Learning lung nodule similarity using a genetic algorithm

Kerry A. Seitz, Jr.[a], Anne-Marie Giuca[b], Jacob Furst[c], Daniela Raicu[c]

[a]Trinity University, One Trinity Place, San Antonio, TX, USA 78212-7200;
[b]Pomona College, 333 North College Way, Claremont, CA, USA 91711-6312;
[c]DePaul University, 243 South Wabash Avenue, Chicago, IL, USA 60604

ABSTRACT

The effectiveness and efficiency of content-based image retrieval (CBIR) can be improved by determining an optimal combination of image features to use in determining similarity between images. This combination of features can be optimized using a genetic algorithm (GA). Although several studies have used genetic algorithms to refine image features and similarity measures in CBIR, the present study is the first to apply these techniques to medical image retrieval. By implementing a GA to test different combinations of image features for pulmonary nodules in CT scans, the set of image features was reduced to 29 features from a total of 63 extracted features. The performance of the CBIR system was assessed by calculating the average precision across all query nodules. The precision values obtained using the GA-reduced set of features were significantly higher than those found using all 63 image features. Using radiologist-annotated malignancy ratings as ground truth resulted in an average precision of 85.95% after 3 images retrieved per query nodule when using the feature set identified by the GA. Using computer-predicted malignancy ratings as ground truth resulted in an average precision of 86.91% after 3 images retrieved. The results suggest that in the absence of radiologist semantic ratings, using computer-predicted malignancy as ground truth is a valid substitute given the closeness of the two precision values.

Keywords: content-based image retrieval, genetic algorithm, computer-aided diagnosis, lung nodules, CT scans, LIDC

1. INTRODUCTION

Lung cancer accounts for the highest number of cancer deaths in the United States each year, constituting 28% and 26% of all cancer deaths in 2011 for males and females, respectively.[1] Computed tomography (CT) scans can assist radiologists in early detection of lung nodules, which increases the likelihood of a patient's survival.[2] In order to improve lung nodule detection, computer-aided diagnosis (CAD) is effective as a second opinion for radiologists in clinical settings.[3] In CAD systems, content-based image retrieval (CBIR) can be used to find images of diagnosed lung nodules similar to an undiagnosed nodule of interest.[4] With the aid of similar images, radiologists' diagnoses of lung nodules in CT scans can be significantly improved.[5] In CBIR, image features are extracted and stored in feature vectors, and a similarity measure is used to calculate the similarity between images based on these features.[6] However, a semantic gap still exists in which the human perception of similarity between images may not correspond to the content-based similarity.[7] Our goal is to bridge this gap by integrating radiologists' semantic characteristic-based ratings of lung nodules with content-based image features in retrieving similar images from the National Institute of Health (NIH) Lung Image Database Consortium (LIDC). To accomplish this, we use a genetic algorithm (GA) to determine the optimum combination of image features that will retrieve images that are most similar, using diagnosis as a ground truth.

Corresponding author for this publication is Kerry A. Seitz, Jr. at kseitz@trinity.edu

2. RELATED WORKS

Extensive work has been done exploring medical applications of CBIR in recent years, particularly with respect to CAD.[4] Using the LIDC for images of thoracic CT scans, Lam et al.[8] developed an image retrieval system known as BRISC and computed texture features using three methods: Haralick Co-occurrence matrices, Gabor filters, and Markov Random Fields. They determined that Gabor and Markov feature extraction techniques gave the best results, using precision, the ratio of relevant images retrieved to total images retrieved, for evaluation. Additionally, they found that precision increased with the number of radiologists who agreed on the rating of a lung nodule, as was confirmed by Datteri et al.[9] To determine which of the retrieved images were relevant matches, Datteri et al.[10] used an objective evaluation, for which images belonging to the same nodule but in a different slice or outlined by a different radiologist were considered relevant, and a subjective evaluation, in which a relevant image is one that appears in a list of most similar images based on radiologists' semantic annotations. In exploring the relationship between content-based similarity and semantic-based similarity, Jabon et al.[11] found a correlation between the content-based image features for LIDC lung nodules and radiologists' semantic ratings of these nodules, with a combination of 64 image features yielding the highest number of matches. Furthermore, Dasovich et al.[12] developed a linear regression similarity model used for a 116 nodule subset of LIDC images using content features as input and semantic characteristics as output. Their model yielded an R^2 value of 0.871, indicating a high correlation between content-based and semantic-based features for this subset. Machine learning approaches, particularly artificial neural networks (ANN), have also been applied to lung CT scans to evaluate the relationship between semantic and content-based similarity.[5, 13] Kim et al.[13] expanded upon the linear regression prediction model using an ANN model to predict semantic similarity from content-based similarity, with a combination of 64 image features. For random nodule pairs, this ANN model resulted in the highest correlation of 0.129, suggesting that for the combination of image features used, the semantic gap still exists.

CBIR has been used in numerous medical applications beyond the scope of lung CT scans, varying in both organ of interest and imaging modality. Continuing with ANN techniques for CBIR, Muramatsu et al.[14] analyzed mammogram images from the Digital Database for Screening Mammography, with five radiologists providing subjective semantic ratings for the masses. The ANN prediction model resulted in a correlation of 0.798 between content-based and semantic-based similarity, where the radiologists' semantic ratings were used as ground truth. Napel et al.[15] used CBIR with liver CT scans, and content-based and semantic-based features were weighted using a method of machine learning known as adaptive boosting (AdaBoost). Two radiologists evaluated each pair of images and rated similarity on a scale of 1 to 3, and images with an average rated similarity of 2.5 or greater were considered similar. Based on this approach, they determined that combining all 209 features resulted in the highest precision, with an average precision above 90%. Syeda-Mahmood et al.[16] integrated multiple modalities, including ultrasounds, ECG, audio data, diagnosis data, and demographic data, in developing the AALIM system for cardiac decision support, with the goal of retrieving similar patient records. In combining the various modalities, they used weighted linear combinations of similarity values.

Another method of determining the best combination of similarity measures is the use of a genetic algorithm (GA) or genetic programming (GP). Both GA and GP are problem-solving artificial intelligence approaches that apply the biological principles of evolution to a population of individuals, or solutions.[17] Genetic transformations including reproduction, cross-over, and mutation are applied to these individuals in order to improve performance in subsequent generations, and fitness functions are applied to determine the most successful individuals. While GA linearly combines features, GP can employ non-linear representations of individuals, such as trees. In applying GP and GA to CBIR, Torres et al.[18] tested combinations of seven fitness functions with five categories of image features and two similarity measures on the fish shape and MPEG-7 databases to find the combination that would result in the highest precision for similar image retrieval. They determined that the combination of similarity measures found using GP and GA resulted in higher precision after 10 images than the baseline, which did not combine similarity measures. Although results for GA and GP implementations of CBIR for non-medical images are promising,[17–21] GA approaches for medical images and CBIR have not been investigated. In the present study, we have incorporated medical data into GA-based CBIR in determining the combination of image features that would yield the highest precision for similar image retrieval.

3. METHODS

The LIDC database used in this study was released in 2009 and contains 399 CT scans of the lungs. Up to four radiologists analyzed each scan by identifying nodules and rating the malignancy likelihood of each nodule on a scale of 1 to 5 (”Highly Unlikely”, ”Moderately Unlikely”, ”Indeterminate”, ”Moderately Suspicious”, or ”Highly Suspicious”, respectively).[22] To reduce the variability among radiologists, the mode of the radiologists' ratings was used; when a unique mode did not exist, the median rounded down to the nearest whole number was used.[11] Nodules with malignancy ratings of 1 or 2 were considered benign, 4 or 5 were considered malignant, and 3 were considered unknown. From these scans, 932 unique nodules were identified, with most nodules appearing on more than one slice. Previous work reduced the number of slices so that each nodule is represented by only one instance - using the boundaries drawn by the radiologists, the slice containing the largest area for a given nodule was used to represent that nodule.[13] Then 63 image features were extracted based on texture (using Gabor filters, Markov Random Fields, and Haralick Co-occurrence matrices), size, shape, and intensity; these features were normalized across all of the nodules using the Z-Score method.[23] The number of nodules was further reduced to 914 by removing nodules smaller than 5 by 5 pixels because information extracted from these smaller nodules is noisy.

The computer-predicted distribution of semantic characteristics for each nodule was obtained using a CAD algorithm described in previous work,[24] where the DECORATE algorithm was used to construct an ensemble of classifiers based on the 63 image features extracted for each nodule. The 914 dataset was further reduced by eliminating nodules with a malignancy rating of 3. When using computer-predicted malignancy as ground truth, this resulted in a total of 387 nodules, and with radiologist-predicted malignancy, 536 nodules. All nodules have an actual diagnosis, whether or not the radiologists were able to predict it. Thus, trying to match nodules rated as unknown with other nodules with the same rating would be inconsistent because some of these nodules are actually malignant while others are actually benign.

For a given nodule, the 63 image features were placed into a feature vector. The Euclidean distance between the feature vectors for a pair of nodules was used to represent their similarity. To evaluate the effectiveness of the image retrieval system, precision was calculated. The baseline precision measurements were obtained by finding the average precision ($\mathrm{P_{AVG}}$) after 3, 5, 10, 20, and 50 images retrieved using all image features in the feature vector. In applying the GA to the image retrieval process, four different configurations were used for calculating precision and assessing the fitness of individuals in the GA population, varying in whether or not nodules with unknown malignancy were used and whether precision was calculated with computer-predicted malignancy or radiologist-predicted malignancy as ground truth (Table 1).

Table 1. Configurations used to test the CBIR system.

Configuration	Include Unknown Malignancy Nodules[b]	Exclude Unknown Malignancy Nodules[b]
Radiologist-Predicted Malignancy[a]	1	2
Computer-Predicted Malignancy[a]	3	4

[a] Indicates whether the radiologist-predicted or computer-predicted malignancy was used as ground truth.

[b] Indicates whether the nodules that were rated 'unknown' with respect to malignancy were used in the dataset.

The GA was written using a framework called Jiva-ng (http://code.google.com/p/jiva-ng). The mutation rate (rate at which any given trait value is randomly altered between generations) was set to 0.1, and the crossover rate (the rate at which two individuals in the current generation are combined to form a new individual in the next generation) to 0.9. These values were chosen because a high mutation rate could result in the loss of good solutions between generations, while a high crossover rate is more likely to produce better solutions by combining good individuals from the previous generation. A population of 500 was used with 50 generations in order to achieve reasonable computation times but still produce good solutions. Individuals were represented as a Boolean list of 63 values, one for each image feature. For a Boolean value at position i in the list, a value of 'true' indicated using the image feature at i when calculating the similarity between pairs of nodules, while a 'false' indicated not using that image feature. Average precision after 3 ($\mathrm{P_{AVG}}$@3), 5 ($\mathrm{P_{AVG}}$@5), and 10 ($\mathrm{P_{AVG}}$@10) images retrieved were tested as fitness functions. In a GA, the fitness function is used to decide which individuals are the best solutions within each generation. The best solutions are then copied directly to the next generation to preserve them, mutated into slightly different individuals, and combined with each other to produce new solutions in an attempt to find better solutions in the following generations. Running the GA on each of the four configurations resulted in four distinct combinations of image features for each fitness function. To evaluate the effectiveness of each of these combinations, precision after 3, 5, 10, 20, and 50 images retrieved was calculated using the configuration that generated that combination.

4. RESULTS

The baseline precision after 3, 5, 10, 20, and 50 images for each configuration is listed in Table 2. The highest precision was 84.24% and was obtained when the computer-predicted malignancy was used and nodules with unknown malignancy ratings were removed. When the radiologist-predicted malignancy was used, a similar precision (82.77%) was obtained, indicating that there is a relationship between image features and semantic characteristics.

Table 2. Baseline precision results for the CBIR system.

Configuration	$\mathrm{P_{AVG}}$@3[a]	$\mathrm{P_{AVG}}$@5[a]	$\mathrm{P_{AVG}}$@10[a]	$\mathrm{P_{AVG}}$@20[a]	$\mathrm{P_{AVG}}$@50[a]
1	51.57%	51.23%	49.78%	50.26%	48.58%
2	82.77%	82.57%	81.94%	80.90%	79.23%
3	65.43%	65.43%	64.15%	62.78%	59.84%
4	84.24%	83.20%	82.76%	82.07%	79.75%

[a] Average precision after 3, 5, 10, 20, or 50 images retrieved using all 63 image features.

After applying the GA, the best precision (86.91%) was obtained using configuration 4 with the computer-predicted malignancy ratings and removal of the unknown malignancies (Table 3). Configuration 2 with the radiologist-predicted malignancy as ground truth resulted in a precision of 85.95%, only slightly worse than configuration 4. These results were obtained using $\mathrm{P_{AVG}}$@3 as the fitness function.

Table 3. GA precision results for the CBIR system with $\mathrm{P_{AVG}}$@3 as the fitness function.

Configuration	$\mathrm{P_{AVG}}$@3[a]	$\mathrm{P_{AVG}}$@5[a]	$\mathrm{P_{AVG}}$@10[a]	$\mathrm{P_{AVG}}$@20[a]	$\mathrm{P_{AVG}}$@50[a]
1	55.76%	53.70%	52.11%	50.69%	49.12%
2	85.95%	84.10%	83.53%	81.95%	80.37%
3	71.30%	69.74%	67.28%	65.45%	62.39%
4	86.91%	85.32%	84.13%	82.58%	80.90%

[a] Average precision after 3, 5, 10, 20, or 50 images retrieved using a reduced set of image features.

The GA selected 29 image features for both configurations 2 and 4 (Table 4). Of these features, 14 are common to both sets. Within each set, all four categories of image features (texture, size, shape, and intensity) are represented. Furthermore, the 14 common features also include all four of these categories, indicating that texture, size, shape, and intensity are all important for CBIR of lung nodules. Using a pair-wise one-tail t-test with $p<0.05$, we determined that the precisions obtained using the GA showed a significant improvement over the baseline in all cases except after 20 images retrieved with configurations 1 and 4. We also calculated recall for each configuration after 3, 5, 10, 20, and 50 images retrieved (Tables 5 and 6), and a pair-wise one-tail t-test with $p<0.05$ confirmed that the recall improvement with GA-reduced features is statistically significant, except for 20 and 50 images retrieved with configuration 1.

Table 4. GA-selected image features using configurations 2 and 4 with $\mathrm{P}_{\mathrm{AVG}}$@3 as the fitness function.

Configuration 2[c]		Configuration 4[c]	
Correlation	**Area**	Cluster Tendency	Markov 2[b]
Entropy	**Circularity**	Energy	Markov 4[b]
Homogeneity	**Perimeter**	**Sum Average**	**Area**
Inverse Variance	**EquivDiameter**	3rd Order Moment	ConvexArea
Max Probability	**MajorAxisLength**	Variance	**Circularity**
Sum Average	Minor Axis Length	Gabor Mean 0 05[a]	**Perimeter**
Gabor Mean 0 04[a]	MinIntensity	**Gabor SD 45 04**[a]	ConvexPerimeter
Gabor Mean 45 04[a]	**MaxIntensity**	Gabor SD 45 05[a]	Roughness
Gabor SD 45 04[a]	MeanIntensity	Gabor Mean 90 04[a]	**EquivDiameter**
Gabor SD 90 05[a]	SDIntensity	**Gabor SD 90 05**[a]	**MajorAxisLength**
Gabor SD 135 04[a]	MinIntensityBG	Gabor SD 135 03[a]	Compactness
Gabor Mean 135 05[a]	**MaxIntensityBG**	Gabor Mean 135 04[a]	**MaxIntensity**
Gabor SD 135 05[a]	**SDIntensityBG**	**Gabor SD 135 04**[a]	**MaxIntensityBG**
Markov 1[b]	Intensity Difference	**Gabor SD 135 05**[a]	**SDIntensityBG**
Markov 3[b]		**Markov 1**[b]	

[a] For Gabor filter features, the first number represents orientation (0°, 45°, 90°, and 135°) and the second is frequency (0.3, 0.4, 0.5).[23] SD is standard deviation.

[b] For Markov Random Field features, the numbers correspond to orientation (0°, 45°, 90°, and 135°, respectively) or variance, for a total of 5 features.[23]

[c] Bold values are common to both configurations.

Table 5. Baseline recall results for the CBIR system.

Configuration	$\mathrm{R}_{\mathrm{AVG}}$@3[a]	$\mathrm{R}_{\mathrm{AVG}}$@5[a]	$\mathrm{R}_{\mathrm{AVG}}$@10[a]	$\mathrm{R}_{\mathrm{AVG}}$@20[a]	$\mathrm{R}_{\mathrm{AVG}}$@50[a]
1	0.49%	0.82%	1.60%	3.21%	7.69%
2	0.87%	1.45%	2.86%	5.59%	13.48%
3	0.59%	0.98%	1.90%	3.71%	8.70%
4	1.25%	2.07%	4.08%	8.07%	19.39%

[a] Average recall after 3, 5, 10, 20, or 50 images retrieved using all 63 image features.

Table 6. GA recall results for the CBIR system with $\mathrm{P}_{\mathrm{AVG}}$@3 as the fitness function.

Configuration	$\mathrm{R}_{\mathrm{AVG}}$@3[a]	$\mathrm{R}_{\mathrm{AVG}}$@5[a]	$\mathrm{R}_{\mathrm{AVG}}$@10[a]	$\mathrm{R}_{\mathrm{AVG}}$@20[a]	$\mathrm{R}_{\mathrm{AVG}}$@50[a]
1	0.53%	0.85%	1.66%	3.21%	7.72%
2	0.92%	1.48%	2.93%	5.67%	13.74%
3	0.65%	1.06%	2.03%	3.91%	9.22%
4	1.30%	2.12%	4.19%	8.20%	19.84%

[a] Average recall after 3, 5, 10, 20, or 50 images retrieved using a reduced set of image features.

5. CONCLUSION

One of the main goals of CBIR is to bridge the semantic gap between the human perception and the computer perception of similarity. We addressed this issue by implementing a genetic algorithm to find a combination of image features for a CBIR system to retrieve lung nodules that radiologists would consider similar with respect to malignancy. The reduced sets of image features determined by the GA increased the precision for this CBIR system. The best precision value was 86.91% and was obtained using computer-predicted malignancy ratings and removing nodules with unknown malignancy ratings (configuration 4), whereas the baseline precision for this configuration was 84.24%. Furthermore, because texture feature extraction methods are computationally intensive, reducing the number of features will decrease computation time when newly obtained images need to be processed. In addition, calculating the similarity between two images will also be faster with fewer features. In the absence of radiologists' semantic ratings, using computer-predicted malignancy as ground truth is a valid substitute since the precision values are very close (computer-predicted: 86.91% versus radiologist-predicted: 85.95%). In future studies, we will attempt to classify the nodules with unknown malignancy ratings using this CBIR system to retrieve nodules with known malignancy ratings. Successfully classifying these unknown nodules would further validate the effectiveness of our CBIR system and the two sets of image features identified by the GA.

ACKNOWLEDGMENTS

This work was supported in part by NSF award 1062909.

REFERENCES

[1] Siegel, R., Ward, E., Brawley, O., and Jemal, A., "Cancer statistics, 2011: The impact of eliminating socioeconomic and racial disparities on premature cancer deaths," *CA: A Cancer Journal for Clinicians* **61**, 212–236 (2011).

[2] Henschke, C. I., McCauley, D. I., Yankelevitz, D. F., Naidich, D. P., McGuinness, G., Miettinen, O. S., Libby, D. M., Pasmantier, M. W., Koizumi, J., Altorki, N. K., and Smith, J. P., "Early lung cancer action project: overall design and findings from baseline screening," *Lancet.* **354**, 99–105 (1999).

[3] Wormanns, D., Fiebich, M., Saidi, M., Diederich, S., and Heindel, W., "Automatic detection of pulmonary nodules at spiral CT: clinical application of a computer-aided diagnosis system," *European Radiology* **12**, 1052–1057 (2002).

[4] Akgul, C. B., Rubin, D. L., Napel, S., Beaulieu, C. F., Greenspan, H., and Acar, B., "Content-based image retrieval in radiology: current status and future directions," *Journal of Digital Imaging* **24**(2), 208–222 (2011).

[5] Li, Q., Li, F., Shiraishi, J., Katsuragwa, S., Sone, S., and Doi, K., "Investigation of new psychophysical measures for evaluation of similar images on thoracic computed tomography for distinction between benign and malignant nodules," *Medical Physics* **30**, 2584–2593 (2003).

[6] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R., "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(12), 1349–1380 (2000).

[7] Datta, R., Joshi, D., Li, J., and Wang, J., "Image retrieval: ideas, influences, and trends of the new age," *ACM Comput. Surv.* **40**, 5:1–5:60 (2008).

[8] Lam, M. O., Disney, T., Raicu, D. S., Furst, J., and Channin, D. S., "BRISC - an open source pulmonary nodule image retrieval framework," *J Digit Imaging* **20 (Suppl 1)**, 63–71 (2007).

[9] Datteri, R., Raicu, D., and Furst, J., "Local versus global texture analysis for lung nodule image retrieval," *Proc. SPIE* **6919**, 691908–691908–9 (2008).

[10] Datteri, R., Raicu, D., and Furst, J., "Texture model comparison for lung nodules interpretation and retrieval," *The 2008 Annual Meeting of the Society for Image Informatics in Medicine (SIIM 2008)* (2008).

[11] Jabon, S. A., Raicu, D. S., and Furst, J. D., "Content-based versus semantic-based retrieval: a LIDC case study," *Proc. SPIE* **7263**, 72631L–72631L–8 (2009).

[12] Dasovich, G., Kim, R., Raicu, D. S., and Furst, J. D., "A model for the relationship between semantic and content based similarity using LIDC," *Proc. SPIE* **762431**, 762431–762431–10 (2010).

[13] Kim, R., Dasovich, G., Bhaumik, R., Brock, R., Furst, J. D., and Raicu, D. S., "An investigation into the relationship between semantic and content based similarity using LIDC," *Proceedings of the international conference on Multimedia information retrieval* , 185–192 (2010).

[14] Muramatsu, C., Li, Q., Schmidt, R., Shiraishi, J., and Doi, K., "Investigation of psychophysical similarity measures for selection of similar images in the diagnosis of clustered microcalcifications on mammograms," *Medical Physics* **35**, 5695–5702 (2008).

[15] Napel, S. A., Beaulieu, C. F., Rodriguez, C., Cui, J., Xu, J., Gupta, A., Korenblum, D., Greenspan, H., Ma, Y., and Rubin, D. L., "Automated retrieval of CT images of liver lesions on the basis of image similarity: method and preliminary results," *Radiology* **256**, 243–252 (2010).

[16] Syeda-Mahmood, T., Wang, F., Beymer, D., Amir, A., Richmond, M., and Hashmi, S. N., "AALIM: Multimodal mining for cardiac decision support," *Computers in Cardiology* **34**, 209–212 (2007).

[17] Faria, F. F., Veloso, A., Almeida, H. M., Valle, E., Torres, R. d. S., Gonçalves, M. A., and Meira, Jr., W., "Learning to rank for content-based image retrieval," *Proceedings of the international conference on Multimedia information retrieval* , 285–294 (2010).

[18] Torres, R. d. S., Falcão, A. X., Gonçalves, M. A., Papa, J. a. P., Zhang, B., Fan, W., and Fox, E. A., "A genetic programming framework for content-based image retrieval," *Pattern Recogn.* **42**, 283–292 (2009).

[19] Li, P. and Ma, J., "Learning to rank for web image retrieval based on genetic programming," *2nd IEEE International Conference on Broadband Network Multimedia Technology, IC-BNMT '09* , 137–142 (2009).

[20] Huang, H. I., Wu, Y. S., Chan, Y. K., and Lin, C. H., "Study on image feature selection: A genetic algorithm approach," *International Conference on Frontier Computing. Theory, Technologies and Applications* , 169–174 (2010).

[21] Santos, J. A. d., Ferreira, C. D., and Torres, R. d. S., "A genetic programming approach for relevance feedback in region-based image retrieval systems," *Proceedings of the 2008 XXI Brazilian Symposium on Computer Graphics and Image Processing* , 155–162 (2008).

[22] Armato, S. G., McLennan, G., McNitt-Gray, M. F., Meyer, C. R., Yankelevitz, D., Aberle, D. R., Henschke, C. I., Hoffman, E. A., Kazerooni, E. A., MacMahon, H., Reeves, A. P., Croft, B. Y., and Clarke, L. P., "Lung image database consortium: developing a resource for the medical imaging research community," *Radiology* **232**, 739–748 (2004).

[23] Lam, M., Disney, T., Pham, M., Raicu, D., and Furst, J., "Content-based image retrieval for pulmonary computed tomography nodule images," *Proc. SPIE* **6516**, 65160N (2007).

[24] Zinovev, D., Raicu, D., Furst, J., and Armato, S., "Predicting radiological panel opinions using a panel of machine learning classifiers," *Algorithms* **2**, 1473–1502 (2009).