Analytical Cellular Pathology 35 (2012) 97–100
DOI 10.3233/ACP-2011-0029
IOS Press

Mitotic figure recognition: Agreement among pathologists and computerized detector

Christopher Malon[a,*], Elena Brachtel[b], Eric Cosatto[a], Hans Peter Graf[a], Atsushi Kurata[c], Masahiko Kuroda[c], John S. Meyer[d], Akira Saito[e], Shulin Wu[b] and Yukako Yagi[b]

[a]*Department of Machine Learning, NEC Laboratories America, NJ, USA*
[b]*Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA*
[c]*Department of Molecular Pathology, Tokyo Medical University, Tokyo, Japan*
[d]*Department of Pathology, St. Luke's Hospital (St. Louis), Chesterfield, MO, USA*
[e]*Innovative Service Solutions Division, NEC Corporation, Tokyo, Japan*

Abstract. Despite the prognostic importance of mitotic count as one of the components of the Bloom – Richardson grade [3], several studies [2, 9, 10] have found that pathologists' agreement on the mitotic grade is fairly modest. Collecting a set of more than 4,200 candidate mitotic figures, we evaluate pathologists' agreement on individual figures, and train a computerized system for mitosis detection, comparing its performance to the classifications of three pathologists. The system's and the pathologists' classifications are based on evaluation of digital micrographs of hematoxylin and eosin stained breast tissue. On figures where the majority of pathologists agree on a classification, we compare the performance of the trained system to that of the individual pathologists. We find that the level of agreement of the pathologists ranges from slight to moderate, with strong biases, and that the system performs competitively in rating the ground truth set. This study is a step towards automatic mitosis count to accelerate a pathologist's work and improve reproducibility.

1. Introduction

Beginning with Greenough's 1925 grading system [7], pathologists have attempted to quantify factors that provide a measure of an invasive breast tumor's locality and prognosis. From the seven factors in Greenough's system, Bloom and Richardson settled on three factors [3], and the 1991 Nottingham revisions gave more stable definitions to these variables [6]. The grades are widely used to inform the selection of high-risk treatments, through the information they provide about survival or the likelihood of distant metastasis.

*Corresponding author: Christopher Malon, PhD, Department of Machine Learning, NEC Laboratories America, 4 Independence Way, Princeton, NJ 08540, USA. Tel.: +1 609 951 2594; Fax: +1 609 951 2482; E-mail: malon@nec-labs.com.

In light of the critical role of grading, many authors have investigated agreement between pathologists on individual components of the grade [2, 9, 10]. The level of agreement may be reported in Cohen's Kappa statistic [5], which equals one in case of perfect agreement and zero in the case of probabililstically independent decisions. The range 0–0.2 is often considered as slight agreement, 0.2–0.4 as fair, 0.4–0.6 as moderate, 0.6–0.8 as good, and 0.8–1 as almost perfect.

Meyer's study of agreement on Bloom-Richardson grading [9] involved groups of five to seven pathologists, with each group examining 10–23 patients' slides of hematoxylin and eosin stained biopsy tissue through an analog microscope. Agreement on the overall grade was moderate, with $\kappa = 0.50$–0.59 for the various groups. The tubularity grade achieved stronger agreement than for the other components ($\kappa = 0.57$–0.83).

Agreement on the pleomorphism component was weakest ($\kappa = 0.27$–0.50). The range of agreement for the mitotic grade was $\kappa = 0.45$–0.67.

The *NEC e-Pathologist Project* aims to provide diagnosis support for anatomical pathology, by providing computerized image analysis for virtual microscopy. It has released modules for analysis of gastric tissue that are in use in a major commercial laboratory in Japan. A breast module for hematoxylin and eosin stained tissue is under development, to target the least two stable parts of the Bloom-Richardson grade. For pleomorphism, the system aims to produce output representing carefully measured statistics of nuclear shape. For mitosis, the system aims to provide a classifier with consistent judgments.

Usage of a computerized mitosis detector could be the continuation of a line of efforts to devise more stable, prognostically useful guidelines for mitosis grading. The original Bloom-Richardson system directed that hyperchromatic nuclei be counted along with true mitotic figures. Cutoffs between grades were vaguely prescribed. In the Nottingham revision, pathologists were directed instead to avoid hyperchromatic nuclei, apoptotic figures, and pyknosis. Mitotic figures in prophase were no longer to be counted because agreement was low. Figures were to be counted in ten high power fields ($25\times$ or $40\times$ magnification), and grading cutoffs were made to depend on the high power field size. Each high power field was to be taken from the tumor's periphery.

The rigor of the grade given in clinical practice appears to vary widely. Comparing pathologists who performed a quick 30-second impression of ten high power fields to those who spent 2–3 minutes applying the rules for the WHO Mitotic Activity Index (MAI), Skaland [11] found that those who followed the MAI procedure gave grades that were prognostic with two more orders of magnitude in p-value than those who did the quick impression. Some see the need for even more concrete guidelines. Baak [1, 12] attempts to describe figures to be counted in image analysis terms (loss of nuclear membrane, presence of "clear, hairy extensions of nuclear material," etc.,). One of the authors (AK) believes a complete decision tree of such image analysis rules could be drawn. However, another author (EB) believes that it may be impossible to apply even such specific guidelines by viewing just one focal plane, because some details of mitotic figures are recognizable only when focusing up and down with the microscope.

Whereas most previous investigations of agreement of mitosis have focused on the mitotic grade, we examine agreement on individual figures. In the largest previous such study we found [9], seven pathologists examined 43 potential mitotic figures. Average pairwise agreement was $\kappa = 0.38$.

2. Methods

In the present study, we asked three of the authors (AK, JM, and SW) who are pathologists actively signing out breast cases, to examine 4,204 potential mitotic figures. These figures were taken from 2,444 high power fields in 94 breast slides, stained in hematoxylin and eosin. Tissues were provided by Massachusetts General Hospital and Tokyo Medical University, and scanned on Hamamatsu Nanozoomer scanners. At full resolution ($40\times$) the scanners afforded a resolution of 4.39 pixels per micron. Because a random selection of nuclei would include too many obvious non-mitotic figures, one author selected the candidates manually, intending to obtain figures that were mitotic or worth a closer look.

Each pathologist examined all 4,204 figures, answering the question "Is this figure mitosis or not?" with "Yes," "No," or "Maybe". For each pair of pathologists, we considered the figures where both pathologists committed to a "Yes" or a "No," and computed Cohen's Kappa on this subset. In this way, figures where the digital image was inadequate for decision should be excluded.

The second part of our investigation concerned the performance of a computerized detector. As discussed above, the description of mitotic figures in terms of decisions about shape and structure may be complex and unclear. Translating those verbal directives into manually coded image analysis rules could be dangerous. Avoiding such a heuristic approach, we developed a detector based on machine learning.

In our system, a simple rule that extracts *blobs* representing nuclei of poss-sible mitotic figures (and many other nuclei, to be rejected later) establishes a set of *candidates* from the digital micrograph of the high power field (HPF). Machine learning is applied in three phases. One phase applies a *support vector regression* [13], which remaps the color palette of the HPF to normalized values. The next phase is a *convolutional neural network* [8], applied at each extracted blob. The convolutional neural network contributes a feature to

a *feature vector*, which also contains many other measurements regarding the shape, color, mass, and texture of the blob and its neighborhood. In the final phase, a *support vector machine* [13] uses the feature vector to classify the area around the blob as a mitotic figure or not.

Machine learning algorithms learn classification rules by taking a set of *training data* where features can be measured and true classifications for each example are known. Using the training data, they find parameters to be used for classification by solving a minimization problem. The learned parameters can then be applied to new examples, where the features can be measured but the classification is not given, to predict the correct classification.

Applying machine learning thus requires that we divide our data into a set for training and a set for testing. We reserved 799 mitotic figures for testing and used the remainder for developing our algorithm. Figures for testing and figures for training never came from the same slides. This careful separation ensures that our test figures reflect tissues totally unseen during development.

Training the machine also requires that we assign one *ground truth* classification to each figure in the training set, but we have three classifications–one from each pathologist. We aggregated the pathologists' decisions using majority voting. Namely, figures with two or three "Yes" labels were taken as ground truth positive, and figures with two or three "No" labels were taken as ground truth negative. Figures for which there were neither two "Yes" verdicts nor two "No" verdicts, such as a set of labels "Yes," "Maybe," and "Maybe," were excluded from the training set. The same procedure was used on the testing data set.

We find that this protocol is useful for investigation purposes because it allows a fair comparison of the machine to pathologists. If the only figures in the data set were those for which all three pathologists agreed, there would be no fair baseline for the machine's performance (trivially, each participating pathologist would perfectly predict all the ground truth labels).

3. Results

Table 1 shows pairwise agreement of our three pathologists, in cases where both committed to "Yes" or "No" decisions. Statistics of agreement are computed in Table 2. The *prevalence index* and the *bias index* express attributes of agreement that do not affect Cohen's Kappa [4]. The prevalence index describes the relative frequency of agreed positives versus agreed negatives. The bias index compares the frequency of positive/negative disagreements and negative/positive disagreements.

Table 1
Pathologist agreement of mitotic figures

Pathologists	Yes/Yes	No/Yes	Yes/No	No/No
A/B	1352	4	789	102
A/C	2705	20	172	83
B/C	1506	756	15	461

Table 2
Statistics of pathologist agreement of mitotic figures

Pathologists	Cohen's κ	Prevalence	Bias
A/B	0.13	0.56	−0.35
A/C	0.44	0.88	−0.05
B/C	0.39	0.38	0.27

We find that strong biases exist between pairs of observers. B is much more likely to reject a figure called mitotic by C or A than to count a figure rejected by C or A. A measurable but much less significant bias exists between C and A: C is more likely to reject a figure counted by A than A is likely to reject a figure counted by C.

In Table 3, we examine how each pathologist's vote predicted the ground truth labels in the test set, which were determined by the result of majority voting. This calculation is intended as a baseline for the binary classifier trained by machine learning, which decides "Yes" or "No" for every candidate mitotic figure. The pathologists had the additional option of saying "Maybe" to a figure, so their binary classification performance may be considered as a range, bounded by the performance levels if all "Maybe" decisions were changed to either "Yes" or "No".

Observer C's performance at first looks surprisingly high, but all observers were advantaged over the machine by contributing a vote to the "majority label." Recall from Table 1 that A and B disagreed on 35% of the cases where they both committed to "Yes" or "No". In each of these cases, C's vote actually defined the majority label. As the tie-breaker in this protocol, it is naturally expected that he has high agreement with the majority label.

Table 3
Prediction of test figures where two pathologists agreed

Observer	Majority label	Observer Yes/Maybe/No	Lower bound Agreement	Upper bound agreement
A	Positive (726)	658/65/3	90.6%	99.6%
	Negative (73)	15/36/22	30.1%	79.4%
B	Positive (726)	394/166/166	54.3%	77.1%
	Negative (73)	0/0/73	100.0%	100.0%
C	Positive (726)	720/4/2	99.2%	99.7%
	Negative (73)	2/2/69	94.5%	97.3%
Machine	Positive (726)	462/0/264	63.6%	63.6%
	Negative (73)	1/0/72	98.6%	98.6%

The last line of Table 3 shows the machine's performance. Compared to A, the machine performs much more strongly on negatives, but not as well on positives. Compared to B, the machine misclassifies only one more negative figure, while falling within the bounds of B's performance on positive figures.

We find this result encouraging, considering several disadvantages of the machine. One disadvantage is that the machine was confined to looking at a small box around each mitotic figure, whereas the pathologist could examine the entire HPF. Another is the lack of a special strategy for telophase figures. Each blob in a telophase figure is regarded separately (although the second blob is visible). Features considering the relationship between the two nearby blobs should improve the performance further.

Although the range of Cohen's Kappa for pathologists on individual mitotic figure recognition is perhaps not surprising, given its range on grade-level agreement, we were surprised to find strong biases. The biases suggest that different pathologists interpret grading guidelines differently. We hope that the development of a computerized mitotic detector may be one step towards the establishment of more stable tissue grading.

References

[1] J. Baak, Mitosis counting in tumors, *Human Pathology* **21**(7) (1990), 683–685.

[2] J.P.A. Baak, E. Gudlaugsson, I. Skaland, L.H.R. Guo, J. Klos, T.H. Lende, H. Soiland, E.A.M. Janssen and A. zur Hausen, Proliferation is the strongest prognosticator in node-negative breast cancer: Significance, error sources, alternatives, and comparison with molecular prognostic markers, *Breast Cancer Res Treat* **115** (2009), 241–254.

[3] H.J. Bloom and W.W. Richardson, Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years, *Br J Cancer* **11** (1957), 359–377.

[4] T. Byrt, J. Bishop and J. B. Carlin, Bias, prevalence and kappa, *Journal of Clinical Epidemiology* **46**(5) (1993), 423–429.

[5] J. Cohen, A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**(1) (1960), 37–46.

[6] C.W. Elston and I.O. Ellis, Pathological prognostic factors in breast cancer, i. the value of histological grade in breast cancer: Experience from a large study with long-term follow-up, *Histopathology* **19**(5) (1991), 403–410.

[7] R.B. Greenough, Varying degrees of malignancy in cancer of the breast, *J Cancer Res* **9** (1925), 425–463.

[8] Y. Le Cun, L. Bottou, Y. Bengio and P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE* **86**(11) (1998), 2278–2324.

[9] J. S. Meyer, C. Alvarez, C. Milikowski, N. Olson, I. Russo, J. Russo, A. Glass, B.A. Zehnbauer, K. Lister and R. Parwaresch, Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: Reproducibility of grade and advantages of proliferation index, *Modern Pathology* **18** (2005), 1067–1078.

[10] P. Robbins, S. Pinder, N. de Klerk, H. Dawkins, J. Harvey, G. Sterrett, I. Ellis and C. Elston, Histological grading of breast carcinomas: A study of interobserver agreement, *Hum Pathol* **26**(8) (1995), 873–879.

[11] I. Skaland, P.J. van Diest, E.A.M. Janssen, E. Gudlaugsson and J.P.A. Baak, Prognostic differences of world health organization-assessed mitotic activity index and mitotic impression by quick scanning in invasive ductal breast cancer patients younger than 55 years, *Hum Pathol* **39**(4) (2008), 584–590.

[12] P. van Diest, J. Baak, P. Matze-Cok, E. Wisse-Brekelmans, C. van Galen, P. Kurver, S. Bellot, J. Fijnheer, L. van Gorp, W. Kwee, J. Los, J. Pe-terse, H. Ruitenberg, R. Schapers, M. Schipper, J. Somsen, A. Willig and A. Ariens. Reproducibility of mitosis counting in 2,469 breast cancer specimens: Results from the multicenter morphometric mammary carcinoma project, *Human Pathology* **23**(6) (1992), 603–607.

[13] V. Vapnik, Statistical learning theory, *Wiley* (1998).