Predicting Malignancy from Mammography Findings and Surgical Biopsies

Pedro Ferreira*, Nuno A. Fonseca*, Inês Dutra*†, Ryan Woods‡ and Elizabeth Burnside§
*CRACS-INESC Porto LA, Porto, Portugal
†CRACS-INESC Porto LA & DCC-FC, Universidade do Porto, Porto, Portugal
‡Department of Radiology, Johns Hopkins Hospital, Baltimore, MD, USA
§University of Wisconsin, Medical School, Madison, WI, USA

***Abstract*—Breast screening is the regular examination of a woman's breasts to find breast cancer earlier. The sole exam approved for this purpose is mammography. Usually, findings are annotated through the Breast Imaging Reporting and Data System (BIRADS) created by the American College of Radiology. The BIRADS system determines a standard lexicon to be used by radiologists when studying each finding. Although the lexicon is standard, the annotation accuracy of the findings depends on the experience of the radiologist. Moreover, the accuracy of the classification of a mammography is also highly dependent on the expertise of the radiologist. A correct classification is paramount due to economical and humanitarian reasons.**

The main goal of this work is to produce machine learning models that predict the outcome of a mammography from a reduced set of annotated mammography findings. In the study we used a data set consisting of 348 consecutive breast masses that underwent image guided or surgical biopsy performed between October 2005 and December 2007 on 328 female subjects. The main conclusions are threefold: (1) automatic classification of a mammography, independent on information about mass density, can reach equal or better results than the classification performed by a physician; (2) mass density seems to be a good indicator of malignancy, as previous studies suggested; (3) a machine learning model can predict mass density with a quality as good as the specialist blind to biopsy, which is one of our main contributions. Our model can predict malignancy in the absence of the mass density attribute, since we can fill up this attribute using our mass density predictor.

***Keywords*-machine learning; mammography; BIRADS;**

I. Introduction

Mammography is considered the cheapest and most efficient method to detect cancer in a preclinical stage and breast screening programs were created precisely with the objective of detecting cancer in earlier stages. The breast screening programs usually generate a huge amount of data, annotated according to the Breast Imaging Reporting and Data System (BIRADS) created by the American College of Radiology. The BIRADS system determines a standard lexicon to be used by radiologists when studying each finding. Although the breast screening programs have helped reducing the number of women with undetected cancer, there is still room for improvement, since recent statistics show that one woman dies of breast cancer every 13 minutes in the U.S. and in 2009, an estimated 40,170 women (15% of all deaths) and 440 men in the U.S. were expected to die from breast cancer. Therefore it is of utmost importance to improve these numbers and raise the life expectancy in the next years.

We applied machine learning methods to 348 consecutive breast masses that underwent image guided or surgical biopsy performed between October 2005 and December 2007 on 328 female subjects. These 348 findings are defined by 14 attributes, with one of them indicating if the finding is malignant or benign. Our main objective is to produce models that can have a good performance at predicting malignancy and a good performance at avoiding to expose healthy women to extra surgical or screening procedures. We are also interested in studying the actual relevance of mass density in the findings, since this is one of the attributes that usually is not regarded relevant by physicians. According to physicians, mass density is a feature usually considered to be difficult to annotate, because of the breast tissue, and fat composition. Previous works have shown that mass density can be an important attribute when predicting malignancy [1], [2], [3]. The 348 mammographies used in this study have annotations of mass density, which allow to (1) investigate in more detail the role played by this feature, and (2) produce models to predict this particular feature and help physicians distinguish between high and iso/low densities.

Much work has been done on applying machine learning techniques to study breast cancer, one of the most common kinds of cancer in the world. In the UCI (University of California, Irvine) machine learning repository[1] there are four data sets whose main target of study is breast cancer. One of the first works on applying machine learning techniques to breast cancer data dates from 1990. At this time, the first data set donated to the UCI repository was created by Wolberg and Mangasarian after their work on a multi-surface method of pattern separation for medical diagnosis applied to breast cytology [4]. Most works in the literature applies artificial neural networks to the problem of diagnosing breast cancer (e.g., [5] and [6]). Others focus on prognosis of the disease using inductive learning methods (e.g., [7]). More recently, Ayer *et al.* [8] have evaluated whether an artificial neural network trained on a large prospectively collected data set of consecutive mammography findings could discriminate

[1] http://archive.ics.uci.edu/ml/datasets.html.

between benign and malignant disease and accurately predict the probability of breast cancer for individual patients. Other works concentrate on the correlation of attributes in the mammographies, for example, the influence of mass density and other features on predicting malignancy [1], [2], [3], [9], [10], [11], [12]. Other recent works focus on extracting information from free text that appears in medical records of mammography screenings [13], and on the influence of age in ductal carcinoma *in situ* (DCIS) findings [14]. Yet other works focus on the mammography images themselves [15], [16]. These are orthogonal to the above mentioned and to our own work, whose focus is on the medical reports.

We use the same data set used by Woods and Burnside [2]. This data set is unique in the sense that all findings were retrospectively assessed and all of them have accurate information about the density of the breast masses. In that work, they showed that high breast mass density is a significant predictor of malignancy, even after controlling for other well-known predictors of malignancy such as mass margin and mass shape. The metric used to evaluate performance was interobserver agreement and they found a moderate k-value for mass density (0.53).

II. Breast Cancer Data

Our study analyzes 348 consecutive breast masses that underwent image guided or surgical biopsy performed between October 2005 and December 2007 on 328 female subjects. Each one of the 348 cases refers to a breast nodule retrospectively classified according to the BIRADS system. On the other hand, a clinical radiologist assessed (at the time of imaging and without biopsy results) the density of 180 of these masses, in an evaluation that can be considered as "performed under stress" (prospective assessment). Pathology result at biopsy was the study endpoint.

Table I shows the main attributes used from these data to learn the models. When learning models to predict malignancy the attribute outcome is the target class. It assumes values malignant and benign and was determined using the results of biopsies. From the 348 cases, 118 are malignant ($\approx$ 34%), and 84 cases have high mass density ($\approx$ 24%) retrospectively assessed Other attributes are mass shape, mass margins, depth, size, among others. For the purpose of our study, we have two attributes that represent the same characteristics of the finding, but with different interpretations. These are retro_density and density_num. Both represent mass densities that can assume values $high$ or iso/low. Retro_density was retrospectively assessed while density_num was prospectively (at the time of imaging) assessed. These two attributes are our target classes when learning models to predict mass density.

III. Methodology

The whole data set (348 findings) was split into two subsets: (1) *training set*: 180 cases, whose mass densities were classified by a radiologist at the exact time of imaging and (2) *test set*: 168 cases, whose mass densities were not annotated at the time of imaging, but instead in a reassessment of all the 348 exams performed by a group of experienced physicians. The attribute corresponding to the prediction of mass density by the specialist is `density_num`. The attribute corresponding to the retrospectively assessed mass density is `retro_density`. We have values for `density_num` for only 180 of the cases, and have values for `retro_density` for all 348 cases. With these train and test datasets, we performed several experiments in order to generate models to (1) predict malignancy (outcome), and (2) to predict mass density.

Table I
Data Attributes.

Attribute	Description
age_at_mammo	Age of the patient when the mammogram was taken
clockface_location	Location of the mass
mass_shape	Shape of the mass
mass_margins	Classification of the margins of the mass
side	Breast where the mass was found (left or right)
depth	Depth of the mass according to a measure from the skin surface to the center of the lesion
mass_margins_worst	Most worrisome mass margin descriptor
quadrant_location_def	Quadrant location of the mass
size	Greatest transverse width of the mass (in mm)
breast_composition	Composition of the breast (e.g., almost entirely fat, scattered fibroglandular densities, heterogeneously dense, extremely dense)
retro_density	Retrospective annotation of mass density
density_num	Prospective annotation of mass density
outcome	Classification of the mass based on the results of the biopsy (malignant or benign)

Table II shows all experiments performed for each task, according to the attributes used to learn mass density or outcome. The first five experiments were performed with 180 findings (training set) while the remaining were performed with 168 findings (test set). From the first five, the first three predict outcome and the other two predict mass density. In a nutshell, the experiments can be described as follows:

- Experiment E_1 aims at finding a classifier to predict outcome using the attribute mass density that was retrospectively annotated (`retro_density`). This classifier would be useful to help physicians make decisions on retrospectively studied patients.
- Experiment E_2 aims at finding a classifier to predict outcome from patients whose mass density was prospectively assessed (using the attribute

Table II
EXPERIMENTS ON THE TRAINING AND TEST SETS. IN EACH LINE, WE GIVE THE CONDITIONS OF THE EXPERIMENT. E.G., E_1, E_2 AND E_3 PREDICT OUTCOME, WHERE E_1 USES MASS DENSITY AS DESCRIBED BY THE ATTRIBUTE RETRO_DENSITY, E_2 USES MASS DENSITY AS DESCRIBED BY THE ATTRIBUTE DENSITY_NUM, AND E_3 DOES NOT USE ANY INFORMATION ABOUT MASS DENSITY

Exp.	outcome	retro_density	density_num	size	output
E_1	class	yes	no	180	classifier for outcome (M_1)
E_2	class	no	yes	180	classifier for outcome (M_2)
E_3	class	no	no	180	classifier for outcome (M_3)
E_4	no	class	no	180	classifier for mass density (M_4)
E_5	no	no	class	180	classifier for mass density (M_5)
E_6	no	class	no	168	test set with mass density filled up by model M_4
E_7	no	no	class	168	test set with mass density filled up by model M_5
E_8	class	yes	no	168	prediction of outcome using actual values of retro_density
E_9	class	yes (E_6)	no	168	prediction of outcome using test set obtained in E_6
E_{10}	class	no	yes (E_7)	168	prediction of outcome using test set obtained in E_7
E_{11}	class	no	no	168	prediction of outcome without mass density

`density_num`). This classifier would be helpful on the clinical daily routine of a physician.

- Experiment E_3 was performed in order to assess the performance of a classifier trained without any mass density information. This experiment was performed in order to assess the relevance of mass density when predicting the outcome. It can be used on new data without any information about the mass density.
- Experiment E_4 generates models to predict mass density based on retrospectively annotated density.
- Experiment E_5 generates models to predict mass density based on prospectively annotated density.

The last two experiments were performed to assess how well an automated classifier can predict the kinds of densities (high or iso/low) when compared to the physician.

We evaluated several classification algorithms available in WEKA [17] and varied their parameters. The experiments were performed with the WEKA's Experimenter module using 10 times 10-fold cross-validation on the training dataset. For each algorithm we selected the combination of parameters that produced the best classifiers, and then selected the top three classifiers for generating models: NaiveBayes [18], DTNB (a decision table algorithm whose leaves are Bayesian networks) and SMO (a support vector machine [19] implementation [20]). A fourth classifier was selected, J48 (decision tree based on Quinlan's C4.5 algorithm), due to its ability to produce readable and easily understandable models.

The last six experiments of Table II apply the models generated (M_1, M_2, M_3, M_4, and M_5 generated by the first five experiments), to the test set containing 168 cases as follows:

1) Experiment E_6 generates the values for mass density using the model trained with the attribute `retro_density` as the class variable (obtained by experiment E_4).
2) Experiment E_7 generates the values for mass density using the model trained with the attribute `density_num` as the class variable, (obtained by experiment E_5).
3) Experiment E_8 predicts outcome using the model M_1 trained with the attribute `retro_density` (obtained by experiment E_1), and uses the actual values of the attribute `retro_density` available in the test set.
4) Experiment E_9 predicts outcome using the model M_1 trained with the attribute `retro_density` (obtained by experiment E_1), and uses the mass density values filled up by experiment E_6 in the test set.
5) Experiment E_{10} predicts outcome using the model trained M_2 with the attribute `density_num` (obtained by experiment E_2), and uses the mass density values filled up by experiment E_7 in the test set.
6) Experiment E_{11} predicts outcome with the model M_3 that does not use any information about mass density, obtained in experiment E_3. For this experiment, no mass density attribute is used in the test set.

We used the metrics CCI (Correctly Classified Instances, a.k.a. accuracy), F-measure (harmonic mean between Precision and Recall) and Kappa statistics to assess the classifiers. Whenever applicable we performed significance tests using paired t-test ($\alpha = 0.05$).

IV. RESULTS

We first investigated the data and calculated simple frequencies to determine if there was some evidence of relationship between attributes, specially if mass density is related to malignancy.

According to the frequencies of attribute values among the classes, from the 348 breast masses, 118 are malignant ($\approx$ 34%), and 84 have high mass density ($\approx$ 24%). If we consider that mass density and malignancy are independent, and take 84 cases from the 348 at random, the probability of these being malignant should still be $\approx$ 34%. However, if it happens that all 84 cases selected at random have high density, then the percentage of malignant cases raises to 70.2% and the probability of this being coincidence is very

Table III
CLASSIFIERS' PERFORMANCE FOR EACH TASK. VALUES NOT IN BOLD ARE STATISTICALLY SIGNIFICANTLY WORSE THAN THE CLASSIFIER WITH HIGHEST ACCURACY (USING PAIRED T-TEST WITH $\alpha = 0.05$).

Exp.	Algorithm	CCI	K	F	AUROC
E1	SMO	**85.6**$\pm$7.3	**0.69**$\pm$0.16	**0.80**$\pm$0.11	**0.84**$\pm$0.08
E1	DTNB	81.6$\pm$8.2	0.60$\pm$0.18	0.74$\pm$0.13	0.88$\pm$0.07
E1	NaiveBayes	81.3$\pm$9.5	0.61$\pm$0.20	0.76$\pm$0.12	0.88$\pm$0.08
E1	J48	80.7$\pm$9.3	0.59$\pm$0.20	0.75$\pm$0.13	0.79$\pm$0.11
E2	SMO	**83.9**$\pm$7.7	**0.66**$\pm$0.17	**0.78**$\pm$0.11	**0.82**$\pm$0.08
E2	NaiveBayes	80.3$\pm$9.3	0.59$\pm$0.19	0.75$\pm$0.12	0.87$\pm$0.09
E2	DTNB	79.8$\pm$9.5	0.56$\pm$0.21	0.72$\pm$0.15	0.86$\pm$0.09
E2	J48	75.4$\pm$9.5	0.47$\pm$0.21	0.65$\pm$0.15	0.73$\pm$0.12
E3	SMO	**83.8**$\pm$7.7	**0.65**$\pm$0.17	**0.78**$\pm$0.11	**0.82**$\pm$0.09
E3	J48	76.3$\pm$9.9	0.49$\pm$0.22	0.67$\pm$0.15	0.76$\pm$0.13
E3	NaiveBayes	76.2$\pm$9.9	0.51$\pm$0.20	0.71$\pm$0.13	0.85$\pm$0.09
E3	DTNB	75.7$\pm$9.0	0.48$\pm$0.19	0.67$\pm$0.13	**0.81**$\pm$0.10
E4	SMO	**81.3**$\pm$8.2	**0.52**$\pm$0.21	**0.64**$\pm$0.17	**0.75**$\pm$0.11
E4	J48	74.4$\pm$8.8	0.32$\pm$0.24	0.47$\pm$0.21	0.67$\pm$0.15
E4	DTNB	73.5$\pm$10.0	0.34$\pm$0.24	0.51$\pm$0.19	**0.76**$\pm$0.12
E4	NaiveBayes	72.8$\pm$9.9	0.37$\pm$0.23	0.56$\pm$0.18	0.77$\pm$0.11
E5	NaiveBayes	**67.2**$\pm$12.1	**0.33**$\pm$0.25	**0.62**$\pm$0.15	**0.72**$\pm$0.14
E5	SMO	**66.8**$\pm$10.7	**0.31**$\pm$0.22	0.55$\pm$0.16	0.65$\pm$0.11
E5	J48	63.6$\pm$10.1	0.26$\pm$0.21	0.56$\pm$0.15	0.62$\pm$0.13
E5	DTNB	62.1$\pm$11.9	0.22$\pm$0.24	0.54$\pm$0.16	0.64$\pm$0.14

Table IV
CLASSIFIERS' PERFORMANCE FOR THE TEST SET.

	Algorithm	CCI	K	F	AUROC
E_6	SMO	84.52	0.46	0.91	0.74
E_7	NaiveBayes	75.60	0.35	0.84	0.81
E_8	SMO	80.95	0.50	0.87	0.74
E_9	SMO	77.98	0.45	0.85	0.80
E_{10}	SMO	79.17	0.49	0.85	0.83
E_{11}	SMO	76.19	0.42	0.83	0.71

low. This simple calculation may already imply that high density has some relationship with malignancy. So may other attributes such as age, mass shape and mass margins. In this work, we do not report on the importance of the other attributes.

A. Performance Analysis

The best models produced for experiments (E_1), (E_2), (E_3) and (E_4) were obtained with the algorithm SMO, with main parameters: polynomial kernel with exponent $E = 1$ and complexity constant $C = 0.05$. For experiment (E_4), the best classifier was obtained with no data normalization/standardization ($N = 2$), while the other 3 experiments used $N = 1$ (the training data was standardized). The parameter C at SMO controls how soft the class margins are. In practice it controls how many instances are used as 'support vectors' to draw the linear separation boundary in the transformed Euclidean feature space. The fact that $C = 0.05$ produces better results seems to indicate that the default value (1.0) somehow generates an over-fitted trained classifier, whose performance is not so good on the cross-validation test sets. For experiment (E_5), the best classifier was obtained using the naive Bayes algorithm with default parameters. Most probably, naive Bayes performed better with this data set because this data is noisy containing errors associated to the prospectively annotated density_num attribute.

Table III shows, for each experiment E_1 to E_5, the best performance of each algorithm after parameter variation (classifiers are sorted in descending order after CCI). The SMO classifier consistently achieves better results for the training data set, even when NaiveBayes wins (experiment E_5, note that there is no statistically significant difference between NaiveBayes and SMO with respect to CCI and K).

All classifiers behave better when trained on retrospectively annotated data (experiment E_1), which seems to indicate that in practical clinical routine, this would be the best classifier to use. However, since it is hard to obtain retrospectively annotated data, the approach followed in E_2, using prospectively annotated mass density values, can also be used with good results. It is important to notice that the SMO obtained with experiment E_2 has performance only slightly lower than the SMO of experiment E_1 and the difference is not statistically significant.

Experiment E_5 is the most difficult as it consists of predicting mass density from noisy data. It is interesting to note that all algorithms achieve lower performance for this experiment than for the other tasks, with NaiveBayes achieving a performance that is close to that of the physician, who has CCI of 70% when compared with the retrospectively annotated mass density.

All results of Table III, with exception of AUROC, are higher for the best classifier. The AUROC is higher for algorithms other than the best.

B. Training to predict outcome

In the three experiments, (E_1), (E_2) and (E_3), the best classifiers found were based on SMO. First of all, these results show that mass density has some influence on the outcome, specially when mass density is the one observed on the retrospective data (experiment E_1). The classifier trained without mass density has an overall performance of 83.8% while the classifier trained with the retrospectively assessed mass has an overall performance of 85.6%. If we look at the K value, we can confirm that the relation between mass density and outcome is not by chance, given the relatively high observed agreement between the real data and the classifier's predicted values. The F-measure balances the values of Precision and Recall and also indicates that the classifiers are behaving reasonably well.

The results obtained with experiments (E_1), (E_2) and (E_3) confirm findings in the literature regarding the relevance of mass density [1], [2], [12], [21], and also show that good classifiers can be obtained to predict outcome (with a high percentage of correctly classified instances and good values of precision and recall, according to F).

Another evidence that mass density is somehow related to malignancy are the decision trees generated by the J48 algorithm, in which `retro_density` and `density_num`

were chosen as the most important attributes appearing in the top of the trees. Despite the fact that J48 was not the best classifier to predict outcome, this fact reveals that the attribute mass density has some influence over all the remaining features. Another important fact to note is that, according to J48, the second most important attribute that helps discriminating between malignant and benign cases is mass margins.

C. Training to predict mass density

Our set of experiments E_4 and E_5 are related to predicting mass density. As the data set has two annotated mass densities, one for the prospective study and another one for the retrospective, we generated two classifiers: one is trained on the prospective values of mass density (density_num), and another one is trained on the retrospective (retro_density) values of mass density. Once more, we used the 180 cases as training set and 10-fold cross-validation. The best classifier for predicting retro_density was SMO and the best to predict density_num was NaiveBayes.

During the prospective study, the radiologist predicted 70% of masses on the 180 findings compared with the annotated masses of the retrospective study. The SMO classifier predicted 81.3% of correct instances when training on the retrospective annotated mass (retro_density) and NaiveBayes predicted 67.2% of correct instances when training on prospective masses annotated by the radiologist. These results are quite good and indicate that either the SMO or the Bayesian classifier generated in this study can be well applied as a support tool to help physicians/radiologists to classify mass density in mammograms.

The values of K and F-measure for this experiment are not so good as the ones obtained with the classifiers that predict outcome. The K value, once more, indicates that both NaiveBayes and SMO have a moderate level of agreement.

D. Performance of the best classifiers on unseen data

Table IV summarizes the results of predicting outcome on the 168 unseen cases as well as the results of filling up the attribute mass density in the test set.

The first two lines of Table IV refer to experiments to fill up values of the attribute mass density in the test set. The CCI indicates how well models M_4 and M_5, obtained respectively with experiments E_4 and E_5, performed on filling up those values, when compared with the actual values of retro_density available in the test set. The SMO classifier, which had a very good performance on the training set (CCI=81.3%), behaves even better when filling up values for retro_density, making mistakes in only 16% of the actual masses. The NaiveBayes classifier (M_5), obtained with experiment E_5, which had CCI=67.2% in the training set, performed very well in the task of filling up the missing values of density_num, correctly classifying 75.6% of the instances. A result that surpasses the result obtained by the specialist, which is 70%.

For the tasks of predicting outcome, the classifiers also perform very well, with the worst predictions being produced by model M_3, which does not use any information about mass density. This result confirms once more the relevance of mass density on predicting outcome. In the absence of this information, the data could be filled up by M_4 or M_5, that, as mentioned, have a good performance on performing this job.

E. MammoClass Application

The best models were integrated into an online application (called MamoClass). It allows a practitioner to quickly and easily assess mammograms by obtaining a prediction for mass density and/or classify a mammography given a reduced set of mammography findings. The application is freely available at http://cracs.fc.up.pt/mammoclass. This application will start to be used at Hospital São João in Porto, Portugal, and at the Medical School, in the University of Wisconsin, Madison, by our collaborators.

V. Conclusions and Future Work

In this work, we were provided with 348 cases of patients that went through mammography screening and biopsies. The objective of this work was twofold: i) find non trivial relations among attributes by applying machine learning techniques to these data, and; ii) learn models that could help medical doctors to quickly assess mammograms.

The conclusions are threefold: (1) automatic classification of a mammography, independent on information about mass density, can reach equal or better results than the classification performed by a physician; (2) mass density seems to be a good indicator of malignancy, as previous studies suggested; (3) machine learning classifiers can predict mass density with a quality as good as the specialist blind to biopsy, which is one of our main contributions. Our classifier can predict malignancy in the absence of the mass density attribute, since we can fill up this attribute using our mass density predictor.

As future work, we plan to extend this work to larger data sets, and apply other machine learning techniques based on statistical relational learning, since classifiers that fall in this category provide a good explanation of the predicted outcomes as well as can consider the relationship among mammograms of the same patient. We would also like to investigate how other attributes can affect malignancy or are related to the other attributes. Yet another stream would be to study why the parameter variation in the WEKA algorithms has a strong impact on the performance of the classifiers. Another important step forward would be to investigate with the physician, why some instances are consistently misclassified by all algorithms.

ACKNOWLEDGMENTS

This work has been partially supported by the projects HORUS (PTDC/EIA-EIA/100897/2008) and DigiScope (PTDC/EIA-CCO/100844/2008) and by the Fundação para a Ciência e Tecnologia (FCT/Portugal).

REFERENCES

[1] R. Woods, L. Oliphant, K. Shinki, D. Page, J. Shavlik, and E. Burnside, "Validation of results from knowledge discovery: Mass density as a predictor of breast cancer," *J Digit Imaging*, pp. 418–419, 2009.

[2] R. W. Woods, G. S. Sisney, L. R. Salkowski, K. Shinki, Y. Lin, and E. S. Burnside, "The mammographic density of a mass is a significant predictor of breast cancer," *Radiology*, 2010. [Online]. Available: http://radiology.rsna.org/content/early/2010/12/18/radiol.10100328.abstract

[3] P. Ferreira, I. Dutra, N. A. Fonseca, R. Woods, and E. Burnside, "Studying the relevance of breast imaging features," in *Proc. of the international Conference on Health Informatics (HealthInf)*, Jan 2011.

[4] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," in *Proceedings of the National Academy of Sciences, 87*, 1990, pp. 9193–9196.

[5] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer," *Radiology*, vol. 187, pp. 81–87, April 1993.

[6] H. A. Abbass, "An evolutionary artificial neural networks approach for breast cancer diagnosis," *Artificial Intelligence in Medicine*, vol. 25, p. 265, 2002.

[7] W. N. Street, O. L. Mangasarian, and W. H. Wolberg, "An inductive learning approach to prognostic prediction," in *ICML*, 1995, p. 522.

[8] T. Ayer, O. Alagoz, J. Chhatwal, J. W. Shavlik, C. E. J. Kahn, and E. S. Burnside, "Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration," *Cancer*, vol. 116, no. 14, pp. 3310–3321, 2010.

[9] V. P. Jackson, K. A. Dines, L. W. Bassett, R. H. Gold, and H. E. Reynolds, "Diagnostic importance of the radiographic density of noncalcified breast masses: analysis of 91 lesions," *AJR Am J Roentgenol*, vol. 157, pp. 25–28, 1991.

[10] E. A. Sickles, "Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases," *Radiology*, vol. 179, pp. 463–468, 1991.

[11] R. C. Cory and S. S. Linden, "The mammographic density of breast cancer," *AJR Am J Roentgenol*, vol. 160, pp. 418–419, 1993.

[12] J. Davis, E. S. Burnside, I. C. Dutra, D. Page, and V. S. Costa, "Knowledge discovery from structured mammography reports using inductive logic programming," in *American Medical Informatics Association 2005 Annual Symposium*, 2005, pp. 86–100.

[13] H. Nassif, R. Woods, E. Burnside, M. Ayvaci, J. Shavlik, and D. Page, "Information extraction for clinical data mining: A mammography case study," in *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, ser. ICDMW '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 37–42. [Online]. Available: http://dx.doi.org/10.1109/ICDMW.2009.63

[14] H. Nassif, D. Page, M. Ayvaci, J. Shavlik, and E. S. Burnside, "Uncovering age-specific invasive and dcis breast cancer rules using inductive logic programming," in *Proceedings of the 1st ACM International Health Informatics Symposium*, ser. IHI '10. New York, NY, USA: ACM, 2010, pp. 76–82. [Online]. Available: http://doi.acm.org/10.1145/1882992.1883005

[15] M. Samulski and N. Karssemeijer, "Optimizing case-based detection performance in a multiview cad system for mammography," *Medical Imaging, IEEE Transactions on*, vol. 30, no. 4, pp. 1001 –1009, april 2011.

[16] J. Lesniak, R. Hupse, M. Kallenberg, M. Samulski, R. Blanc, N. Karssemeijer, and G. Székely, "Computer aided detection of breast masses in mammography using support vector machine classification," in *Proc. SPIE 7963*, ser. SPIE 2011, 2011.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, November 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[18] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann, 1995, pp. 338–345.

[19] L. Wang, *Support Vector Machines: Theory and Application*. Springer, 2005.

[20] J. Platt, "Machines using sequential minimal optimization," in *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT Press, 1998.

[21] J. Davis, D. Page, E. Burnside, I. Dutra, R. Ramakrishnan, J. Shavlik, and V. Santos Costa, *Introduction to Statistical Relational Learning*. MIT Press, 2007, ch. Learning a New View of a Database: With an Application in Mammography.