Graphical Oracles to Assess Computer-Aided Diagnosis Systems: A Case Study in Mammogram Masses and Calcifications Detection

Vagner Mendonça Gonçalves[1,3], Márcio E. Delamaro[2,3], Fátima L. S. Nunes[3]
[1]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, São Paulo, Brazil
[2]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brazil
[3]Laboratory of Computer Applications for Health Care,
Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, Brazil
vagner.goncalves@ifsp.edu.br, delamaro@icmc.usp.br, fatima.nunes@usp.br

***Abstract*—Computer-Aided Diagnosis (CAD) systems provide a second opinion to health professionals about the possible existence of an anomaly. Evaluation of CAD systems is a challenge and most of the traditional metrics requires the constant participation of experts. This paper presents an approach for evaluating CAD systems using concepts of Content-Based Image Retrieval and graphical oracles. After implementing feature descriptors and selecting three similarity functions, two metrics are proposed to measure the efficiency of CAD systems. A case study was conducted considering three simulated CAD systems to detect masses and calcifications in mammographic images. The results indicated that the our approach is as robust as traditional metrics with respect to performance evaluation. However, our approach is more flexible than traditional metrics because evaluators can choose the more adequate features to assess a particular CAD system.**

***Keywords*—Computer-Aided Diagnosis, Graphical Oracles, Content-Based Image Retrieval, Medical Images**

I. Introduction

Computer-Aided Diagnosis (CAD) systems provide a second opinion to radiologists, thus contributing to decision-making about a diagnosis. In general, CAD systems involve activities of segmentation, detection and/or classification of structures in medical images [1], [2].

One of the greatest difficulties found during the development of CAD systems is related to evaluation. It is not easy to know whether a technique is effective or not, because results can vary according to datasets used. Even though image databases suitable to evaluate the performance of CAD systems are available, this evaluation happens with constant participation of human resources in order to determine whether the system result is correct or not [3]. Therefore, to define a more objective way to assess this modality of systems constitutes an important contribution to this research area as it can aid in the establishment of standards to measure the performance of these systems.

Evaluation metrics, such as accuracy, sensitivity, specificity, and the Receiver Operating Characteristic (ROC) curve have been used by many researchers to assess CAD systems [1], [4]–[6]. A new software testing approach to programs with graphical outputs was proposed in [7]. This technique applies concepts of Content-Based Image Retrieval (CBIR) to automate graphical oracles in software testing. These oracles are able to compare an output of the program under test with an image provided as reference for executing the program. Thereby, from criteria defined by a tester, they provide a verdict about the correctness of the output considered. Additionally, the researchers propose a tool to support the definition and the use of graphical oracles in test activities of programs with graphical outputs, the Oracle for Images (O-FIm) framework.

This paper proposes, applies and validates an approach to assess CAD systems using graphical oracles. In order to validate the proposed approach, a case study was conducted simulating three CAD systems with different characteristics and performance, according to traditional metrics widely used in the area. The systems evaluated simulate mammogram masses and calcifications detection.

The remaining of this paper is organized as follows. Section II presents a background in relation to CAD systems evaluation. In Section III the materials and methods are defined. In Section IV results are available. Finally, Section V presents the conclusions.

II. Background

According to the results of a Systematic Review conducted previously [5], the main metrics applied to assess CAD systems are shown in Table I. Sensitivity and specificity measures are part of this set of traditional statistical metrics. They are based on concepts of diagnosis rate considered true-positive (TP), true-negative (TN), false-positive (FP) e false-negative (FN) [8]. Diagnosis rate refers to the number of cases that the CAD system processed correct or incorrectly. TP and TN are the number of cases – positive and negative, respectively – that the system detected correctly. FP and FN refer to the number of cases that the system detected incorrectly.

In any CAD system, the performance is directly proportional to its sensitivity and specificity. However, evaluating a CAD system using only one sensitivity value and one specificity

978-1-7281-7539-3/20/$31.00 ©2020 IEEE

TABLE I
MAIN METRICS APPLIED TO ASSESS CAD SYSTEMS.

Metrics	Formula	Description
Accuracy (Correct classification / detection rate)	$\frac{TP + TN}{TP + TN + FP + FN}(\times 100\%)$	Percentage of abnormalities and normal structures correctly classified/detected.
Negative predictive value	$\frac{TN}{TN + FN}(\times 100\%)$	Percentage of normal structures detected / classified that does not really represent abnormalities.
Precision (or Positive predictive value)	$\frac{TP}{TP + FP}(\times 100\%)$	Percentage of detected structures that are actually abnormalities.
Sensitivity (or Recall)	$\frac{TP}{TP + FN}(\times 100\%)$	Percentage of abnormalities correctly detected/classified.
Specificity	$\frac{TN}{TN + FP}(\times 100\%)$	Percentage of normal structures not incorrectly detected/classified as possible abnormalities.

Source: [5].

value is not enough to testify its quality, mainly when the objective is to compare several systems or techniques. Usually, CAD systems are parameterizable to define different levels of sensitivity and specificity. In this scenario the concept of Receiver Operating Characteristic Curve or, simply, ROC Curve arises. Such curve represents the sensitivity as a function of the false-positive fraction (FP fraction $= 1 -$ specificity) [9], [10]. An ideal CAD should present maximum sensitivity and minimum FP fraction. As the attainment of ideal systems is unlikely, an acceptable system reaches the greater sensitivity possible, with a FP fraction acceptable within the medical context.

A. Graphical Oracles and the O-FIm Framework

A test oracle is an effective mechanism that indicates to the tester if the output of a program, obtained using test data, is acceptable or not [11], [12]. The O-FIm framework is a tool for programmer testers. It uses concepts of CBIR to configure oracles in order to support program testing with graphical outputs. Such oracles were defined in [7] as graphical oracles.

CBIR is defined in [13] as a technology which aids digital images organization by using their visual content. The main components of the systems that use this technology consist in feature descriptors, similarity functions and the image database itself.

Feature descriptors are used to compose a feature vector for each image. In general, after comparisons between feature vectors, CBIR systems return the images most similar to an image provided as a model. This concept was adapted in the context of the framework aforementioned aiming at enabling the distance computation between feature vectors of two images, identifying how similar they are. The model image is related to an image defined as a reference and the "test image" (output of a program under test) is the one the oracle will indicate as similar or different to the model image.

The O-FIm kernel provides an Application Programming Interface that allows the creation of oracles in a simple way. To perform a test, the tester must provide a textual description (oracle description) that indicates which components (similarity functions and feature descriptors) will be applied in the test as well as their parameters when required. Then, O-FIm uses this description to create the oracle. Another parameter indicated in the oracle description is the threshold value, which indicates the maximum acceptable distance between two compared images, so that they are considered equivalent.

III. MATERIAL AND METHODS

Our evaluation approach consists of defining feature descriptors and similarity functions in order to compose a graphical oracle. This graphical oracle is used to compare output images of a CAD system with their respective images considered correct. The performance of a CAD system can be measured by the percentage of output images considered correct. One image is considered correct if its distance to its model image is smaller than a threshold value.

However, to define this threshold value is not an easy task. Thus, we propose additional metrics for measuring the CAD performance, both of them aiming at attributing a grade in the $[0, 10]$ interval to the system.

After defining the graphical oracle and the performance metrics, we conducted a case study considering CAD systems to detect masses and calcifications in mammograms. The characteristics of these images and the systems are detailed next, as well as our proposal to evaluate the systems.

A. Case Study

To apply our approach based on graphical oracles we used cases extracted from the DDSM mammographic image database [14]. Mammograms of this database are accompanied, when appropriate, by the manual demarcation of the suspicious region (mass or calcification) performed by an experienced radiologist.

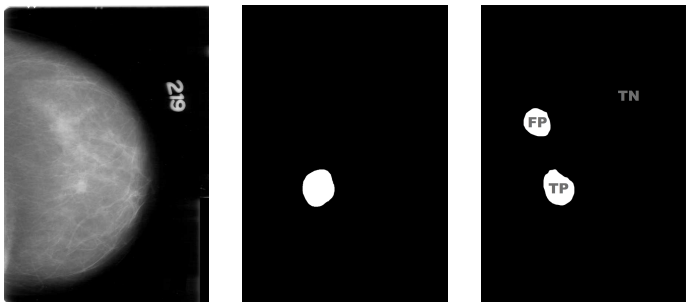
The DDSM basis contains 2620 cases, arranged in 43 volumes. For this case study we selected two volumes, indexed in the database by names `benign_10` and `cancer_13`. The first consists of 21 benign cases containing images acquired with 12 bits of contrast resolution. The latter also contains 21 malignant cases with images with the same contrast resolution.

In each case we selected a mammogram containing a suspicious region (mass or calcification) demarcated by the radiologist, to compose the set of images processed by simulated CAD systems. Thus, 42 mammograms were used in the tests described herein. It is important to mention that the objective was to compose the set of images with the same number of images representing each of the diagnosed cases (benign and cancer).

We developed three simulated CAD systems, each generating different results. The designation of simulated systems refers to the fact that we used classical techniques of image processing to detect suspicious regions on mammograms, without claiming whether the results of the system were satisfactory. Thereby,

the first and the second systems used histogram equalization, thresholding and mathematical morphology operators (erosion and dilation) in different ways. In the third case, the images were manually segmented by using a general image processing program. In this first moment, we chose to use simulated CAD systems to guarantee total comparability among them, given the difficulty in obtaining three systems with equivalent objectives and fully comparable for the same image database.

A set of output images was created for each simulated CAD system. To systematize the evaluation results, we defined the following criteria. Every region detected whose intersection with the area demarcated by the radiologist is equal or greater than 70% was considered as a true-positive (TP). Any other region detected was considered as a false-positive (FP). The background with no structures segmented corresponded to a true negative (TN). A region that belongs to the background of the output image of the simulated CAD system, but which was demarcated as a suspicious region by the radiologist on the original mammogram, was considered a false negative (FN). To exemplify the adopted criteria, Figure 1 presents the images obtained after processing a mammogram selected from malignant case 4034.



(a) Original mammogram.

(b) Suspicious region manually demarcated by a radiologist.

(c) Simulated output of a CAD system to detect suspicious regions with signals TP, TN and FP.

Fig. 1. Example of images from malignant case 4034 of the DDSM database.

Systematizing the sets of images used in this work, we define the set GT as the set of reference images containing the regions manually demarcated by the radiologist provided by DDSM database:

$$GT = \{G_1, G_2, G_3, \ldots, G_{42}\}$$

The sets of simulated outputs of each of the three CAD systems define the sets $Output_1$, $Output_2$ and $Output_3$.

$$Output_1 = \{O_{1,1}, O_{1,2}, O_{1,3}, \ldots, O_{1,42}\}$$

$$Output_2 = \{O_{2,1}, O_{2,2}, O_{2,3}, \ldots, O_{2,42}\}$$

$$Output_3 = \{O_{3,1}, O_{3,2}, O_{3,3}, \ldots, O_{3,42}\}$$

The graphical oracles were configured from the definition of similarity functions and feature descriptors. Therefore, to propose and to implement an approach based on such oracles for evaluating CAD systems, we implemented and included different similarity functions and feature descriptors in the O-FIm framework.

Each experiment was repeated three times, using a different similarity function in each execution in order to allow comparison of results obtained by using distinct similarity functions. Results of a study previously conducted [15] defined three groups of similarity functions with same behavior. For this study, we selected a similarity function from each of these groups, namely: Canberra distance, χ^2 distance, and Euclidean distance.

The feature descriptors applied in each experiment were: area, center of mass, height, number of objects, perimeter, and width. All the features were implemented in order to normalize the computed values in the range $[0, 1]$.

B. General View of the CAD Evaluation Approach

Figure 2 presents a diagram with stages of the approach for CAD evaluation based on graphical oracles.

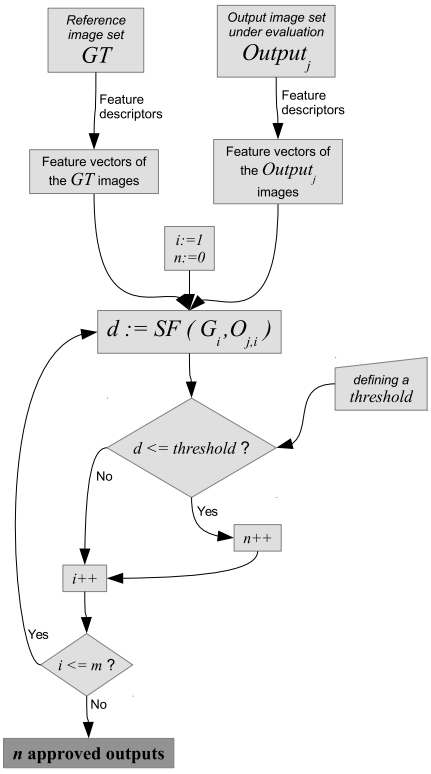


Fig. 2. Stages of the approach for CAD evaluation.

To evaluate each output image of each simulated CAD system ($O_{j,i}, j = 1 \ldots 3, i = 1 \ldots 42$), comparing them to a reference image ($G_i, i = 1 \ldots 42$), we defined the graphical oracle textually presented in Figure 3. The format of this graphical oracle was built using O-FIm framework (see Subsection II-A). The feature descriptors used were defined specifically for our case study, but they can be easily adapted for any CAD system, since some characteristics of the output images are known.

In addition, all the applied feature descriptors consider as background image all the pixels with zero value in grayscale

```
similarity SimilarityFunction
extractor Area { thr = 0 }
extractor CenterOfMassX { }
extractor CenterOfMassY { }
extractor Height { thr = 0 }
extractor NumObjects
{ normalizer = 10 thr = 0 }
extractor Perimeter { thr = 0 }
extractor Width { thr = 0 }
precision threshold
```

Fig. 3. Textual description of the graphical oracle defined for the experiments.

(meaning of the parameter `thr` that appears in the description of the graphical oracle).

For each similarity function used, a set of threshold values were defined with which the results were obtained and compared, defining an approval or a disapproval. Considering a similarity function SF and the oracle configured (Figure 3), we performed the procedures described below.

First, we calculated the maximum distance (MAX_{SF}) between vectors that can be computed according to the graphical oracle described. Since the values of the features used were normalized in the range $[0, 1]$, this value corresponds to the maximum difference between the null vector V_0, where all the features have value 0, and the vector maximum V_1, where all the features have value 1.0. Hence, the value MAX_{SF} corresponds to distance $SF(V_0, V_1)$ and it was used to define the threshold values as shown in Equation 1. MAX_{SF} is the maximum possible threshold that would approve any image under evaluation.

$$threshold = \beta \times MAX_{SF} \tag{1}$$

The values of MAX_{SF} computed were: 7.00 for Canberra distance, 3.5 for χ^2 distance, and 2.65 for Euclidean distance.

In our case study, we consider values of β varying from 0.05 to 0.5, taking intervals of 0.05. This value represents a percentage of the maximum possible threshold. Using different values of threshold in the tests we could evaluated the influence of this parameter in the results, as shown in the next section.

C. Defining Performance Metrics

A direct way of using our approach for evaluating a CAD system performance is comparison the percentages of output images approved, applying each of the similarity functions and threshold values for different values of β (Equation 1).

Although this measure is direct and easily calculated, it is dependent on a threshold value, as aforementioned. The objective was define a way to compare different systems without using threshold values. Then, we defined additional metrics to compute a grade in the $[0, 10]$ interval for the CAD system, where zero is the worst grade.

First, we define the set of distances obtained by comparing each output image of set $Output_j$ $(j = 1 \dots l)$ to its reference image of GT set, as shown in Equation 2.

$$Dist_{SF\,,\,j} = \{\, SF(O_{j,1}\,,\, G_1),\, SF(O_{j,2}\,,\, G_2),\, \dots,\, SF(O_{j,k}\,,\, G_k) \,\} \quad j = 1 \dots l, k = 1 \dots m \tag{2}$$

In our case study $l = 3$ (three CAD systems are compared) and $m = 42$ (42 images are considered in the test). After obtaining the distances set, two metrics were defined as follows.

The first performance metric calculated for CAD system j ($\texttt{Performance}_{1,j}$) is shown in Equation 3. The basic idea is to get a grade to the system in the range $[0, 10]$, with zero being the worst possible performance.

$$\texttt{Performance}_{1,j} = 10 - \frac{10}{MAX_{SF}} \times \texttt{AVERAGE}(Dist_{SF\,,\,j}) \tag{3}$$

where MAX_{SF} is the greater distance obtained considering each similarity function used computed by Equation 2 and $\texttt{AVERAGE}(Dist_{SF\,,\,j})$ is the average of distances obtained for each CAD system under evaluation.

All systems have initially grade equal 10. Thus, the grade of each system is decreased according to how much its average distance is close to the MAX_{SF}. When the average is equal to MAX_{SF}, i.e., all the distances are maximal, the system has the worst possible performance. In this way, 10 points are discounted from the system grade and it gets the minimum performance grade (zero). On the other hand, when the average is zero, i. e., all the distances obtained are minimal, system has the best possible performance, no point is deducted from the system grade. Thus, it gets the maximum performance grade (ten).

The second performance metrics ($\texttt{Performance}_{2,j}$), also defined based on MAX_{SF} value, uses the same principle of the $\texttt{Performance}_{1,j}$ metric, but considers the median of the set of distances, rather than the average (Equation 4). However, it is important to note that due to the definition of median, it is not necessary for all the distances to be equal to zero for the system to obtain the best score (ten). Also is not necessary for all the distances to be maximal for the system to get the worst possible score (zero). Furthermore, the median has the property of excluding non-standard results, i. e., defective images or outliers images have a smaller influence on the result computed.

$$\texttt{Performance}_{2,j} = 10 - \frac{10}{MAX_{SF}} \times \texttt{MEDIAN}(Dist_{SF\,,\,j}) \tag{4}$$

IV. Results and Discussion

Each of the simulated CAD systems was initially evaluated according to the traditional metrics most often cited in the literature and presented in Section II.

All the output images were manually analyzed and compared with the expected outputs provided by the DDSM database. The amount of TP, FP, FN and TN were recorded and used in the computation of the metrics. The values obtained for each of the three systems are shown in Table II.

According to the results in Table II the best system detection is CAD 3 because it reached best values for all the measurements when compared to the other two systems. However, choosing the best system between CAD 1 and CAD 2 is not as trivial either. The two had almost the same accuracy. CAD 2, meanwhile, showed better values for negative predictive value, precision, and sensitivity. CAD 1, when compared to CAD 2,

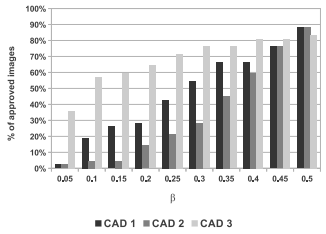
TABLE II
VALUES OBTAINED WITH TRADITIONAL EVALUATION METRICS.

Metric	CAD 1	CAD 2	CAD 3
Accuracy	0.37	0.37	0.67
FP fraction	0.60	0.69	0.28
Negative predictive value	0.59	0.69	0.71
Precision	0.17	0.20	0.61
Sensitivity	0.31	0.55	0.60
Specificity	0.40	0.31	0.72

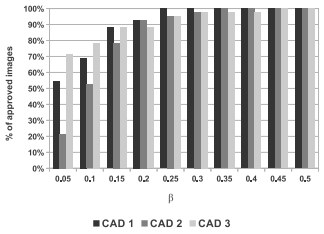
showed better values for false-positive fraction and specificity. Therefore, defining the best in this case, based on such metrics would depend on the requirements defined for the system by the users.

A. Application of the Approach Based on Graphical Oracles

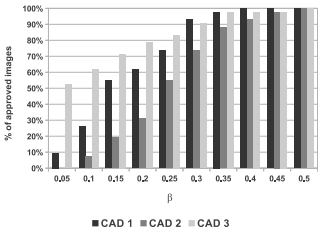
Figure 4 shows the percentage of output images approved for each CAD system, considering the approach for CAD evaluation described in Subsection III-B.



(a) Using Canberra distance.



(b) Using χ^2 distance.



(c) Using Euclidean distance.

Fig. 4. Comparison among the performances of the three systems simulated considering percentage of approved outputs.

Figure 4a shows the performance obtained using the Canberra distance. For most values of β tested, the CAD 3 performed better than the other two systems. For stricter thresholds (computed with the lowest values of β), CAD 3 showed much higher performance than the other two. Such result is very consistent with traditional metrics calculated.

Only for $\beta = 0.5$, CAD 3 had a lower performance. However, through the various experiments conducted in this study and in a previous work about evaluation of similarity functions [15] it was found that higher threshold values tend to equalize the assessment of different CAD systems. This occurs because the higher threshold values become the evaluation stricter, i. e., the resulting image of the CAD must be very close to its respective oracle for the output to be considered correct. Thus, these high thresholds consider small differences in the output images as errors and thus all systems tend to be poorly evaluated. This behavior was also observed in the results presented here.

Comparing CAD1 to CAD 2, still using results obtained with Canberra distance, for most values of β, CAD 1 usually performed better, with some exceptions when performance of both were similar. This result demonstrates that the feature descriptors used in this case can minimize the difficulty in assessing which of these two systems performed better. However, it is worth noting that depending on the feature descriptors used, the criteria may change and the assessment results as well.

The performance obtained when we applied the χ^2 distance is shown in Figure 4b. As observed in the graph, only for the two lowest values of β, the CAD 3 presented better performance. Using this similarity function the performance obtained grew faster, reaching the maximum value (100%) for relatively low thresholds. By the observed behavior, stricter threshold values (with β <0.15) continue to indicate that CAD 3 had the best performance. It can be concluded, therefore, that to obtain the same assessment rigidity as that obtained with Canberra distance, but using the χ^2 distance, it is necessary to determine a threshold level relatively lower than the one used in the former.

Also using the χ^2 distance, the CAD 1 presented better performance than CAD 2. The performances of these two systems were equal only when values of β produced thresholds that approved all the outputs under evaluation.

The performances obtained with the Euclidean distance are shown in Figure 4c. Again, for more stringent threshold values (with $\beta \leq 0.25$), CAD 3 showed better performance. From the value of $\beta = 0.3$ CAD 3 no longer showed the best performance, but the performance of all the systems began to converge to the maximum possible performance. The results obtained with this similarity function also revealed the better performance of CAD 1 when compared to CAD 2.

B. Application of the Proposed Performance Metrics

The grades for the three systems evaluated by the $\texttt{Performance}_{1,j}$ and $\texttt{Performance}_{2,j}$ metrics are presented in Table III. Considering the $\texttt{Performance}_{1,j}$ metric, the CAD 3 system performed better for all the three similarity functions. Comparing CAD 1 to CAD 2, CAD 1 was verified to perform better for the three functions used. With respect to the CAD 3 system, results are consistent with the traditional metrics calculated. The grade of the three systems are consistent with the performance obtained by the percentage of approved images. Furthermore, the performance metric proposed allows objectively comparing the performance of CAD 1 to the performance of CAD 2, which was not possible using traditional metrics.

With the $\texttt{Performance}_{2,j}$ metric, for the three similarity functions used, CAD 3 obtained the best performance. Comparing CAD 1 to CAD 2, again CAD 1 had better performance for the three similarity functions used. The consistency with the traditional metrics (with respect to the best performance

TABLE III
GRADES OBTAINED TO DETECTION SYSTEMS EVALUATED.

CAD	Metric	Canberra	χ_2	Euclidean
1	$\texttt{Performance}_{1,1}$	6.83	9.34	8.29
	$\texttt{Performance}_{2,1}$	7.27	9.56	8.64
2	$\texttt{Performance}_{1,2}$	6.40	8.98	7.56
	$\texttt{Performance}_{2,2}$	6.37	9.06	7.70
3	$\texttt{Performance}_{1,3}$	7.39	9.48	8.95
	$\texttt{Performance}_{2,3}$	9.28	9.98	9.68

obtained by CAD 3) and the performances obtained by the percentage of approved images (considering the grades of the three systems) was again evident.

C. Limitations and Strengths of the Proposed Approach

A limitation of our approach is the computational cost involved. The more images to be tested, the more feature vectors are computed. The higher the dimensionality of the vector, the higher is the processing required. This disadvantage can be minimized by optimizing the algorithms implemented with efficient indexing techniques. For example, when there is a known set of images with fixed size, it is not necessary to calculate all the same features if their values are efficiently stored. Additionally, techniques to reduce the dimensionality of the vectors can be used. Even with this limitation, the advantage of being able to adapt the approach to the evaluation criteria of each particular system, with definition of specific feature descriptors, makes our approach interesting and powerful, as demonstrated by the results presented in this article.

A key feature of this approach is that the evaluator can determine which criteria are important for testing the outputs of the system and then turn them into feature descriptors. Furthermore, it is possible to assess the overall quality of the detection system with only one performance metric, calculated from the results of a configured graphical oracle. Each traditional metric used in the area, by contrast, evaluates the performance of the system according to its own criteria (TP fraction, FP fraction, among others), requiring the analysis of various metrics together to reach a conclusion about the system performance.

V. CONCLUSIONS

We presented, discussed and validated an approach for evaluating CAD systems. This approach is based on graphical oracles which, in turn, are based on the concept of CBIR. In the tests we applied the O-FIm framework, a supporting tool for configuring and applying graphical oracles.

The results demonstrated the validity of the approach proposed and its consistency with the traditional metrics used for evaluating CAD systems for detection and classification of anomalies, such as sensitivity and specificity. Through experiments conducted, the approach was verified to be robust regarding the similarity function used, as well as being flexible and adaptable to evaluate CAD systems effectively. Additionally, it can be implemented without great efforts, since there is a framework with feature descriptors and similarity functions available to aid in this task.

ACKNOWLEDGMENT

The authors are grateful to Brazilian National Council of Scientific and Technological Development (CNPq): grants #309030/2019-6 and #308615/2018-2; São Paulo Research Foundation (FAPESP): grants #2010/01496-1 and #2019/06937-0; and National Institute of Science and Technology – Medicine Assisted by Scientific Computing (INCT-MACC): grant #157535/2017-7.

REFERENCES

[1] M. A. Al-antari, M. A. Al-masni, M.-T. Choi, S.-M. Han, and T.-S. Kim, "A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification," *Int J Med Inform*, vol. 117, pp. 44–54, sep. 2018.

[2] K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Comput Med Imaging Graph*, vol. 31, no. 4-5, pp. 198–211, 2007.

[3] E. Y. Jeong, H. L. Kim, E. J. Ha, S. Y. Park, Y. J. Cho, and M. Han, "Computer-aided diagnosis system for thyroid nodules on ultrasonography: diagnostic performance and reproducibility based on the experience level of operators," *Eur Radiol*, vol. 29, pp. 1978–1985, apr. 2019.

[4] L. G. Falconí, M. Pérez, and W. G. Aguilar, "Transfer learning in breast mammogram abnormalities classification with mobilenet and nasnet," in *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, jun. 2019, pp. 109–114.

[5] V. M. Gonçalves, M. E. Delamaro, and F. L. S. Nunes, "A systematic review on the evaluation and characteristics of computer-aided diagnosis systems," *Rev Bras Eng Bioméd*, vol. 30, no. 4, pp. 59–72, oct./dec. 2014.

[6] H.-P. Chan, K. Doi, S. Galhotra, C. J. Vyborny, H. MacMahon, and P. M. Jokich, "Image feature analysis and computer-aided diagnosis in digital radiography. I. automated detection of microcalcifications in mammography," *Med Phys*, vol. 14, no. 4, pp. 538–548, jul. 1987.

[7] M. E. Delamaro, F. L. S. Nunes, and R. A. P. Oliveira, "Using concepts of content-based image retrieval to implement graphical testing oracles," *Softw Test Verif Rel*, vol. 23, pp. 171–198, 2013.

[8] R. Garnavi, M. Aldeen, and M. E. Celebi, "Weighted performance index for objective evaluation of border detection methods in dermoscopy images," *Skin Res Technol*, vol. 17, no. 1, pp. 35–44, 2011.

[9] C. E. Metz, "Evaluation of CAD methods," in *Computer-Aided Diagnosis in Medical Imaging*, ser. International Congress Series, K. Doi, H. MacMahon, M. L. Giger, and K. R. Hoffmann, Eds., vol. 1182. Amsterdam, Netherlands: Elsevier Science BV, 1999, pp. 543–554.

[10] R. F. Wagner, C. E. Metz, and G. Campbell, "Assessment of medical imaging systems and computer aids: a tutorial review," *Acad Radiol*, vol. 14, no. 6, pp. 723–748, jun. 2007.

[11] L. Baresi and M. Young, "Test oracles," University of Oregon, Dept. of Computer and Information Science, University of Oregon, Dept. of Computer and Information Science, Eugene, Oregon, USA, Technical Report CIS-TR-01-02, aug. 2001.

[12] D. Hoffman, "A taxonomy for test oracles," in *Proceedings of the 11th International Quality Week*, San Francisco, CA, USA, may 1998, pp. 1–8.

[13] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: ideas, influences, and trends of the new age," *ACM Comput Surv*, vol. 40, no. 2, pp. 1–60, 2008.

[14] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The Digital Database for Screening Mammography," in *Proceedings of the Fifth International Workshop on Digital Mammography*, M. J. Yaffe, Ed. Medical Physics Publishing, 2001, pp. 212–218.

[15] F. L. S. Nunes, M. E. Delamaro, V. M. Gonçalves, and M. S. Lauretto, "CBIR based testing oracles: an experimental evaluation of similarity functions," *Int J Softw Eng Know*, vol. 25, no. 8, pp. 1271–1306, 2015.