Using Machine Learning to Identify Benign Cases with Non-Definitive Biopsy

Finn Kuusisto*, Inês Dutra†, Houssam Nassif*, Yirong Wu*, Molly E. Klein*,
Heather B. Neuman*, Jude Shavlik* and Elizabeth S. Burnside*
*University of Wisconsin–Madison, Madison, WI, USA
†CRACS INESC-TEC LA, University of Porto, Porto, Portugal

***Abstract*—When mammography reveals a suspicious finding, a core needle biopsy is usually recommended. In 5% to 15% of these cases, the biopsy diagnosis is non-definitive and a more invasive surgical excisional biopsy is recommended to confirm a diagnosis. The majority of these cases will ultimately be proven benign. The use of excisional biopsy for diagnosis negatively impacts patient quality of life and increases costs to the healthcare system. In this work, we employ a multi-relational machine learning approach to predict when a patient with a non-definitive core needle biopsy diagnosis need *not* undergo an excisional biopsy procedure because the risk of malignancy is low.**

I. Introduction

When a screening mammogram presents a suspicious finding, a follow-up diagnostic mammogram is performed to further define the abnormality. If the finding remains suspicious, a core needle biopsy (CNB) may be recommended. In this procedure, a hollow needle is inserted into the breast under imaging guidance to remove small samples (“cores”) of the abnormal breast tissue. In most cases, pathologic review of the biopsy confirms the presence or absence of cancer [1]. However, in 5% to 15% of cases, the results are not definitive [2], and surgical excisional biopsy is recommended to determine the final pathology and rule out the presence of malignancy. If a malignancy is subsequently confirmed, the case is “upgraded” from non-definitive to malignant. In the US, women over the age of 20 have an annual breast biopsy utilization rate of 62.6 per 10,000, translating to over 700,000 women undergoing breast core biopsy in 2010 [3], [4]. Approximately 35,000 to 105,000 of these women then likely underwent excision, a more invasive procedure. Ultimately, a majority of these women received a benign diagnosis.

Breast cancer diagnosis is an ideal domain to develop and test machine learning methods for risk prediction because 1) a standardized lexicon with probabilistic underpinnings has been established to summarize imaging features, 2) risk factors are generally available, and 3) accurate outcomes exist through cancer registries. In the mid-1990s, the American College of Radiology developed the mammography lexicon, Breast Imaging Reporting and Data System (BI-RADS), to standardize mammogram feature distinctions and the terminology used to describe them [5]. Studies show that BI-RADS descriptors are predictive of malignancy [6], [7], [8], specific histology [9], [10], and prognostic significance [11], [12], [13].

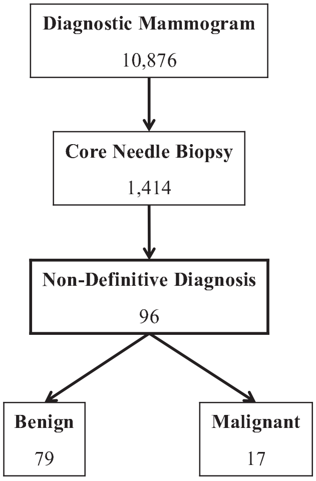
In this study, we investigate the use of machine learning to predict benign entities in cases where CNB has produced a non-definitive diagnosis. Our study considers demographic risk factors and mammographic features, as well as biopsy and pathology characteristics, to estimate the risk of upgrade. These factors and features are organized in multiple tables, which makes the dataset suitable for relational learning [14]. We generate interpretable classifiers, based on first-order logic, that capture the correlation between features included in this study to predict when a patient need *not* undergo excision.

II. Materials and Methods

Institutional review board approval was obtained prior to the commencement of this retrospective study. Written informed consent of patients was not required. We included a population of patients that underwent 1,414 consecutive CNB, as a result of a diagnostic mammogram, from Dec 31, 2005 to Dec 31, 2009. Of these biopsies, 96 were prospectively given a non-definitive diagnosis after discussions in clinical conference meetings. We limited our dataset to this subset. For all 96 cases, we collected information related to the pathological diagnoses, technical biopsy procedure and materials, as well as patient history, information about previous mammograms, and BI-RADS descriptors associated with the biopsied tissue from our multi-relational database. A diagram of our case inclusion process can be seen in Figure 1. Each of the top three boxes represent a step in our inclusion filtration process and show the number of cases included at that step. The bottom two boxes represent our malignant and benign groups.

We use the inductive logic programming (ILP) system, Aleph [15], to predict when a patient should *not* undergo excision. ILP is a machine learning approach that learns a set of rules in first-order logic that explain a given dataset [16]. We use ILP because it is well suited for our multi-relational dataset and because the logical rules produced can be easily interpreted by a human. Previous work using a similar dataset also showed that other methods produced worse results than ILP [17]. We make benign cases our “positive” class because we wish to find highly accurate rules that predict when this procedure is not needed. Unlike most machine learning approaches, ILP treats its positive and negative training asymmetrically, focusing on inducing rules that match many positive examples and few (ideally zero) negative examples. Readers should be aware of this wording (“positive” is benign), as it is somewhat counter-intuitive, but it is motivated by the machine learning approach we employ.

Fig. 1. Case Inclusion Diagram



We considered few training parameters and tried to select values in line with our clinical objective of identifying benign cases without missing malignancies:

minpos The minimum number of positive examples that a rule is required to cover.

noise The maximum number of negative examples that a rule is allowed to cover.

evalfn The rule cost evaluation function.

We chose 2 for minpos in order to require rules that generalize beyond a single case in the training set at minimum. For noise we chose 0, disallowing any rules that misclassify even a single malignant case in the training set, due to the high cost of missing cancer [18]. For evalfn we use the well-known F_β measure as it allows us to balance the importance of true positives (TP), false positives (FP), and false negatives (FN):

$$F_\beta = \frac{(1+\beta^2)\times \text{TP}}{(1+\beta^2)\times \text{TP} + \beta^2 \times \text{FN} + \text{FP}}$$

We chose a value of 0.1 for β, effectively making precision 10 times as important as recall. This is again because we deem it more important to avoid calling malignant cases benign.

ILP generates a theory that may consist of many different rules, where each rule is a conjunction of features that together predict the chosen positive class (i.e. benign in this task). To reduce overfitting on such a small dataset, we prune the output theory to a single rule. Our pruning process selects the rule with the best F_β score (as described above) on the training set. By pruning this way, we hope to reduce each theory to its best performing rule.

We use stratified 17-fold cross-validation for evaluation, each fold including a single malignant case in its test set. Multiple biopsies of the same patient were all placed in the same fold.

III. Results

We first present the results from 17-fold cross-validation in Table I. Recall that, because we are trying to predict which patients should *not* go to surgery, true positives are the benign cases that are correctly classified as benign, and false positives are malignant cases that are misclassified as benign. We also report aggregate precision, recall, and $F_{0.1}$ [19].

TABLE I
17-Fold Cross Validation Results

TP	FP	FN	TN	Precision	Recall	$F_{0.1}$
25	**2**	**54**	**15**	**0.93**	**0.32**	**0.91**

Each of the 17 folds produced a single theory that was then pruned to a single rule. In many of the folds, the rule produced was identical to that of another fold. What follows are the five unique rules that were produced amongst all the folds, sorted by the number of folds that produced them. They have been translated from first-order logic to English to make them easier to read. The performance of each unique rule on the full dataset can be found in Table II, along with the number of folds in which each rule was learned.

The 5 unique learned rules say that a non-definitive case is benign if:

1) **The patient did not have a previous surgery, imaging did not present a spiculated mass margin, the abnormality remained in post-biopsy imaging**
2) **Imaging did not present an indistinct mass margin, imaging did not present a spiculated mass margin, the abnormality remained in post-biopsy imaging**
3) **Imaging did not present a spiculated mass margin, the abnormality remained in post-biopsy imaging**
4) **Imaging did not present an indistinct mass margin, the abnormality remained in post-biopsy imaging**
5) **The patient has no personal history of breast cancer, the abnormality remained in post-biopsy imaging**

TABLE II
Individual Rule Performance on Full Dataset (# Folds is the number of folds in which a rule was learned)

Rule	# Folds	**TP**	FP	FN	TN	Precision	Recall	$F_{0.1}$
1	10	**30**	0	49	17	1.00	0.38	0.98
2	4	**29**	0	50	17	1.00	0.37	0.98
3	1	**34**	1	45	16	0.97	0.43	0.96
4	1	**31**	1	48	16	0.97	0.39	0.95
5	1	**28**	0	51	17	1.00	0.35	0.98

IV. Discussion

In this project, we demonstrate that ILP can derive rules that accurately predict when a woman may not require excision after a non-definitive core breast biopsy. All five rules predict a substantial number of cases that are benign, and only two miss a single malignancy each. Two important themes are inherent in this project. First, ILP, an approach that can construct rules from multi-relational data, appeared to be important because multiple rules contain both imaging and clinical factors. Considering all five of the rules together, the features included fall into three main categories: post-biopsy imaging (a standard part of the CNB process), mass margin descriptors, and patient history. All of the included features also have some clinically significant explanation as confirmed by our multidisciplinary (radiology, pathology, and surgery) team.

Second, the choice to predict a benign outcome (rather than a more common approach of predicting malignancy) appears to be an effective strategy in this clinical situation. The cross-validation results indicate that we can potentially reduce the total number of patients with non-definitive diagnosis from undergoing excision by around 28% (25 true benigns identified) with 93% precision (2 malignancies misclassified).

Importantly, the two rules that missed single malignant cases were each only learned in a single fold, whereas the strongest rule that misses no malignancies (rule 1) was learned in ten different folds. Similarly, the second strongest rule that misses no malignancies (rule 2) was learned in four different folds. This lends support to the idea that these two rules capture a significant signal across the entire dataset. When choosing rules to implement clinically, clinicians would undoubtedly prefer rules that do not miss a cancer. Our results may indicate that the combination of fold coverage and clinical judgement could serve as criteria on which to select the most advantageous rules. In our project, this approach designates the first two rules as the most useful. Whether these rules will be generalizable to new data remains future work.

Despite a small dataset, our approach was able to infer highly accurate rules. We are working on testing our method on a larger population of patients, which should allow us to learn better rules. We also note that, while several of the rules are derived from imaging features, the pathology features are poorly utilized. This is likely because, in our database, the imaging features are well populated and standardized using BI-RADS, but most of our pathology results are stored in free text. We are working on improving our data collection process, which may be reflected in an increased use of pathology features in future work.

V. Conclusion and Future Work

In this work, we use an ILP rule-learner to develop classifiers that successfully predict when a patient with a non-definitive core needle biopsy may not need to undergo excisional biopsy. The unique contribution of this project involves the use of a multi-relational dataset containing features from multiple disciplines (radiology, pathology, and surgery) to predict the most appropriate outcome: benignity. Additionally, our approach has the advantage of generating interpretable rules, enabling clinicians to more easily consider them in practice.

In order to validate our rules and learn more robust models, we are actively collecting additional retrospective data in order to greatly increase our patient population. In addition, we have instituted prospective collection of a richer set of variables on which to predict outcomes. These preliminary rules will shortly be tested on new data to get a better understanding of their general performance. Future work will build on preliminary research that indicates that including expert advice as background knowledge can improve performance [17].

Acknowledgements

This work is supported by the US National Library of Medicine (NLM) grant 1R01LM010921-01A1. Inês Dutra would like to acknowledge the FCT project PTDC/EEI-SII/2094/2012.

References

[1] W. Bruening, J. Fontanarosa, K. Tipton, J. R. Treadwell, J. Launders, and K. Schoelles, "Systematic review: comparative effectiveness of core-needle and open surgical biopsy to diagnose breast lesions," *Ann Intern Med*, vol. 152, no. 4, pp. 238–246, Feb, 16th 2010, epub 2009 Dec 14.

[2] L. Liberman, "Percutaneous imaging-guided core breast biopsy: state of the art at the millenium," *Am J Roentgenol*, vol. 174, no. 5, pp. 1191–1199, 2000.

[3] "Projections of the total resident population by 5-year age groups, and sex with special age categories: Middle series, 2001 to 2005," United States Census Bureau, Tech. Rep., 2000, population Projections Program, Population Division, U.S. Census Bureau, Washington, D.C. 20233. [Online]. Available: http://www.census.gov/population/projections/

[4] K. Ghosh, L. J. Melton, V. J. Suman, C. S. Grant, S. S, and et al., "Breast biopsy utilization: a population-based study," *Arch Intern Med*, vol. 165, pp. 1593–1598, 2005.

[5] D'Orsi, C. J. and Bassett, L. W. and Berg, W. A. and et al., *BI-RADS®: Mammography*, 4th ed. American College of Radiology, Inc., 2003, reston, VA.

[6] L. Liberman, A. F. Abramson, F. B. Squires, J. R. Glassman, E. A. Morris, and D. D. Dershaw, "The breast imaging reporting and data system: positive predictive value of mammographic features and final assessment categories," *Am J Roentgenol*, vol. 171, pp. 35–40, 1998.

[7] M. Moskowitz, "The predictive value of certain mammographic signs in screening for breast cancer," *Cancer*, vol. 51, pp. 1007–1011, 1983.

[8] J. A. Swets, D. J. Getty, R. M. Pickett, C. J. D'Orsi, S. E. Seltzer, and B. J. McNeil, "Enhancing and evaluating diagnostic accuracy," *Med Decis Making*, vol. 11, pp. 9–18, 1991.

[9] E. S. Burnside, D. L. Rubin, R. D. Shachter, R. E. Sohlich, and E. A. Sickles, "A probabilistic expert system that provides automated mammographic-histologic correlation: initial experience," *AJR Am J Roentgenol*, vol. 182, pp. 481–488, 2004.

[10] 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.

[11] M. G. Thurfjell, A. Lindgren, and E. Thurfjell, "Nonpalpable breast cancer: mammographic appearance as predictor of histologic type," *Radiology*, vol. 222, pp. 165–170, 2002.

[12] L. Tabar, H. H. T. Chen, M. F. A. Yen *et al.*, "Mammographic tumor features can predict long-term outcomes reliable in women with 1-14mm invasive breast carcinoma," *Cancer*, vol. 101, pp. 1745–1759, 2004.

[13] R. Nakayama, Y. Uchiyama, R. Watanabe, S. Katsuragawa, K. Namba, and K. Doi, "Computer-aided diagnosis scheme for histological classification of clustered microcalcifications on magnification mammograms," *Med Phys*, vol. 31, pp. 789–799, 2004.

[14] L. D. Raedt, *Logical and Relational Learning*. Springer, 2008.

[15] A. Srinivasan, *The Aleph Manual*, 2001.

[16] N. Lavrac and S. Dzeroski, *Inductive Logic Programming: Techniques and Applications*. New York: Ellis Horwood, 1994.

[17] I. Dutra, H. Nassif, D. Page *et al.*, "Integrating machine learning and physician knowledge to improve the accuracy of breast biopsy," in *AMIA Annual Symposium Proceedings*, Washington, DC, 2011, pp. 349–355.

[18] M. Petticrew, A. Sowden, and D. Lister-Sharp, "False-negative results in screening programs: medical, psychological, and other implications," *Int J Technol Assess*, vol. 17, no. 2, pp. 164–170, 2001.

[19] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement," *SIGKDD Explor Newsl*, vol. 12, no. 1, pp. 49–57, Nov. 2010.