Interpretable Models to Predict Breast Cancer

Pedro Ferreira*, Inês Dutra*†, Rogerio Salvini‖, Elizabeth Burnside¶
*CRACS-INESC TEC, Porto, Portugal
†DCC-FC, Universidade do Porto, Porto, Portugal
¶University of Wisconsin, Medical School, Madison, WI, USA
‖Federal University of Goiás, Goiás, Brazil

***Abstract*—Several works in the literature use propositional ("black box") approaches to generate prediction models. In this work we employ the Inductive Logic Programming technique, whose prediction model is based on first order rules, to the domain of breast cancer. These rules have the advantage of being interpretable and convenient to be used as a common language between the computer scientists and the medical experts. We also explore the relevance of some of variables usually collected to predict breast cancer. We compare our results with a propositional classifier that was considered best for the same dataset studied in this paper.**

***Index Terms*—inductive logic programming; machine learning; mammography; BI-RADS**

I. Introduction

Breast cancer is a disease in which malignant cells form in the tissues of the breast. Once it is detected, depending on its gravity, and suitable treatment, most women can continue a normal life. However, despite screening programs for early detection of tumors, breast cancer is still the second leading cause of cancer death among women, according to the American National Breast Cancer Organization [1]. Several works in the literature search for a solution to this problem trying to better characterize patterns of breast cancer both in the medical area and in computer-based systems. Their aim is to anticipate patterns of the disease in an early stage in order not to miss false negatives. In that area, computer-based systems can aid, specially when helping to build prediction models for cases hard to discriminate or for early screening. Much work has been done on applying machine learning techniques to the area of breast cancer. In the UCI (University of California, Irvine) machine learning repository[1], there are four datasets 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 dataset 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 [2]. Most works in the literature applies artificial neural networks to the problem of diagnosing breast cancer (e.g., [3] and [4]). Others focus on prognosis of the disease using inductive learning methods (e.g., [5]). Ayer *et al.* [6]) have evaluated whether an artificial neural network trained on a large prospectively collected dataset of consecutive mammography findings could discriminate between benign and malignant disease and accurately predict the probability of breast cancer for individual patients. Recently, a new method has been developed for automatically explaining prediction/classification results for any machine learning model without degrading accuracy [7], [8]. Other works concentrate on the correlation of attributes in the mammograms, for example, the influence of mass density and other features on predicting malignancy [9], [10], [11], [12], [13], [14], [15]. Other works focus on extracting information from free text that appears in medical records of mammography screenings [16], [17], and on the influence of age in ductal carcinoma *in situ* (DCIS) findings [18]. Yet other works focus on the mammography images themselves [19], [20]. The latter are orthogonal to the above mentioned and to our own work, whose focus is on the medical reports.

In a previous work, we generated models to predict malignancy from a small set of variables annotated from mammography images [21]. A tool, MammoClass[2] was built and is publicly available to be used and tested. MammoClass was trained using a set of variables collected from mammography reports and other information related to the patient history, biopsies, among others. A subset of these variables (Side, Depth, Clockface and Quadrant) are considered to be non indicative of malignancy by expert radiologists, but studies show that for some populations there can be a prevalence of breast cancer according to the value of some of these variables. For example, the GEC-ESTRO Handbook of Brachytherapy (Section Breast Cancer [22]) says that the upper outer quadrant is the most common site of origin of breast cancer. It also says that breast cancer is more common in the left than in the right breast. Other studies on laterality also confirm this tendency [23]. In this work, we study the relevance of these variables on our dataset and also investigate alternatives for the SVM model used by MammoClass by training our data using Inductive Logic Programming (ILP) [24] and generating more interpretable models based on first-order logic, which, by its turn, is more expressive and compact than decision trees, another interpretable model. Both models (SVM and ILP) agree that Side, Depth, Clockface and Quadrant are not relevant to predict breast cancer on our dataset. However, the quantitative performance of ILP is below the SVM's. We then performed several experiments with ILP and concluded that it is possible to reach performances as good as the SVM's with

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

[2]http://cracs.fc.up.pt/mammoclass

the benefit of having an interpretable model based on first order logic. Another fact is that ILP presents better results than decision trees.

In the next section, we present our experimental methodology. In Section III, we present and discuss performance results for all experiments. Finally, we present our conclusions and perspectives of future work.

II. Methodology

We used the same dataset and settings used in [21]: 348 consecutive breast masses that underwent image guided core biopsies 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 BI-RADS system. The whole dataset (348 findings) was split into two subsets: (1) *training set*: 180 cases (71+/109-), and (2) *test set*: 168 cases (47+/121-). This split is the same used in our previous settings [21]. Test set is independent from the training set.

Experiments were performed with ILP system Aleph [25] and with the WEKA toolkit [26]. The Aleph system was developed to be a prototype to explore ideas in ILP and was written in Prolog. Aleph has a powerful representation language that allows to represent complex expressions and incorporate new background knowledge easily. Aleph also let choose the order of generation of the rules, change the evaluation function and the search order. Allied to all these characteristics the Aleph system is open source making it a powerful resource to all ILP researchers [27]. WEKA is a collection of machine learning algorithms for data mining tasks. It was written in Java and developed at the University of Waikato, New Zealand. It is a free software. For WEKA we used the best classifier obtained in our previous work [21] for predicting malignancy. This classifier was obtained with the SMO (a support vector machine [28] implementation [29]) and is used by MammoClass. MammoClass uses a small set of variables to predict a probability of a finding being malignant or benign. These are the patient's age, mass size, breast composition, mass shape, mass clockface location, mass margins, mass density, side, quadrant and depth. From these, clockface location, side, quadrant and depth are considered to be not important by expert radiologists and their suggestion was to remove them from the tool. Our first hypothesis is: can we remove these variables and still obtain the same results with the test set? Our second hypothesis is: can we produce more interpretable classifiers that can have as good or better performance than the SVM?

In order to answer the first question, we trained our SVM on the 180 examples removing the four variables: side, depth, clockface and quadrant, and compared the performance on the 168 test cases with the performance we had obtained in [21], when training with all variables.

In order to obtain more interpretable classifiers, we trained the 180 examples using Aleph, with (1) all variables and (2) removing the subset of four variables aforementioned. We also further explored Aleph by varying its parameters with the goal of searching different portions of the search space looking for better and more meaningful hypotheses.

These experiments are described as follows:

- Experiment A trains the SVM classifier on the 180 training cases, without the four variables, and evaluates its performance on the 168 test set. Results from this experiment are compared with results published in [21], when the same classifier was trained on the same 180 cases and tested on the same 168 cases, using all variables. This is called "$Prev$" in the results section.
- Experiment $B1$ trains Aleph on the 180 training cases, with all variables, and evaluates the performance of the classifier on the 168 test cases.
- Experiment $B2$ trains Aleph on the 180 training cases, without the four variables, and evaluates the performance of the classifier on the 168 test cases.
- Experiment C searches for Aleph classifiers that can be better than the SVM. For this experiment, we varied Aleph's internal parameters: noise and evalfn. Noise controls the maximum number of false positives allowed by the model during training. Evalfn controls the evaluation function used to assess the quality of each hypothesis generated. Noise variation goes from zero to one hundred (0-100). This range was defined according to the number of cases in the training set. The evaluation function is set as: *coverage*, *mestimate*, *cost*, *entropy*, *gini* and *wracc*. *Coverage's* clause utility is *P-N*, where *P* and *N* are the numbers of positive and negative examples covered by a clause, respectively. *Mestimate's* clause utility is described in detail in Dzeroski and Bratko's paper [30]. The value of *m* is set by *set(m,M)*. *Cost* depends on a user defined function. In our setting the cost of each rule is given by the number of correct positive examples covered (trying to maximize Recall). *Entropy's* clause utility is $plogp + (1-p)log(1-p)$ where $p = \frac{P}{P+N}$ and *P*, *N* are the numbers of positive and negative examples covered by the clause. *Gini's* clause utility is $2p(1-p)$, where $p = \frac{P}{P+N}$ and *P*, *N* are the numbers of positive and negative examples covered by the clause. *Wracc's* clause utility is calculated using the weighted relative accuracy function described in Lavrac *et al.* article [31]. The objective of this experiment is to compare Aleph's performance (number of true positives and accuracy) to WEKA's performance by varying Aleph's internal parameters.

Statistical significance tests were applied to compare the models, using GraphPad's McNemar's test, since we use binary outcomes (malignant or benign) for evaluation (Aleph only produces binary results). To binarize the outcomes of the probabilistic SVM values we used threshold of 0.5.

For experiments $B1$ and $B2$, we used Aleph's default parameters: $noise = 0$ and $evalfn = coverage$.

III. Results

A. Performance of Classifiers

Table I shows the performance of our classifiers on the 168 test set. We used the metrics Correctly Classified Instances (CCI, a.k.a. accuracy), Kappa statistics (K), F-measure (F, harmonic mean between Precision and Recall), Area Under the ROC Curve (AUROC), True Postive Rate (TPR), Precision (P) and True Negative Rate (TNR) to assess the classifiers.

TABLE I
Performance of Classifiers on Test Set

Platform	Exp.	CCI	K	F	AUROC	TPR	P	TNR
Aleph	B1	77.4	0.37	0.52	–	0.43	0.65	0.91
Aleph	B2	79.8	0.41	0.52	–	0.40	0.76	0.95
WEKA	Prev	79.2	0.47	0.62	0.82	0.60	0.64	0.87
WEKA	A	81.0	0.51	0.64	0.85	0.60	0.68	0.89

The first lines ($B1$ and $B2$) are experiments with Aleph, with all variables and without the four variables, respectively. McNemar test for the binary predictions of these two outcomes gave a p-value of 0.18, which indicates that most probably those four variables are not relevant. The same happens to experiments $Prev$ and A (respectively, SVM trained on all variables and SVM trained without the four variables), with p=0.55. Aleph and WEKA agree that the classifier's performance are similar using the four variables (Side, Depth, Clockface and Quadrant) or not using the four variables. The issue is controversial: these results, on our dataset, reinforce the non-relevance of these features, although some studies show the opposite [22], [23]. Comparing $B1$ with $Prev$, we get p=0.02, which means that the Aleph classifier has worse performance than that of the SVM. We investigate this issue in the next section by training Aleph using parameters other than the default in order to reach better performance.

B. Interpretable Classifiers

Figure 1 shows ROC points comparing our SVM classifier on the test set with various Aleph performances obtained by varying noise and evalfn described in Experiment C (notice that one of the points of the SVM curve, with threshold 0.5, represents the performance shown in Table I, line $Prev$). Points for Aleph were obtained by fixing one evalfn and varying noise from 0 to 100. Because Aleph produces discrete results, one for each parameter variation, we plotted the discrete Aleph results as points and not as lines.

For low false positive rate the best results are shown by evalfn *coverage*, *cost* and *mestimate*. However, these metrics guides the search space in a way that there is not much variance in performance. For several values of noise, results are always the same. *Mestimate*, though, is the only clause evaluation function that can reach performance very close to the SVM curve for some values of noise.

For best Recalls, *gini* and *wracc* are ideal and have the advantage of having more variance, which can be explored to further improve performance.

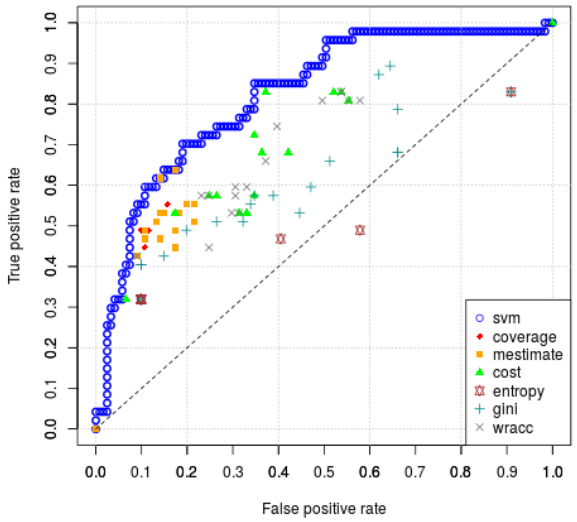

Fig. 1. ROC points for SVM and Aleph (Experiment C)

Surprisingly, entropy does not contribute to performance gain. It seems that its effect is to classify every example in the same class as we increase the noise value.

We performed McNemar tests to compare the binary predictions of the points closer to the SVM curve and concluded that the *mestimate* points that fall on the SVM curve (that use noise = 19 and noise = 93) had p-values of 0.84 and 0.23, respectively. These are very encouraging results since we can reach closer performance to our best classifier with the advantage of having interpretable rules that explain the reasons for a case being malignant or not.

Next, we present examples of clauses generated by Aleph in the **training set** to illustrate the interpretability of results (in the Prolog syntax). "Pos cover" and "Neg cover" in these examples refer to the number of positive and negative examples, respectively, covered by the rules. In total, 17 rules were generated.

```
[Rule 6] [Pos cover = 6 Neg cover = 0]
is_malignant(A) :-
      shape(A,'Round'),
      depth(A,'Middle'),
      density(A,high).

[Rule 12] [Pos cover = 17 Neg cover = 0]
is_malignant(A) :-
      shape(A,'Irregular'),
      margins(A,'Spiculated').
```

In the **training set**, rule 6 covers 6 (out of 71+) positive examples and none of the negative examples (109-), whereas in the **test set** covers 1 (out of 47+) positive example and none of the negative examples (121-). It says that if a finding A has mass shape round and is middle depth and has high density,

there is a risk of the finding being malignant. Round masses are usually benign, but this added by high mass density may be indicative of malignancy. Depth shows here, but it is not relevant. This rule is somewhat weak, since it covers only 6 cases on the training set. Nevertheless, it is a three-variable combination pattern that appears on this data on 6 malignant cases and in none of the 109 negative (benign) cases.

In the **training set**, rule 12 covers 17 (out of 71+) positive examples and none of the negative examples (109-), whereas in the **test set** covers 7 (out of 47+) positive examples and none of the negative examples (121-). It says that if a finding A has mass shape irregular and mass margins spiculated then it is suspicious of malignancy. In fact, the medical literature says that spiculated masses are 90% indicative of malignancy while irregular margins have also high risk malignancy. Although this is a trivial rule, it shows that the classifier is finding rules consistent with the medical literature.

C. Malignant Rules

We generated eight malignant rules for the training set. For each one of the 180 cases from the training set we verified if it was covered by any of the rules generated. We counted the number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). Having the confusion matrices for each rule we computed the percentage of corrected classified instances (CCI). The rules generated for the training set were then applied to the test set of 168 cases in order to investigate how many cases from the test set were covered by the rules generated in the training set. For each case that was covered by one of the malignant rules, the CCI percentage computed in the training set for that rule, was assigned to the case covered in the test set. If one case in the test set was covered by more than one rule, the CCI value chosen was the maximum registered for that case. For each case of the 168 we obtained a CCI maximum value of percentage for being covered by a malignant rule. Those CCI values were then used to plot the malignant rules lines in the ROC curves in Figures 2 and 3.

Figure 2 shows ROC points comparing our SVM classifier (weka_svm) with malignant rules generated by Aleph. For sensitivity values until 0.8 and false positive rate until 0.5, malignant rules closely approach our best SVM classifier's performance (weka_svm). In Figure 3 malignant rules clearly surpass the performance of our best decision tree classifier (weka_j48) for almost all sensitivity and false positive rate values. These are extremely interesting results from Aleph since we can reach closer performance to our best SVM classifier and even overcome our best decision tree classifier with the advantage of having interpretable rules that help to explain the reasons for a case being malignant.

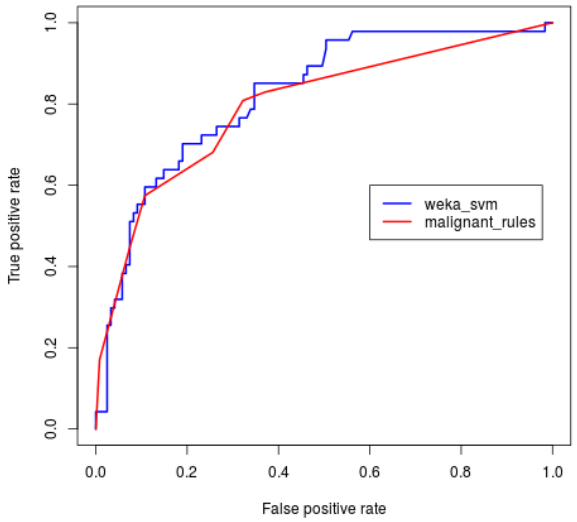

Fig. 2. ROC points for SVM and malignant rules from Aleph

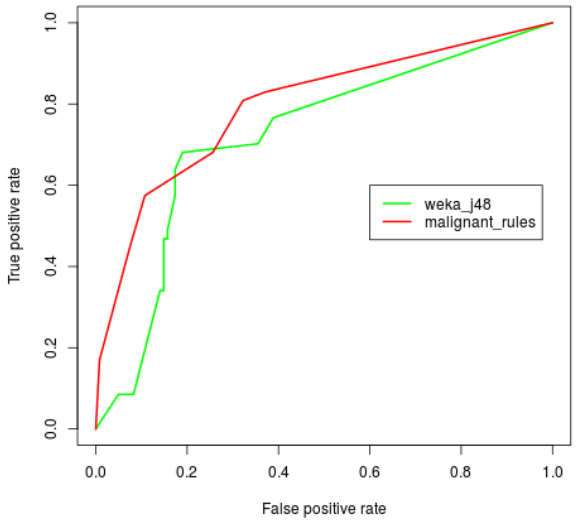

Fig. 3. ROC points for malignant rules from Aleph and decision tree classifier

IV. Conclusions and Future Work

We explored alternatives to our best SVM classifier and have shown that it is possible to obtain more interpretable classifiers with the same performance on the test set. We have shown also that we can generate interpretable classifiers with higher performance than our best decision tree classifier, which help to explain the reasons for a case being malignant. We also studied the relevance of Side, Clockface, Depth and Quadrant on predicting breast cancer. It is not clear from the literature if these variables are important. We performed experiments by training classifiers using all variables and removing these four variables and concluded that for our dataset they are not important.

Our next step is to search for a smoothing function that can produce less discrete results for Aleph, and use the techniques

and methodology applied to this work in larger and more varied datasets.

ACKNOWLEDGEMENTS

This work is supported by project "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016", financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

REFERENCES

[1] N. B. C. Foundation. (2015) Breast Cancer facts. [Online]. Available: http://www.nationalbreastcancer.org/breast-cancer-facts

[2] 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 USA*, vol. 87, December 1990, pp. 9193–9196.

[3] 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, no. 1, pp. 81–87, April 1993.

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

[5] W. N. Street, O. L. Mangasarian, and W. H. Wolberg, "An inductive learning approach to prognostic prediction," in *Proceedings of the Twelfth International Conference on Machine Learning*. Morgan Kaufmann, 1995, pp. 522–530.

[6] 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.

[7] G. Luo, B. L. Stone, F. Sakaguchi, X. Sheng, and M. A. Murtaugh, "Using computational approaches to improve risk-stratified patient management: Rationale and methods." *JMIR research protocols*, vol. 4, no. 4, pp. e128–e128, 2014.

[8] G. Luo, "Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction," *Health information science and systems*, vol. 4, no. 1, p. 1, 2016.

[9] 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*, vol. 23, no. 5, pp. 554–561, 2010.

[10] 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*, vol. 258, no. 2, pp. 417–425, February 2011. [Online]. Available: http://radiology.rsna.org/content/early/2010/12/18/radiol.10100328.abstract

[11] 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.

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

[13] E. A. Sickles, "Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases," *Radiology*, vol. 179, no. 2, pp. 463–468, 1991, pMID: 2014293. [Online]. Available: http://pubs.rsna.org/doi/abs/10.1148/radiology.179.2.2014293

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

[15] 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.

[16] 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

[17] H. Nassif, F. Cunha, I. C. Moreira, R. Cruz-Correia, E. Sousa, D. Page, E. S. Burnside, and I. de Castro Dutra, "Extracting bi-rads features from portuguese clinical texts," in *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2012*, 2012, pp. 1–4.

[18] 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

[19] 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.

[20] 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: Medical Imaging Computer-Aided Diagnosis*, ser. SPIE 2011, 2011, pp. 79 631K–79 631K–7.

[21] P. Ferreira, N. A. Fonseca, I. de Castro Dutra, R. W. Woods, and E. S. Burnside, "Predicting malignancy from mammography findings and image-guided core biopsies," *IJDMB*, vol. 11, no. 3, pp. 257–276, 2015. [Online]. Available: http://dx.doi.org/10.1504/IJDMB.2015.067319

[22] E. S. for Radiotherapy and Oncology. (2016) Handbook of brachytherapy. [Online]. Available: http://www.estro.org/binaries/content/assets/estro/about/gec-estro/handbook-of-brachytherapy/j-18-01082002-breast-print_proc.pdf

[23] M. H. Amer, "Genetic factors and breast cancer laterality," *Cancer Manag Res*, vol. 16, no. 6, pp. 191–203, April 2014.

[24] N. Lavrac and S. Dzeroski, *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood, New York, 1994, out-of-print, but available from http://www-ai.ijs.si/SasoDzeroski/ILPBook/.

[25] A. Srinivasan, *The Aleph Manual*, 2007.

[26] 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

[27] J. P. D. Conceição, "The aleph system made easy," 2008.

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

[29] 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.

[30] S. Dzeroski and I. Bratko, "Handling noise in inductive logic programming," in *Proceedings of the 2nd International Workshop on Inductive Logic Programming*. Report ICOT TM-1182, 1992, pp. 109–125.

[31] N. Lavrac, P. Flach, and B. Zupan, "Rule evaluation measures: A unifying view," in *Proceedings of the 9th International Workshop on Inductive Logic Programming (ILP-99*. Springer-Verlag, 1999, pp. 174–185.