if they can improve the retrieval system performance.

Finally, when breast cancer risk assessment is performed by means of the CBIR that uses parenchymal features, it did not show a significant improvement in the AUC, compared with the classical parenchymal analysis. In contrast, the use of the BoW CBIR in the hierarchical regression approach, shows a significant improvement in risk assessment when applied into the dataset that has images from both mamographic systems. In the consensus scoring approach, even though the CBIR results are not statistical significant, it can be seen that it exists the possibility of improving retrieval precision in a wider range of the retrieved images. Arguably, this facilitates obtaining a better performance in risk prediction, due to here risk score is given by the mean of the scores of the retrieved images.

BIBLIOGRAPHY

AIMILIA GASTOUNIOTI, Emily F. Conant; KONTOS, Despina. “Beyond breast den- sity: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment”. In: *Breast Cancer Research* 18*.1* (Sept. 2016). DOI: 10 . 1186 / s13058-016-0755-8.

BUGATTI, P. H.; RIBEIRO, M. X.; TRAINA, A. J. M.; JUNIOR, C. T. “Content-Based Retrieval of Medical Images by Continuous Feature Selection”. In: *2008 21st IEEE International Symposium on Computer-Based Medical Systems*. 2008, pp. 272–277.

CHEDDAD, Abbas; CZENE, Kamila; ERIKSSON, Mikael; LI, Jingmei; EASTON, Douglas; HALL, Per; HUMPHREYS, Keith. “Area and volumetric density estimation in processed full-field digital mammograms for risk assessment of breast cancer”. In: *PLoS One* 9.10 (Oct. 2014). DOI: 10.1371/journal.pone.0110690.

DA SILVA, S. F.; TRAINA, A. J. M.; RIBEIRO, M. X.; DO E. S. BATISTA NETO, J.; TRAINA, C. “Ranking evaluation functions to improve genetic feature selection in content-based image retrieval of mammograms”. In: *2009 22nd IEEE International Symposium on Computer-Based Medical Systems*. 2009, pp. 1–8.

ELIZABETH R. DELONG, David M. DeLong; CLARKE, Daniel L. “Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach”. In: *Biometrics* 44*.3* (Sept. 1998), 837–845*. DOI: 10 . 2307/2531595. URL: https://www.jstor.org/stable/2531595.

FENG, David Dagan. “Biomedical Information Technology”. In: 1st. Elsevier/Academic Press, 2008. Chap. 4, pp. 83–113. DOI: 10.1016/C2017-0-03305-0.

GUYON, Isabelle; ELISSEEFF, Andre. “An Introduction to Variable and Feature Se- lection”. In: *Journal of Machine Learning Research* 3 (Apr. 2003), pp. 1157–1182. DOI: 10.1162/153244303322753616.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. *Data mining: Data mining concepts and techniques*. 3rd. 225 Wyman Street, Waltham, MA 02451, USA: Morgan Kaufmann Publishers, 2012.

HOWELL, D Gareth R Evansand Anthony. “Breast cancer risk-assessment models”. In: *Breast Cancer Research* 9.05 (2007), 1–8*.

JOSE, S.; CHANDY, D. A. “Content based mammogram retrieval using biorthogo- nal wavelet filters in DDSM database”. In: *2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)*. 2014, pp. 1–6.

LI, Jingmei; SZEKELY, Laszlo; ERIKSSON, Louise; BOEL HEDDSON, Ann Sund- bom. “High-throughput mammographic-density measurement: a tool for risk predic- tion of breast cancer”. In: *Breast Cancer Research* 14.4* (July 2012). DOI: 10.1186/ bcr3238.

LI, Xi-Zhao; WILLIAMS, Simon; BOTTEMA, Murk J. “Texture and region dependent breast cancer risk assessment from screening mammograms”. In: *Pattern Recog- nition Letters* 36 (2014), pp. 117–124. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2013.10.001. URL: http://www.sciencedirect.com/science/article/pii/ S0167865513003747.

LING LIU, M. Tamer Özsu. *Encyclopedia of Database Systems*. 1st. 233 Spring Street, New York, NY 10013, U.S.A: Springer Science+Business Media, LLC 2009 (USA), 2009.

LIU, Huan. “Evolving Feature Selection”. In: *Trends and Controversies, IEEE Com- puter Society* (Nov. 2005), pp. 64–76.

LIU, J.; CHEN, J.; HUANG, Z.; LIU, X. “Design and implementation of content-based medical image retrieval system on mammograms”. In: *2011 4th International Con- ference on Biomedical Engineering and Informatics (BMEI)*. Vol. 1. 2011, pp. 237– 240.

LIU, J.; ZHANG, S.; LIU, W.; DENG, C.; ZHENG, Y.; METAXAS, D. N. “Scalable Mammogram Retrieval Using Composite Anchor Graph Hashing With Iterative Quan- tization”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 27.11 (2017), pp. 2450–2460.

LOSEE, Robert M. *Text Retrieval and Filtering: analytic models of performance*. 1st*. Springer Science+Business Media Dordrecht, 1998. DOI: 10.1007/978-1-4615-5705-0.

MEHARBAN, M.S.; PRIYA, Dr.S. “A Review on Image Retrieval Techniques”. In: *Bon- fring International Journal of Advances in Image Processing* 6*.2* (Apr. 2016), 07– 10*. DOI: 10.9756/bijaip.8136.

PERTUZ, S.; SASSI, A.; ARPONEN, O.; HOLLI-HELENIUS, K.; LÄÄPERI, A.; RINTA-KIIKKA, I. “Do Mammographic Systems Affect the Performance of Computerized Parenchymal Analysis?” In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019, pp. 4863–4866.

PERTUZ, S.; TORRES, G. F.; TAMIMI, R.; KAMARAINEN, J. “Open Framework for Mammography-based Breast Cancer Risk Assessment”. In: *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*. 2019, pp. 1–4.

PERTUZ, Said; SASSI, Antti; HOLLI-HELENIUS, Kirsi; KÄMÄRÄINEN, Joni; RINTAKIIKKA, Irina; LÄÄPERI, Anna-Leena; ARPONEN, Otso. “Clinical evaluation of a fully-automated parenchymal analysis software for breast cancer risk assessment: A pilot study in a Finnish sample”. In: *European Journal of Radiology* 121 (2019), p. 108710. ISSN: 0720-048X. DOI: 10.1016/j.ejrad.2019.108710. URL: http://www.sciencedirect.com/science/article/pii/S0720048X19303602.

SALVADOR GARCÍA, Julián Luengo; HERRERA, Francisco. *Data Preprocessing in Data Mining*. Vol. 72*. 978-3-319-10246-7. Dirección del editor: Springer Cham Hei- delberg New York Dordrecht London, 2015. DOI: 10.1007/978-3-319-10247-4.

SHAMNA, P.; GOVINDAN, V.K.; NAZEER], K.A. [Abdul. “Content-based medical image retrieval by spatial matching of visual words”. In: *Journal of King Saud Univer- sity - Computer and Information Sciences* (2018). ISSN: 1319-1578. DOI: 10.1016/ j . jksuci . 2018 . 10 . 002. URL: http : / / www . sciencedirect . com / science / article / pii / S131915781830750X.

SILVA, Anthony Mihirana De; LEONG, Philip H.W. *Grammar-based feature genera- tion for time-series prediction*. Springer Singapore Heidelberg New York Dordrecht London, 2015. DOI: https://doi.org/10.1007/978-981-287-411-5.

SINGH, V. P.; GUPTA, A.; SINGH, S.; SRIVASTAVA, R. “An efficient content based image retrieval for normal and abnormal mammograms”. In: *2015 IEEE UP Section Conference on Electrical Computer and Electronics (UPCON)*. 2015, pp. 1–6.

TAN, Maxine; PU, Jiantao; LIU, Samuel Cheng andHong; ZHENG, Bin. “Assessment of a Four-View Mammographic Image Feature Based Fusion Model to Predict Near- Term Breast Cancer Risk”. In: *Annals of Biomedical Engineering* 43 (2015), pp. 2416– 2428. DOI: 10.1007/s10439-015-1316-5.

THE MATHWORKS, Inc ®. *Image Retrieval with bag of visual words*. Access on 04 may 2020. 2018. URL: https://es.mathworks.com/help/vision/ug/image- retrieval-with-bag-of-visual-words.html.

TSOCHATZIDIS, Lazaros; ZAGORIS, Konstantinos; ARIKIDIS, Nikolaos; KARAHA- LIOU, Anna; COSTARIDOU, Lena; PRATIKAKIS, Ioannis. “Computer-aided diagno- sis of mammographic masses based on a supervised content-based image retrieval approach”. In: *Pattern Recognition* 71 (2017), pp. 106–117. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2017.05.023. URL: http://www.sciencedirect.com/ science/article/pii/S0031320317302194.

WEI YANG, Kuanquan Wang; ZUO, Wangmeng. “Neighborhood component feature selection for high-dimensional data”. In: *Journal of Computers* 7.1* (Jan. 2012), 162– 168*. DOI: 10.4304/jcp.7.1.161-168.

WEI, Chia-Hung; CHEN, Sherry Y.; LIU, Xiaohui. “Mammogram retrieval on similar mass lesions”. In: *Computer Methods and Programs in Biomedicine* 106.3 (2012), pp. 234–248. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2010.09.002.