Investigating the performance of a CAD_x scheme for mammography in specific BIRADS categories

Andreadis I., Nikita K.
Department of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
iandr@biosim.ntua.gr, knikita@ece.ntua.gr

Spyrou G.
Informatics Laboratory-Biomedical Research Foundation
Academy of Athens
Athens, Greece
gspyrou@bioacademy.gr

***Abstract*— A Computer Aided Diagnosis (CAD_x) pipeline has already been introduced to discriminate between benign and malignant clusters of microcalcifications (MCs). In this study, we evaluate the specific methodologies using cases from publicly available databases of mammograms, the MIAS database and the Digital Database of Screening Mammography (DDSM). Specifically, we investigate various subsets of regions of interest (ROIs) containing cluster of MCs, following the BIRADS assessment performed by radiologists who have participated in the preparation of the databases. The obtained results indicate that there are specific BIRADS categories where the proposed system provides high classification results. Furthermore, it seems that the best results are achieved in obscure cases, where the radiologists have doubts about their diagnosis and recommend extra medical examinations and short follow-up. The reported results indicate the potential of the proposed CAD_x system to assist the diagnostic task of radiologists by providing a reliable second opinion concerning the diagnosis of a cluster of MCs.**

Keywords—microcalcifications; computer aided diagnosis; support vector machines; BIRADS standard

I. INTRODUCTION

Microcalcifications (MCs) are one of the most important radiographic findings associated to the existence of breast cancer [1]. Despite the fact that mammography is considered during the last years the most prominent technique for screening the breast cancer, there are still important inherent limitations of the method that may lead to missed findings or false diagnostic decisions [2]. Additionally, the subtle nature of these findings makes their interpretation a difficult task, even for experienced radiologists [1].

These factors led to the development of computer aided diagnosis (CAD_x) systems, whose role is to assist the diagnostic task of the radiologists by providing a reliable second opinion [3]. The basic pipeline implemented in a CAD_x system includes independent steps, ending in a classification phase that discriminates between benign and malignant findings. In other words, the final output of such systems is usually a binary prediction concerning the benignity or malignancy of the region of interest (ROI) containing the cluster of MCs. However, this is not the common practice followed by the radiologists when analyzing mammograms. The radiologists follow the BIRADS standard for their analysis [4], according to which their final assessment is not a binary value but a discrete value ranging from 0 to 5, corresponding to different risk of malignancy and alternative medical actions.

An extended review of the studies focusing on the CAD_x diagnosis of breast MCs may be found in [5]. Despite the great number of such studies reported in the literature and the variety of CAD methodologies proposed, it is hard to extract secure conclusions about their contribution on the improvement of the diagnostic process. First of all, the majority of the published studies in the literature use small subsets of mammograms (<200) [5]. As a result no conclusions with high generalization ability may be extracted and the straightforward comparison of their performance is infeasible. Additionally, in the majority of studies, no information on the composition of the dataset is usually provided and therefore the grade of difficulty for the considered cases is unknown.

To overcome these limitations, public databases should be investigated in order to be able to extract conclusions with high generalization ability, as well as to make the comparison between different studies feasible. The most commonly used databases containing digitized mammograms are the MIAS database [6] and the Digital Database of Screening Mammography (DDSM) [7]. The former contains 22 cases with annotated clusters of MCs, while the latter is the largest publicly available database, since it contains more than 1700 corresponding cases.

In a previous work [8], we have already proposed CAD_x methodologies and evaluated them on all the available cases of the DDSM database in order to study the effect of the breast density and the subtlety of findings on the performance of the CAD_x system. We have indicated that the system presents steady behavior and works towards the right direction, since its performance was affected by the same factors that affect the radiologists in daily clinical practice. It is noteworthy that the CAD_x system outperformed the radiologists' performance in the majority of the considered subsets of mammograms. However, we have not studied the performance of the proposed framework using the MIAS database, which is the other popular choice in such applications. Furthermore, despite having already evaluated the role of mammograms' inherent

properties to the performance of the proposed pipeline, we have not studied how the system performs on subsets of cases, depending on the initial BIRADS assessment of the radiologists who participated in the preparation of the DDSM database. As discussed earlier, a CAD_x system should not be used independently from the radiologists, but assist instead their task by providing an additional diagnosis. The a priori knowledge of critical radiologists' recommendations may be exploited in the CAD_x pipeline to refine the diagnostic process. Therefore, in this work we evaluate the role of the radiologists' categorization in terms of BIRADS rating. To the best of our knowledge, such a study has not been performed in large scale, using a variety of different mammograms that have been classified in various BIRADS categories by experienced radiologists. Through the evaluation of the CAD_x pipeline in specific BIRADS categories, we may further investigate the possibility to improve the performance of the computer aided diagnosis component, by using alternative computational algorithms.

As a result, in this study, we work on two main axes: firstly, we apply for reasons of completeness the CAD_x framework on the MIAS database in order to investigate whether the system discriminates satisfactorily benign and malignant cases. Secondly, we apply the proposed CAD_x framework on all the available cases of the DDSM database in order to study the effect of the initial BIRADS assessment on system's performance and investigate the potential to create classification schemes oriented for cases of specific BIRADS categories.

II. Methods and materials

A. CAD_x pipeline

The CAD_x methodologies evaluated in the current study have been introduced in a previous work [8]. Initially, a segmentation algorithm is applied on the initial ROI in order to segment the MCs from the surrounding breast tissue and omit the background. The feature extraction phase is then implemented, where a great variety of image features are extracted. To this end, the analysis of the ROI includes shape analysis of the individual MCs, analysis of the morphology of the whole cluster and texture analysis (first order and second order statistics based on the grey level occurrence matrices) of the surrounding tissue. A total of 188 different image features are computed.

Due to the great number of features extracted, a feature selection step is prerequisite in order to retain the most significant features for the discrimination task. The wrapper method Sequential Forward Selection (SFS) is used to locate the best subset of features. For the classification phase, a Support Vector Machines classifier has been implemented, using the Gaussian radial basis function (rbf) kernel. For the parameterization of the classifier, 10-fold cross validation has been followed to adjust the regularization parameter C and the parameter g of the rbf kernel. For performance evaluation, the Leave One Out (LOO) method has been used and receiver operating characteristic (ROC) analysis has been performed. The LIBSVM library has been used for our measurements [9].

B. BIRADS standard

As discussed previously, the radiologists perform their diagnosis based on the BIRADS standard [4]. According to the provided guidelines, a case is classified using a rating ranging from 0 to 5. According to the specific rating, different risk of malignancy is provided and different medical actions are recommended. Specifically:

- *BIRADS 0*: incomplete diagnosis, extra medical examinations are required.
- *BIRADS 1*: negative, no findings observed.
- *BIRADS 2*: benign findings.
- *BIRADS 3*: probably benign findings, short follow up after six months is recommended.
- *BIRADS 4*: suspicious abnormality, biopsy recommended.
- *BIRADS 5*: Highly suggestive of malignancy findings, biopsy recommended.

Cases with obvious signs for benignity or malignancy are classified as BIRADS 2 or 5 respectively. On the other hand, cases classified as BIRADS 3 are considered probably benign but seem to present some suspicious features that force the radiologist to recommend short follow up. These are obscure cases that may need a second opinion to support their categorization. Cases where the radiologist needs extra medical examinations to perform his diagnosis are classified as BIRADS 0. Finally, cases classified as BIRADS 4 are considered suspicious of malignancy and immediate biopsy is followed.

It is obvious that there are cases where the radiologists may have doubts about their diagnosis and as a result an extra diagnostic component may be of valuable importance for the diagnostic process. For this reason, we want to investigate the performance of the proposed CAD_x scheme in specific BIRADS categories.

C. Databases of mammograms

The first database considered is the MIAS database that consists probably the most popular choice for studies related to the automated image analysis of mammograms. The specific database presents many advantages, such as consistent format of the provided mammograms and precise description of the ROI containing the annotated finding. However, the available information is not in accordance with the BIRADS standard, since no assessment by radiologists is provided and the density rating does not follow the official guidelines. Additionally, the size of the specific database is quite small, as it contains only 22 ROIs with annotated clusters of MCs. Due to the specific disadvantages, the MIAS database has been used in the specific study only for reasons of completeness, in order to evaluate the proposed CAD_x pipeline and compare the achieved performance with the corresponding values reported in the literature.

The second database considered is the DDSM database that currently consists the largest publicly available database containing digitized mammograms. All cases contain

mammograms from both mediolateral (MLO) and craniocaudal (CC) views, as well as information on the boundaries of the annotated regions, where the lesion has been detected and biopsy has been done. For each case, there is also extra information concerning: (i) the density of the breast, according to the BIRADS standard, (ii) the subtlety rating, which is a subjective measure determined by experienced radiologists, who participated in the preparation of the DDSM database and (iii) the BIRADS assessment performed by the corresponding radiologist. We used almost all the available ROIs containing cluster of MCs, ending in a dataset consisted of 1715 ROIs that consists, to the best of our knowledge, the largest subset that has been used for the computer aided diagnosis of MCs.

Since the BIRADS assessment is provided in the files of DDSM database, we exploit the specific source in order to investigate in this study the performance of the proposed methodologies in different subsets of cases, depending on the BIRADS assessment of the radiologists.

III. Results and Discussion

A. Evaluation on the MIAS database

The MIAS database has been extensively used in the literature for the automated analysis of mammograms. We exploited all the 22 available ROIs to evaluate the performance of the proposed CAD_x methodologies on the specific subset of mammograms. Consequently, we applied then the proposed CAD_x framework discussed in section II, paragraph A and we performed ROC analysis, recording the values of accuracy (ACC), sensitivity (SN), specificity (SP) and Area Under Curve (A_z). In Table I, we present the number of cases (#Cases, inside the brackets we fill the number of benign and malignant cases respectively) and the results obtained for the considered metrics.

TABLE I
CAD_x PERFORMANCE ON THE MIAS DATABASE

#Cases	*CAD_x*			
	ACC	*SN*	*SP*	*A_z*
22 (10/12)	0.909	0.833	1.000	0.942

The reported results are quite high, achieving A_z value greater than 0.9. This fact reveals the potential of the proposed pipeline towards the correct discrimination between benign and malignant cases. Due to the fact that the size of the subset is quite small, further evaluation is required using larger datasets.

B. Effect of the BIRADS assessment

As discussed previously, the MIAS database could be exploited to evaluate the performance of the proposed CAD_x framework. However, there was no additional information concerning the assessment that has been prior performed by radiologists. For this reason, the current section includes the results observed when using the DDSM database for performance analysis.

Each cluster in the DDSM database contains the assessment according to the BIRADS standard that has been performed by experienced radiologists who participated in the preparation of the database. Our aim is to investigate the performance of the framework when using cases of a specific BIRADS category. For this reason, we divided the initial dataset of 1715 cases on four subsets, based on the BIRADS assessment value of each cluster. Specifically, since BIRADS 2 and BIRADS 5 category contains almost exclusively benign and malignant cases respectively, it is not feasible to train new classifiers using the specific categories independently. For this reason, we combined the cases of these two categories, so as to form a relatively balanced dataset.

TABLE II
CAD_x PERFORMANCE DEPENDING ON RADIOLOGISTS' BIRADS ASSESSMENT

BIRADS category	*#Cases*	*CAD_x*			
		ACC	*SN*	*SP*	*A_z*
0	76 (37/39)	0.697	0.703	0.692	0.691
2+5	323 (95/228)	0.82	0.886	0.662	0.838
3	116 (61/55)	0.905	0.927	0.885	0.955
4	1200 (684/516)	0.634	0.779	0.525	0.668

We applied then the proposed CAD_x framework and we performed ROC analysis for each different BIRADS category. As in the case of the MIAS database, we present in table II the performance measures for each category. Fig. 1 presents the ROC curves achieved by the classifier for the cases of each different BIRADS category.

C. Discussion

As far as the results using the MIAS database are concerned, we observe that the proposed methods performs indeed towards the right direction. The achieved results are quite high, indicating the potential of the proposed CAD_x pipeline to discriminate correctly benign and malignant cases. The fact that the number of considered cases is not quite large discourages us from extracting conclusions with high generalization ability. However, since the MIAS database is a publicly available source of mammograms, we may proceed on a straightforward comparison with other related studies reported in the literature that have used the same cases.

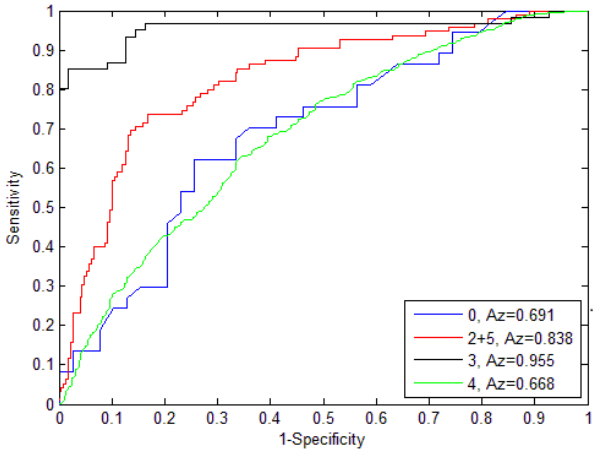

Fig. 1. ROC curves for each BIRADS category

Specifically, Papadopoulos et al. [10] reported a value for A_z equal to 0.81, while Chen et al. [11] reported $A_z = 0.92$. Jian et al. [12] did not include the A_z values in their study, that is the most powerful metric for comparison in two class classification problems. Instead, they reported values for the accuracy that was found to be equal to 95.8% without using wavelets and 100% using a dual-tree complex wavelet transform.

Concerning the effect of BIRADS assessment, we may extract valuable conclusions analyzing Table II. The achieved A_z value in all subsets is greater than the threshold of randomness (0.5). This fact indicates that the proposed system does not perform randomly but towards the right direction for the correct discrimination between benign and malignant cases. Specifically, the value of A_z for the subset of BIRADS 2+5 is almost equal to 0.84. We remind that if the ROI is classified as BIRADS 2 or 5 by the radiologist, the case is highly suggestive for benignity or malignancy respectively. As a result, the value 0.84 is quite satisfactory and reveals that the system discriminates correctly the cases when more "obvious" cases are under consideration. We may proceed on corresponding observations when considering the subsets of cases classified as BIRADS 0. These cases are considered of higher grade of difficulty, since the radiologists are not able to perform diagnosis and extra medical examinations are recommended. Despite the fact that the performance is not quite high (A_z = 0.73), it is encouraging that the specific value is achieved for difficult cases, where the radiologists need a reliable second opinion.

The lowest performance (A_z = 0.668) is observed when considering cases classified as BIRADS 4. This fact was expected since the current subset contains a great number of cases (1200) and as a result there is a great variety of different ROIs without specific trends. We remind that the specific category includes the cases that are considered suspicious for malignancy and have been prompted for biopsy by the radiologists. However, we observe that the majority of cases proved to be benign (57%), after the biopsy test. This fact indicates that the radiologists present high sensitivity aiming at early diagnosis of breast cancer but achieve simultaneously low levels of specificity, as there appears to exist a great number of cases that were falsely recommended for biopsy.

The most important result concerns the cases of BIRADS 3 category. The CAD_x framework achieves for the specific category the highest classification performance (A_z equal to 0.955). This performance reveals the potential of the automated algorithms to discriminate the benign and malignant clusters, in cases where the radiologists present serious concerns about their diagnosis. We remind that cases classified as BIRADS 3 are considered probably benign, but short follow-up is recommended. After the biopsy performed, it appears that a great percentage of them (47.4%) are actually malignant. The high discriminative value achieved by the CAD_x pipeline is quite promising, since it seems that the system may provide a reliable second opinion in obscure diagnostic cases where the radiologists have doubts for their diagnosis.

IV. Conclusions

In this study, we evaluated computer aided diagnosis methodologies for clusters of MCs using publicly available databases of mammograms. The achieved results reveal that there are subsets of cases where the proposed methods present high discrimination ability and perform properly towards the diagnosis of breast cancer. However, the role of a CAD system is not to act independently but assist the radiologist by providing a reliable second opinion. For this reason, we investigated the performance of the CAD_x framework depending on the BIRADS assessment performed by radiologists who examined the same cases. We indicated that the system performs satisfactorily in subsets of obscure cases, such as cases classified as BIRADS 3 or BIRADS 0, where the radiologists recommend short follow up or extra medical examinations. This observation implies that the proposed CAD_x framework may provide valuable help to the radiologists, if adopted during the diagnostic process.

The evaluation performed in the current study is quite important, as we investigate the proposed methodologies in a framework of daily clinical practice, where the radiologist assess mammograms in terms of BIRADS rating. As we have already discussed, a CAD_x system should interact with the radiologist during the diagnostic process. As a result, critical radiologists' recommendation, and particularly their final assessment, should be exploited in the CAD_x pipeline. Through this interaction, the diagnostic process of the CAD_x system is refined, as the best computational algorithms are selected for each different mammogram, while the radiologist may exploit the second opinion provided by the CAD_x system to confirm or reassess his diagnosis. The main question that has to be faced in the future is how we may simulate the interaction between radiologists and CAD_x system in order to improve the diagnostic role of the physician. Important factors, such as the density of the breast or the initial assessment of the radiologist without the use of the CAD_x system may be exploited towards this direction. As a result, the conclusions extracted in this study concerning the role of the initial BIRADS assessment of the radiologist provide us a proper baseline to work towards the refinement of the diagnostic process, focusing on the simulation of the interaction between radiologists and CAD systems.

References

[1] M. Lanyi, " Microcalcifications in the breast—a blessing or a curse? A critical review," Diagn. Imaging Clin. Med., vol. 54, pp. 126–145, 1985.

[2] H. H. Ng and M. Muttarak, "Advances in mammography have improved early detection of breast cancer," J. HK Coll. Radiol., vol. 6, no. 3, pp. 126–131, 2003.

[3] M. Giger, H. Chan and J. Boone, "Anniversary paper: history and status of CAD and quantitative image analysis: the role of medical physics and AAPM," Med. Phys., vol. 35, pp. 5799–5820, 2008.

[4] American College of Radiology (ACR), ACR Breast Imaging Reporting and Data System, Breast Imaging Atlas, 4th Edition, Reston, VA. USA, 2003.

[5] M. Elter and A. Horsh, "CADx of mammographic masses and clustered microcalcifications: a review," Med. Phys., vol. 36, pp. 2052–2068, 2009.

[6] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt and C. Boggis, “The mammographic images analysis society digital mammogram database,” Exerpta Med., vol. 1069, pp. 375-378, 1994.

[7] M. Heath, K. Bowyer, D. Kopand, R. Moore and W. Kegelmeyer. “The Digital Database for Screening Mammography,”in 2001 Proc. of the 5th IWDM, Yaffe M. Medical Physics Publishing, pp. 212-218.

[8] I. Andreadis, G. Spyrou, P. Ligomenides and K. Nikita, “Variations on breast density and subtlety of the findings require different computational intelligence pipelines for the diagnosis of clustered microcalcifications,” presented at the 13th Int. IEEE Conf. BioInformatics and BioEngineering, Chania, Greece, Nov. 10-13, 2013.

[9] Chang CC and Lin CJ2001 LIBSVM:alibraryforsupport vector machines Software available at http://www.csie.ntu.edu.tw/□cjlin/libsvm

[10] A. Papadopoulos, D. Fotiadis, A. Likas, “Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines,” Artif. Intell. Med., vol. 34, pp. 141-50, 2005.

[11] Z. Chen, A. Oliver, E. Denton, C. Boggis and R. Zwiggelaar, “Classification of microcalcification clusters using topological structure features,” Medical Image Understanding and Analysis, pp 37-42 Swansea, Wales, UK. July 2012.

[12] W. Jian, X. Sun and S. Luo, “Computer-aided diagnosis of breast microcalcifications based on dual-tree complex wavelet transform,” Biomedical Engineering, BioMedical Engineering OnLine 2012 11:96, 2012.