Journal of Digital Imaging (2019) 32:702–712
https://doi.org/10.1007/s10278-019-00181-8

Computer-Assisted Nuclear Atypia Scoring of Breast Cancer: a Preliminary Study

Ziba Gandomkar[1] · Patrick C. Brennan[1] · Claudia Mello-Thoms[1,2]

Published online: 4 February 2019

Abstract

Inter-pathologist agreement for nuclear atypia scoring of breast cancer is poor. To address this problem, previous studies suggested some criteria for describing the variations appearance of tumor cells relative to normal cells. However, these criteria were still assessed subjectively by pathologists. Previous studies used quantitative computer-extracted features for scoring. However, application of these tools is limited as further improvement in their accuracy is required. This study proposes COMPASS (COMputer-assisted analysis combined with Pathologist's ASSessment) for reproducible nuclear atypia scoring. COMPASS relies on both cytological criteria assessed subjectively by pathologists as well as computer-extracted textural features. Using machine learning, COMPASS combines these two sets of features and output nuclear atypia score. COMPASS's performance was evaluated using 300 images for which expert-consensus derived reference nuclear pleomorphism scores were available, and they were scanned by two scanners from different vendors. A personalized model was built for three pathologists who gave scores to six atypia-related criteria for each image. Leave-one-out cross validation (LOOCV) was used. COMPASS was trained and tested for each pathologist separately. Percentage agreement between COMPASS and the reference nuclear scores was 93.8%, 92.9%, and 93.1% for three pathologists. COMPASS's performance in nuclear grading was almost identical for both scanners, with Cohen's kappa ranging from 0.80 to 0.86 for different pathologists and different scanners. Independently, the images were also assessed by two experienced senior pathologists. Cohen's kappa of COMPASS was comparable to the Cohen's kappa for two senior pathologists (0.79 and 0.68).

Keywords Breast · Breast cancer · Microscopy · Nuclear atypia grading · Nuclear pleomorphism grading · Pattern recognition

Introduction

Breast cancer is a heterogeneous disease and different treatment options are available for the women diagnosed with it. Prognostic factors, which represent the aggressive potential of the tumor, could provide valuable information for the selection of a treatment regimen. For example, hormonal treatment and adjuvant chemotherapy, which are used to increase patient survival, are expensive and could cause serious side effects, and hence are only advisable for high-risk patients [1].

Previous studies have shown that the Nottingham modification of the Scarff-Bloom-Richardson (NSBR) breast cancer grading system provides useful prognostic information [2]. However, application of the NSBR score is still limited in routine patient management due to various reasons. Among them, the considerable inter-pathologist variability and subjectiveness are major hindrances. In [3], it was shown that inter-reader variation impacts on a patient's risk assessment for hormonal treatment and adjuvant chemotherapy.

The NSBR grading system has three contributing components, namely, the magnitude of nuclear pleomorphism, the degree of gland formation, and the number of mitotic figures [2]. The overall NSBR score is an average of scores of these three components. Nuclear pleomorphism (or atypia) score represents the variations in size, shape, and appearance of tumor cells relative to normal cells. Although different criteria have been proposed to compare appearance of the tumor cells to normal cells, the assessment of these criteria is qualitative and subject to inter-pathologist discrepancies. In the clinical practice, with lack of quantitative measurements, the

✉ Ziba Gandomkar
ziba.gandomkar@sydney.edu.au

[1] Discipline of Medical Imaging and Radiation Sciences, Medical Image Optimisation and Perception Group (MIOPeG), The University of Sydney, 512/Block M, Cumberland Campus, Sydney, NSW, Australia

[2] Carver College of Medicine, Department of Radiology, University of Iowa, Iowa City, IA, USA

pathologist must decide how to categorize a nucleus with mixed features (for example, small but with an irregular shape) and that might explain why the agreement among readers is very poor.

Previous studies investigated the magnitude of interobserver variability in NSBR grading and its components. The percentage agreement among pathologists in previous studies ranged from 43 to 74%, with Cohen's kappa ranging from 0.19 to 0.74 [4–8]. It was also shown that among these three components, the agreement on the nuclear pleomorphism score was the weakest, with percentage agreement of 55–68% and Cohen's kappa ranging from 0.27 to 0.5 [4, 7].

In addition to be a contributing factor of the NSBR grade, the nuclear atypia score might be a more useful prognostic tool compared to the overall NSBR grade for patients with invasive lobular carcinoma, as mitotic activity and tubule formation vary little in these patients [9]. Due to the importance of the nuclear atypia grade and the lack of agreement among pathologists for grading it, recently a few studies aimed at devising automatic algorithms for nuclear pleomorphism scoring [10]. However, application of these algorithms has been limited as their accuracy should be further improved.

In this paper, we propose a method for reproducible nuclear pleomorphism scoring called COMPASS (COMputer-assisted analysis combined with Pathologist's ASSessment). Unlike previous algorithms which aimed at providing an independent second opinion to the pathologists, COMPASS combines the pathologist's assessment of six criteria related to the nuclear atypia with computer-extracted features and assigns a nuclear pleomorphism score to the image based on both subjective scores and objective features. Another novelty of COMPASS is being a hybrid segmentation-based and texture-based approach to extract the computer-related features from the digitized slides. In the previous automatic nuclear grading methods, the features were either extracted from the segmented nuclei (segmentation-based methods [11]) or from the entire tissue [12]). However, COMPASS involves a coarse segmentation to restrict further analysis to a few regions of interest followed by textural feature extraction from these areas. An additional uniqueness of COMPASS is that, being a personalized model, it considers each individual's unique perceptual pattern, and eliminates systematic over- or under-estimating of each grader. In [13], it was shown that some pathologists are prone to under-grading while others systematically over-grade the cases. Junior pathologists are target users for COMPASS as in general less experienced pathologists have lower agreement levels with a consensus of expert readers [13, 14] and could benefit significantly from such an algorithm. Unfortunately, due to lack of expert pathologists with subspecialty training in reading breast biopsies, many specimens are currently interpreted by less experienced or general pathologists. This paper aims to investigate the possibility of improving junior pathologists' performances to a level comparable to the expert readers' performance by using computer-extracted features combined with a systematic evaluation of cytological features by the pathologists.

Materials and Methods

Dataset

Three-hundred images were obtained from the Mitosis Atypia challenge 2014 dataset [15], which is publicly available. Three of the images were excluded as there was no tumor region present in them, and hence, no atypia grade was associated with them.

Nuclear pleomorphism scores were given by two experienced senior pathologists. In case of disagreement, a third pathologist scored the image and the final score was obtained based on a vote of the majority. Therefore, for each image, a consensuses-driven ground truth was provided in the database. Based on NSBR, a score of 1 is given to an image when there is little increase in the size of nuclei in comparison with normal breast epithelial cells, the outlines of nuclei are regular, and the nuclear chromatin is uniform. When the cells are larger than normal with visible nucleoli and have open vesicular nuclei, with moderate variations in size and shape among cells, a score of 2 is assigned. A score of 3 is appropriate when nuclei are vesicular with prominent, often multiple nucleoli, have noticeable variations in shape and size, and large and bizarre nuclei are present in the sample [2]. All images were scanned by two different scanners, namely, Aperio Scanscope XT and Hamamatsu Nanozoomer 2.0-HT. The pathologists graded images at ×20 magnification, which covered approximately 0.511 mm^2 of tissue. The area located inside tumors and was selected by an experienced pathologist prior to the experiment. The nuclear grades were given at ×20 magnification level as this level of magnification is mostly used to grade nuclear atypia in the clinical practice. For each ×20 image, the database included the scores given by two original senior pathologists and consensuses-driven ground truth.

In addition, three junior pathologists were asked to evaluate six criteria related to nuclear atypia and give a score from one to three for each criterion. These criteria were nuclei size, nucleoli size, anisonucleosis (size variation within a population of nuclei), chromatin density, regularity of nuclear contour, and membrane thickness. Some of these criteria (nuclei size, nucleoli size, and regularity of nuclear contour) are explicitly mentioned in NSBR grading [2]. These criteria are also components of some other nuclear grading systems [16] which try to quantify other factors that contribute the pathologists' judgments about nuclear atypia grading. For example, in Fisher's modification of Black's nuclear grading, anisonucleosis, nuclear membrane, chromatin density, and nucleoli size are taken into account [17], while in Robinson's nuclear grading system, which showed high level of concordance with NSBR grade [16], nuclei size, nucleoli size, cell uniformity, regularity of nuclear contour, and membrane thickness were taken into account [18].

For each image at ×20 magnification, the junior pathologists evaluated the criteria on four sub-images at ×40 magnification (resolution of 0.2455 μm/pixel for Aperio and horizontal resolution of 0.2273 μm/pixel and vertical resolution of 0.2275 μm/pixel for Hamamatsu). The four sub-images contained the same region in the specimen. For scoring these criteria, ×40 frames were used as the detailed description of relevant nuclear atypia criteria; i.e., size of nuclei, size of nucleoli, density of chromatin, thickness of the nuclear membrane, regularity of nuclear contours, and anisonucleosis (i.e., size variation within a population of nuclei) might require a higher magnification level. Hence, for an image at ×20 magnification, each junior pathologist gave 24 (6 criteria × 4 images) scores describing criteria relevant to nuclear atypia.

COMPASS

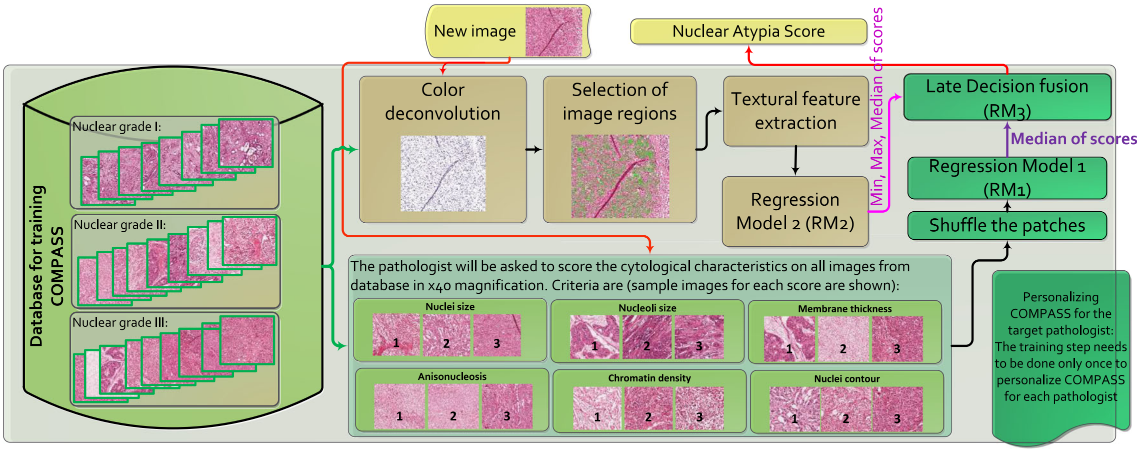
Overview

The steps of COMPASS are depicted in Fig. 1. As shown, COMPASS consists of two modules. The first module generates a score based on the pathologist's assessment of the four images at ×40 magnification. The images at this magnification level were used to ensure that the cytological features were visible to the pathologists. The second module generates atypia scores based on textural features from the image at ×20 magnification. We assumed that the cytological features and texture features provide complementary information as one describes the appearance of individual cells while the other one describes global appearance of a group cells. In the last stage of COMPASS, the scores corresponding to each image from both modules are combined by using an ensemble of trees for regression and a single score is given to the image.

Computer-Extracted Features

In order to produce the computer-extracted features, the first ten image patches (or sub-images) containing epithelial cells were automatically chosen in each image. Then, textural features were extracted from these image patches. Their size was 251 × 251 pixels. The nuclear atypia score describes the appearance of epithelial cells. Therefore, it was desirable that patches were centered at locations with high density of cancerous epithelial cells. To find the centers of the patches, the stain normalization method suggested in [19] was utilized to minimize inconsistencies in staining of different images and a set of image-processing steps (discussed below) was used to find epithelial cells. Finally, locations contained a high number of epithelial were determined and used as the center of patches.

Color deconvolution was utilized to separate H and E channels of the stained-normalized image [19]. A sample image along with separated H and E channels are shown in Fig. 2a–c. The complement of the H channel was then processed with morphological closing (a dilation followed by an erosion) using a disk of radius 2. This was followed by filling holes within the image to generate H_P (the processed image). In the context of greyscale images, holes are areas of dark pixels surrounded by lighter pixels. Finally, the candidate locations for epithelial cells are then detected by thresholding H_P and removing the connected components whose areas are less than 30 pixels. The threshold value was found empirically and set to 80. The H_P corresponding to the image shown in Fig. 2a is shown in Fig. 2 d and the thresholded image (HTh_1) overlaid on the original image using green color is indicated in Fig. 2e.



The pathologist will be asked to score the cytological characteristics on all images from database in x40 magnification. Criteria are (sample images for each score are shown):

Personalizing COMPASS for the target pathologist: The training step needs to be done only once to personalize COMPASS for each pathologist

Fig. 1 The steps of COMPASS. In each iteration of leave-out-out cross validation, one image (shown as "new image") served as the test image. The rest of images (database for training COMPASS) served as the training set for estimating the parameters and hyperparameters of COMPASS. Training involved estimating the parameters for regression model 1 (RM1), regression model 2 (RM2), and regression model 3 (RM3); 80% of training data was used to estimate the parameters of RM1 and RM2 while the rest of it was used to estimate the parameters of RM3

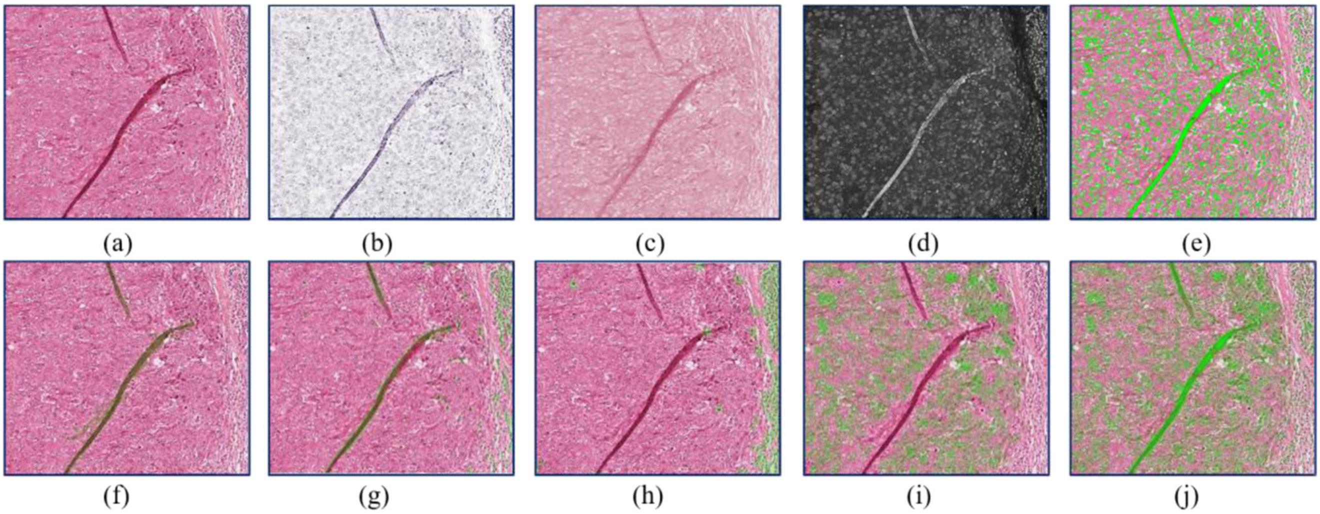


Fig. 2 **a** Original image. **b**, **c** Outputs of color deconvolution separated H and E channels, respectively. **d** The H channel image after being processed. **e** The thresholded image in the first step. **f**–**h** Three masks. **i** HF. **j** HF if the masks were not subtracted from the thresholded image

In order to extract appropriate image patches, we needed to ensure that the imperfect areas (e.g., folded tissues) and areas with normal epithelial and lymphocyte cells were excluded from HTh_1. To eliminate these areas, three different masks were generated and subtracted from HTh_1. The first mask was obtained by thresholding the complement of the E image followed by removing all connected components whose areas were smaller than 5000 pixels. Next, the holes were filled to generate $Mask_1$. To generate $Mask_2$, the complement of the H channel was filtered by a Gabor filter bank with the wavelength of 20 pixel/cycle and eight equally spaced orientations. Next, the maximum filter response was recorded for each pixel. Finally, the maximum response image was thresholded to $Mask_2$. $Mask_3$ included areas with normal epithelial tissue and lymphocytes which are darker, smaller in size, rounder, and without irregularities or broken areas in their membrane. Therefore, filtering the HP with a Laplacian of Gaussian (LoG) followed by thresholding of the filtered image was used. Previously, LoG was utilized to detect epithelial cells [20] and mitotic figures [21]. The standard deviation of the filter determines the size of the structure which is detected by the LoG. Here we found the appropriate size empirically and set it to 20 pixels. The output of LoG filter was then thresholded and the connected components with an area smaller than 2000 pixels were eliminated from $Mask_3$. All three masks were subtracted from HTh_1 to generate HTh_2. As stated previously, we want to find hypercellular areas. To do so, HTh_2 was convolved with a Gaussian filter to generate HF. Therefore, when multiple cells are present in a neighborhood of a pixel, it will have a high value in HF. Three masks and HF are shown in Fig. 2f–i. Figure 2j depicts HF if the masks were not subtracted from HTh_2. As shown, the subtraction is essential to restrict the analysis to the tumor areas. Finally, HF was normalized and ten pixels whose intensity was at least 0.75 were randomly selected from HF. The distance of the selected points should be greater than 100 pixels to ensure that most of the image have been sampled by these image patches (although some amounts of overlap have been permitted). These ten pixels were selected as the center for ten patches per each image.

Next, the textural features listed in Table 1 were extracted from each patch. The textural features [22] were extracted

Table 1 Extracted features from each patch

Feature type (feature name)
First-order statistics features
(AVE, STD, 1st, 5th, 25th, 50th, 75th, 95th, 99th percentile of intensity)
Haralick texture features averaged over four directions for $d = 3$ pixels
(Contrast, correlation, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity, sum of squares, sum average, sum entropy, difference variance, difference entropy, information measure of correlation 1 and 2, inverse difference normalized, inverse difference moment normalized)
Local binary patterns
(Uniform local binary patterns with number of number of neighbors = 8)
Features from gray-level run length matrix
(Short-run emphasis, long-run emphasis, gray-level non-uniformity, run percentage, run length non-uniformity, low gray-level run emphasis, high gray-level run emphasis)
Gabor-based features
(AVE energy of filtered image using Gabor filter bank in one scale and six orientations)
Features based on maximum response filters
(AVE energy of in eight filtered images)

AVE and STD are average and standard deviation, respectively from H channel, blue-ratio channel, and each one of three RGB channels. The images were also converted to Lab, YUV, HSL, and LMS color spaces, and the features were extracted from each channel of these color spaces.

Regression Models

As shown in Fig. 1, in the intermediate steps of COMPASS, there are two regression models, namely, regression model 1 (RM1) and regression model 2 (RM2). The inputs of RM1 were the scores given by the pathologists while RM2 relied on the textural features. Both RM1 and RM2 were ensembles of trees for regression which comprised of a weighted combination of multiple regression trees.Pathologists scored six atypia-related criteria on four images at ×40 magnification for each image in the dataset. This resulted in a 24-dimensional feature vector for each image. If bizarre nuclei were present in one of the four images at ×40 magnification, the grade of the image is 3, and this does not depend on the arrangement of the four images. Therefore, for an input image, all 24 possible combinations of the shuffling of these four ×40 images were generated. Then RM1 assigns a score to each of these 24 possible permutations, and the final score of the image is the median of these values. For each test image, ten patches were selected as suggested in 2–2-2. Each one of these patches was inputted to RM2.

For training RM1 all 24 possible combinations of four ×40 images were generated for each ×20 image in the training set. This increased the size of the training set by 24 times and made RM1 invariant to the spatial layout of the structures within the image. For training RM2, each patch was considered as an instance, and the grade of the image (from which the patch was selected) was considered as the grade of the patch. Hence, the size of the training set for RM2 was ten times larger than the number of the images.

One of the main challenges in using ensemble models is setting the hyperparameters of the model because they could affect the performance of the model. We used Bayesian optimization for hyperparameter tuning [23]. Here the optimization searched over the ensemble method, namely, either Bag (bootstrap aggregation) or LSboost (least squares boosting), over the number of weak learners, over the learning rate for shrinkage of the LSBoost method, over the minimum number of leaf node observations in the template tree, and over the number of features to select at random for each split in the tree.

Late Decision Fusion

As shown in Fig. 1, the median of 24 values given by RM1 to 24 possible permutations of four ×40 images, along with minimum, median, and maximum scores given by RM2 to ten patches of each image, built the feature vector for RM3. RM3 was an ensemble of trees for regression as well. In order to find the cutoff values to threshold the scores from regression models and produce three-scale atypia grades, two receiver operating characteristic (ROC) curves were generated, one for detecting high-grade images (grade 3 against combined grade 1 and 2) and one for low-grade images (grade 1 against combined grade 2 and 3) and their optimal operating points were found.

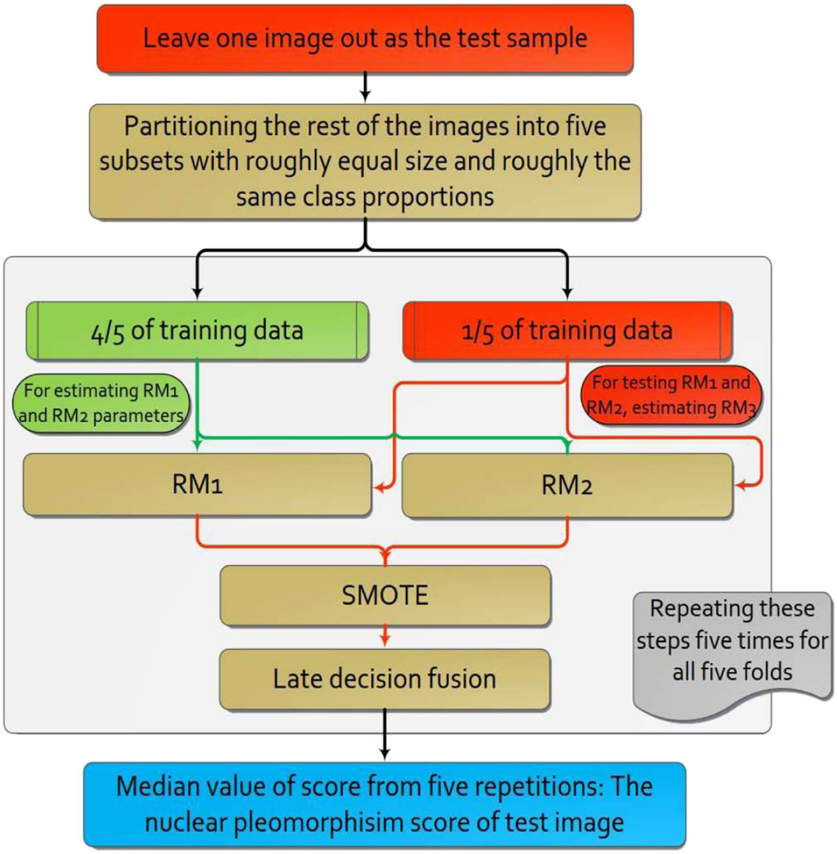
For training RM3, the instances from grade 1 and 3 were upsampled by applying the Synthetic Minority Oversampling TEchnique (SMOTE) [24]. The numbers of nearest neighbors to use were set to 3 and 5 for grade 1 and 3, respectively, and the percentages of SMOTE instances to create were set to 200% and 400%. The hyperparameters of RM3 were also set by using Bayesian optimization [23].

Evaluation of COMPASS

As COMPASS is a personalized tool, first the parameters of the model should be estimated for each pathologist by asking the readers to assign scores to six nuclear atypia criteria on the images for which the expert-consensus derived reference nuclear pleomorphism scores are available. After this training stage, COMPASS can be used to score new images. Therefore, for evaluating COMPASS's performance, we need to first train the model and then test the trained model on unseen data. As the size of the dataset was small, it was not possible to perform hold-out validation. Therefore, we used leave-one-image-out cross validation (LOOCV). Hence, each time one of ×20 images (and four corresponding ×40 images) served as the test data and the rest of the images (training data) were utilized for estimating the parameters of COMPASS. The percentage agreement and Cohen's kappa [25] were calculated for each junior pathologist. The percentage agreement indicates the number of concordance cases divided by the total number of cases. Cohen's kappa measures interobserver agreement for categorical items and is a more robust measure than the agreement rate, as it considers the possibility of the agreement happening by chance. It usually interpreted as follows: kappa ≤ 0 shows no agreement, 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–0.99 as almost perfect agreement [25].

In each iteration of LOOCV, the training data was partitioned into five subsets with roughly identical size and roughly the same class proportions as in the original dataset. Four subsets were utilized to estimate the parameters of RM1 and RM2. Next, the images in the remaining subset were inputted to RM1 and RM2. As stated earlier, four features were extracted from the scores given by RM1 and RM2 to each instance in this subset and used to train RM3. Finally, a score was given to the test data by the trained model. Figure 3 shows the procedure for training COMPASS. This procedure was repeated five times; each time one of the subsets was used to estimate the parameters of RM3, and the rest of them were

Fig. 3 The evaluation procedure of COMPASS


Leave one image out as the test sample
Partitioning the rest of the images into five subsets with roughly equal size and roughly the same class proportions
4/5 of training data
1/5 of training data
For estimating RM1 and RM2 parameters
For testing RM1 and RM2, estimating RM3
RM1
RM2
SMOTE
Late decision fusion
Repeating these steps five times for all five folds
Median value of score from five repetitions: The nuclear pleomorphisim score of test image

used to estimate the parameters of RM1 and RM2. Therefore, five scores were given to each test data. The median value of all these scores was assigned to each image. In order to set the hyperparameters of regression models, in each of the five repetitions, ten-fold cross validation was used for Bayesian optimization. To achieve a result robust to partitioning noise, at every iteration, the cross validation was repartitioned.

Results

Performance of COMPASS

As described in the "Evaluation of COMPASS" section, leave-one-image-out cross validation was used to evaluate the performance of COMPASS for each scanner. COMPASS is personalized; hence, it should be trained and tested for each reader separately. Table 2 shows the confusion matrices of COMPASS for each scanner and each junior pathologist. In the table, the upper triangular part of the matrix represents "under-graded instances" and the lower part represents "over-graded instances" based on COMPASS.

The paired Mann-Whitney U test was used to compare the grades given by COMPASS when tested on Aperio images with the given grades for Hamamatsu images. The given grades were not significantly different (junior pathologist 1: $z = -1.1$, $p = 0.29$; junior pathologist 2: $z = 0.48$, $p = 0.63$; junior pathologist 3: $z = 0.86$, $p = 0.39$). Also, Spearman's rank-order correlation coefficients between the scores (before thresholding it to produce

Table 2 Confusion matrices

	Pathologist 1 Aperio Scanscope			Pathologist 2 Aperio Scanscope			Pathologist 3 Aperio Scanscope		
	G1	G2	G3	G1	G2	G3	G1	G2	G3
G1	**18**	1	0	**20**	3	0	**16**	3	0
G2	5	**215**	5	3	**213**	6	7	**214**	8
G3	0	6	**47**	0	6	**46**	0	5	**44**
	Pathologist 1 Hamamatsu			Pathologist 2 Hamamatsu			Pathologist 3 Hamamatsu		
	G1	G2	G3	G1	G2	G3	G1	G2	G3
G1	**16**	0	0	**19**	3	0	**16**	1	0
G2	7	**212**	5	4	**207**	5	7	**217**	5
G3	0	10	**47**	0	12	**47**	0	4	**47**

Columns are true labels (based on the consensuses of pathologists) while rows are labels from COMPASS. COMPASS's performance for images scanned by Aperio scanner Hamamatsu scanner is shown separately. G stands for grade. COMPASS is a personalized tool, so each performance was evaluated and reported for each pathologist

The correct predictions are located in the diagonal of the table(shown in bold) three scale grades) given to the images from two scanners were 0.74, 0.77, and 0.75 for three junior pathologists.

Comparison of COMPASS With Senior Pathologists

We retrospectively simulated the adoption of COMPASS by the junior pathologists and compared the performance of COMPASS to that of senior pathologists. Although in real clinical practice, COMPASS would provide feedback to a pathologist, who would give the final score to the case, we assumed that the junior pathologists would accept the decision of COMPASS as our data has been retrospectively collected. The average Correct Classification Rate (CCR) per each grade is shown in Tables 3 and 4 for all junior pathologists. The values are an average of the two scanners. Similarly, on the right side of the table, CCRs are shown for the senior pathologists. As shown, the overall performance of COMPASS was comparable to that of the senior pathologists.

Cohen's kappa was also calculated to measure the magnitude of agreement of COMPASS with the ground truth. The value is calculated for each pathologist and each scanner separately and shown in Tables 3 and 4. Similarly, Cohen's kappa was calculated for the senior pathologists. For nine images of the databases, one of the senior pathologists could not assign a grade. As shown the agreement level is almost perfect except when COMPASS was adopted for the third junior pathologist and images from Aperio scanner were used. The Cohen's kappa value is substantial for this arrangement. For the senior readers, the agreement level was substantial to almost perfect.

The joint distribution of grades from COMPASS and each of the senior pathologists was investigated to find the percentage of the images that both graded correctly, only one of them graded correctly, or both graded incorrectly. The result is shown in Fig. 4. Results for both scanners were combined to generate the plots. Each row in the plot shows COMPASS as adopted by one of the junior pathologists, and each column represents one of the senior pathologists. Among misclassified images, the percentage of images which were graded incorrectly by both COMPASS and senior pathologists (orange areas in the plots) were lower than those which were graded correctly by one of them. Hence, to some extent, COMPASS could complement the senior pathologist's performance.

Computerized tools for providing a second opinion to pathologists are usually more useful with the difficult cases. We assumed that cases, which were misclassified by one of the senior pathologists, are more difficult than other cases and investigated the performance of COMPASS in this subset of images. For these images, a CCR of 85.7%, 81.6%, and 83.7% was achieved between COMPASS and the reference nuclear grade for three pathologists. Therefore, more than 80% of these relatively difficult cases were correctly classified by COMPASS.

Table 3 Cohen's kappa and percentage agreement of senior pathologists with the consensuses-driven ground truth

	1	2
I (Cohen's kappa)	0.79	0.68
E (Cohen's kapp)	0.85	0.73
G1 (%)	78.3%	82.6%
G2 (%)	90.1%	91.4%
G3 (%)	**98.1%**	69.2%
T (%)	90.6%	86.9%

The highest accuracy is shown in bold. I (E): the cases to which the readers could not assign a grade were included (excluded). G and T stand for grade and total

Table 4 Cohen's kappa and percentage agreement of COMPASS with the consensuses-driven ground truth

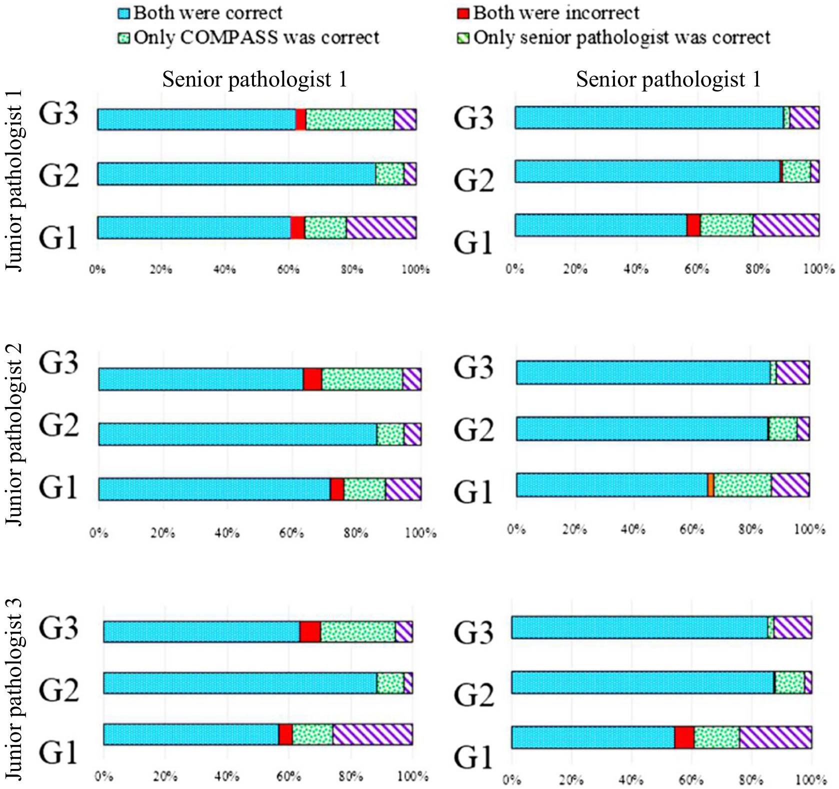
	JP1	JP2	JP3
SA (Cohen's kappa)	0.86	0.85	0.80
SH (Cohen's kapp)	0.81	0.81	0.85
G1 (%)	73.9%	**84.8%**	69.6%
G2 (%)	96.2%	94.6%	**97.1%**
G3 (%)	90.4%	89.4%	87.5%
	93.4%	92.9%	93.3%

The highest accuracy is shown in bold. JP stands for junior pathologist. SA: Aperio. SH: Hamamatsu Nanozoomer 2.0-HT Scanner

Added Benefits of Textural Features

The added benefit of textural features was investigated by comparing the performance of COMPASS against that of two baseline approaches that only used the scores given by the pathologists to the cytological characteristics of images. The first one was grading based on the total cumulative score given to all criteria. Most of the nuclear grading systems produce the final nuclear grade of each sample by summing up all scores given to the considered criteria [16]. Hence, we also used this approach for comparison. The second approach was based on the score assigned by RM1. One possible benefit of COMPASS is to use a complex non-linear regression model to associate the scores given by the pathologists to the nuclear grade. This part is done by RM1 in COMPASS. Therefore, we compared the performance of COMPASS with that of RM1 to investigate the importance of the added textural features. The boxplots, which display the distribution of scores among different grades, are shown in Fig. 5. The plots were generated from two baseline approaches as well as COMPASS for both scanners. The Kruskal-Wallis test resulted in p values < 0.0001 for all approaches and all pathologists, except for the first approach (sum) when adopted for the first pathologist, which led to a p value of 0.007 (χ^2 (2297) = 5.03). The rank-based version of Tukey's HSD (Tukey-Kramer) test showed that differences between all possible multiple pairs for

Fig. 4 The percentage of concordant and discordant cases for each atypia category based on scores given by COMPASS and the senior pathologists. The values are the average of two scanners. G1, G2, and G3 indicate grades 1, 2, and 3, respectively. Each row represents one of the junior pathologists



all approaches were significant ($p < 0.05$) except for grade 1 against grade 2 for the first and second approaches when adopted by the first reader. The results of the rest of the comparisons are indicated in the figure.

Also, the AUC for detecting high-grade images (i.e., grade 3) and low-grade images (i.e., grade 1) is reported in Table 5. As shown, the cumulative score led to the poorest results for detecting grade 3 for all pathologists; however, the differences

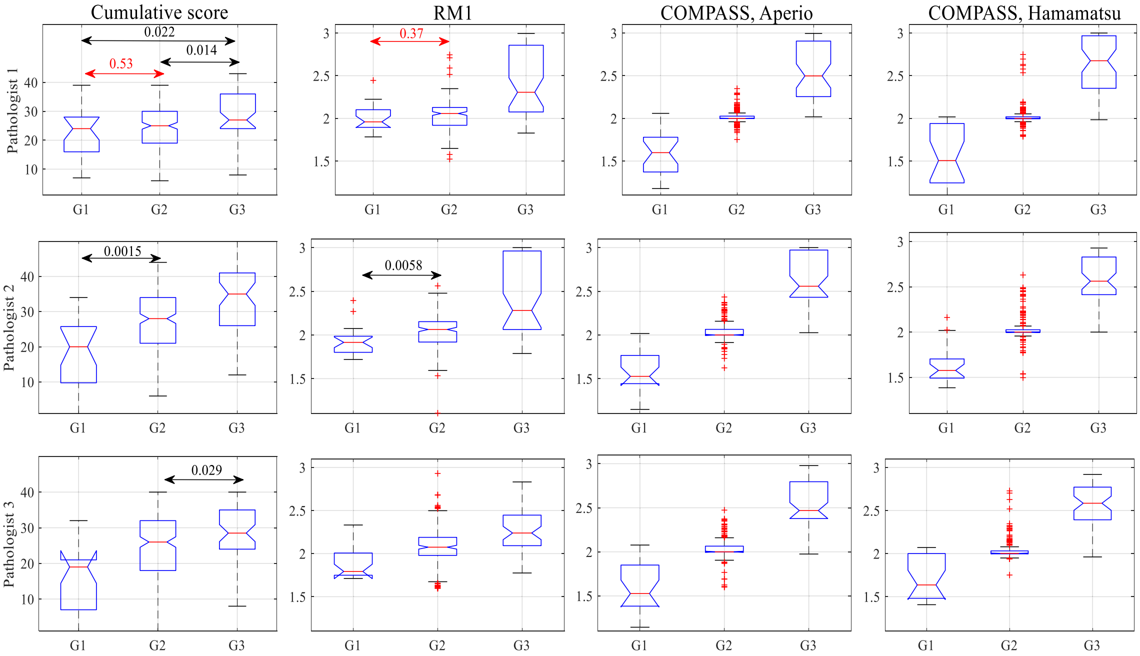


Fig. 5 Boxplots for displaying the distribution of scores given by each approach among three grades. G1, G2, and G3 represent grades 1, 2, and 3. Each row of plots represents one of the junior pathologists. Numbers above the arrows in the figure show the p values of Tukey-Kramer test for each pair. When the p values were not shown between a pair, it means p value < 001. The red numbers show insignificant differences between a pair

Table 5 AUC values for detection of grades 3 and 1

	Detection rate for grade 3				Detection rate for grade 1			
JP	**Sum**	**RM1**	**SA**	**SH**	**Sum**	**RM1**	**SA**	**SH**
1	0.631	0.784*	0.977*	0.941*	0.591	0.638	0.934*	0.888*
2	0.708	0.786	0.975*	0.959*	0.756	0.746	0.948*	0.907*
3	0.635	0.744*	0.963*	0.926*	0.754	0.779	0.935*	0.850

between AUC of the cumulative score and RM1 were not significant for the second pathologist. For detecting low grade (i.e., grade 1) RM1 outperformed the cumulative score for pathologists 1 and 3 while the two approaches resulted in an almost similar AUC values for pathologist 2.

JP represents the junior pathologist for whom COMPASS was trained and tested. SA and SH represent the performance of COMPASS for Aperio and Hamamatsu scanners. Asterisk shows that AUC value for cumulative score (sum column) is significantly lower than the compared AUC. The *p* values were calculated based on [26]

Discussion

In this paper, COMPASS, a personalized algorithm for reproducible nuclear pleomorphism grading, was introduced. The leave-one-out cross validation was used and a percentage agreement of 93.4%, 92.9%, and 93.3% was achieved between COMPASS and the reference nuclear grade for three pathologists. Therefore, the percentage agreement was almost identical for three junior pathologists. The results also suggested that the performance of COMPASS was approximately similar for both scanners. However, the CCRs of COMPASS varied among different nuclear grades. As shown in Tables 2, 3, and 4, CCRs were the highest in grade 2 and the lowest in grade 1 for all three junior pathologists. Similarly, on average, senior pathologists achieved the highest CCR in grade 2.

Table 2 represents the magnitude of over- and under-grading when COMPASS was adopted for three pathologists. In practice, the nuclear grade is one of the three building components of breast cancer grade, which is used to recommend an appropriate treatment pathway for a breast cancer patient. Previously, it was shown that inter-reader variation of breast cancer grading impacts on a patient's risk assessment for hormonal treatment and adjuvant chemotherapy [3]. For example, over-grading might result in treating women with lymph node-negative breast cancer with chemotherapy, hormonal therapy, and/or targeted therapy, while some of these women are likely to be cured by surgery and radiotherapy alone [27]. Therefore, these women will be over-treated with the adjuvant therapy. Avoiding over-grading of these women will prevent unnecessary exposure to the toxicity of adjuvant therapy. On the other hand, under-grading of lymph node-negative breast cancer might result in treating women by surgery and radiotherapy alone while some of these women, who are under-graded and do not receive adjuvant therapy, might benefit from this treatment [27]. Given the fact that the agreement on the nuclear atypia score was the weakest among three components of breast cancer grading system, developments of reproducible nuclear atypia scoring can reduce inter-pathologist variation in breast cancer grading, which is used for determining candidate patients for receiving adjuvant therapy.

Most of the previous algorithms aiming at automatic nuclear grading segmented the cells within an image and then extracted features from the segmented areas. COMPASS also detects the nuclei; however, it does not improve the coarse segmentation and extracts the textural features from neighborhoods with a high density of nuclei. Therefore, the impact of segmentation errors on the features extracted by COMPASS has been compensated to some extent. A recent fully automatic algorithm based on textural features was proposed in [12]. It achieved CCRs of 65.22%, 90.09%, and 69.23% for the three nuclear grades and a Cohen's kappa of 0.6123 (substantial agreement) on the same images that we used here. By taking advantage of scores from pathologists and restricting the analysis to the areas with high nuclear density, COMPASS obtained average (across three junior pathologists) CCRs of 76.1%, 96.0%, and 89.1% for the three nuclear grades and a Cohen's kappa of 0.83 (almost perfect agreement). As shown in Table 5, the scores given by RM1 (which relies on features from pathologist's assessment) achieved an average AUC of 77% for detecting high-grade cases. This shows the higher discriminative ability of scores given by the pathologists in detecting high-grade images compared to low-grade images (average AUC of 72%).

Cohen's kappa ranged from 0.80 to 0.86 for different pathologists and different scanners. The values were comparable to the Cohen's kappa for senior pathologists while assessing the same dataset. Figure 5 shows the percentage of images on which COMPASS agreed with each one of the senior pathologists. It should be noted that even two senior pathologists did not agree with each other on all images. Their agreement rate was 74%, 84%, and 69% for the three nuclear grades. Estimating the internal parameters of COMPASS was computationally expensive, and this procedure should be done for each pathologist separately. Specifically, this was due to the fact that we set the hyperparameters of COMPASS by using Bayesian optimization and repeated the late decision fusion step five times to avoid partitioning noise. However, the training step should be done only once and after that COMPASS can be used to assess new images.

The study has a number of limitations. First, the prevalence of different grades in the dataset was different from real clinical practice. Therefore, the agreement rate reported here and Cohen's kappa will change if the class proportions change. However, having a balanced dataset may improve the CCRs of the grades 1 and 3, as more data will be available for training the model. Here, we dealt with the class imbalance problem by adopting the SMOTE [24] algorithm for upsampling of minority classes (i.e., grades 1 and 3); nonetheless, COMPASS could benefit from larger sample size. Moreover, Mitosis-Atypia database is a relatively small database and only included 300 images. The main advantages of using this database were being publicly available, providing high-quality ground truth based on the consensus of three pathologists, and including six atypia-related into the database. Using publicly available databases makes our study comparable to the previous studies in the literature and facilitates replicating our work in future studies. However, it should be noted that the sample size is relatively small, and availability of certain data types was limited in the database. Further validation of this preliminary study on a larger dataset will be required in the future. In addition, with a larger dataset, the performance of deep-learning-based features to extract textural characteristics of tissue can be also investigated. In recent years, deep learning has been used for various tasks in breast pathology such as benign/malignant classification [28], detecting mitotic figures [29], and determining different cancer subtypes [30], and very promising results have been obtained. Therefore, using deep features as an input of RM2 can be a potential avenue for future work. Secondly, intra-pathologist variability in scoring six atypia-related criteria should be investigated. Thirdly, the reported results were based on 300 images from 11 patients. Here we used leave-one-image-out cross validation to evaluate the performance of COMPASS. However, the results would be more realistic if the test images were from different patients. In the publicly available challenge dataset, 124 test images from different patients were provided; however, the junior pathologists did not asses those images. Hence, COMPASS could not be used for grading them.

Fourthly, as the ultimate goal of breast cancer grading is utilizing it as a prognostic factor in patient management, investigating the association between the nuclear grade outputted by COMPASS and patient survival would strengthen the study. Relating COMPASS's output to patients' prognosis could be a future step of this study. Also, the internal parameters of COMPASS are estimated based on the current performance of the junior pathologist. However, the scores given by the junior pathologists could change as they gain experience. Therefore, the parameters of the model should be updated on a regular basis. Investigating paradigm for updating the parameters and algorithmic considerations (e.g., whether the hyperparameters should be updated or not) could be a possible avenue for future work.

Moreover, in the publicly available database utilized in this study, information regarding the number of years since board examination of senior and junior pathologists and viewing condition of images were not provided. Based on the provided description [15], the pathologists have been recruited from Pitie-Salpetriere Hospital, Paris, in 2014 and were categorized as senior or junior. Investigating the added benefit of COMPASS based on expertise level of pathologist could be a possible future work.

Finally, COMPASS was tested retrospectively, and we assumed that the junior pathologists would accept the nuclear grade given by COMPASS. However, in a more realistic setup, the junior pathologists would score six atypia-related criteria and then COMPASS would combine these scores with the computer-extracted textural features using previously trained non-linear regression models and output the nuclear grade to the junior pathologists, who would assign the final nuclear grade to the image.

As COMPASS is a personalized tool, in order to use COMPASS in practice, the parameters of three regression models should be first estimated at the individual level in a training phase. Therefore, a database of approximately 300 images (i.e., the size of our database) with known nuclear atypia grade will be required. For each image, the pathologist, who will be using COMPASS, should score six atypia-related criteria on four ×40 magnification factor. After the training phase for personalizing COMPASS to each individual pathologist, the tool is ready to be used. Moreover, re-estimating parameters (re-calibrating COMPASS) might be required as junior pathologists are gaining more expertise as it is well known in the discipline of medical image perception that perceptual skills [31, 32] and error-making patterns [33, 34] change with expertise development. For a new test image and during the training phase, textural features should be extracted from ×20 magnification level as RM2 was previously trained based on this magnification factor. For six nuclear-atypia-related scores, in the database utilized here, scores were given based on ×40 magnification level. However, one possible future work could be investigating the effect of different magnification factor on pathologist's perception of six atypia-related criteria and evaluating differences among pathologists in terms of optimal magnification for scoring these criteria. In the utilized database, it was argued that detailed features of nuclei might not be visible at ×20 magnification level.

In summary, COMPASS, which is a personalized tool, potentially can assist junior pathologists in nuclear grading of breast cancer and achieved a performance that was comparable to that of the senior pathologists. This study has also demonstrated that COMPASS, if it had been adopted by the junior pathologists, could play the role of the second reader, and it could also complement the senior pathologist's performance to some extent. The findings also underscore the importance of textural computer-extracted features to supplement the junior pathologist's assessment of the case.

Acknowledgments We would like to thank Mitosis-Atypia challenge organizers who collected the data utilized in the study and kindly provided us the access to their dataset after the challenge time. We also acknowledge the University of Sydney HPC Service at the University of Sydney for providing the high-performance computing resources that have contributed to the research results reported within this paper.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Mook S, Schmidt MK, Rutgers EJ, van de Velde AO, Visser O, Rutgers SM, Armstrong N, van't Veer LJ, Ravdin PM: Calibration and discriminatory accuracy of prognosis calculation for breast cancer with the online adjuvant! Program: A hospital-based retrospective cohort study. Lancet Oncol 10(11):1070–1076, 2009
2. Elston CW, Ellis IO: 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): 403–410, 1991
3. Bueno-de-Mesquita JM, Nuyten D, Wesseling J, van Tinteren H, Linn S, van De Vijver M: The impact of inter-observer variation in pathological assessment of node-negative breast cancer on clinical risk assessment and patient selection for adjuvant systemic treatment. Ann Oncol 21(1):40–47, 2009
4. Frierson, Jr HF, Wolber RA, Berean KW, Franquemont DW, Gaffey MJ, Boyd JC, Wilbur DC: Interobserver reproducibility of the Nottingham modification of the bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma. Am J Clin Pathol 103(2):195–198, 1995
5. Harvey JM, de Klerk NH, Sterrett GF: Histological grading in breast cancer: Interobserver agreement, and relation to other prognostic factors including ploidy. Pathology 24(2):63–68, 1992
6. Longacre TA, Ennis M, Quenneville LA, Bane AL, Bleiweiss IJ, Carter BA, Catelano E, Hendrickson MR, Hibshoosh H, Layfield LJ: Interobserver agreement and reproducibility in classification of invasive breast carcinoma: An NCI breast cancer family registry study. Mod Pathol 19(2):195–207, 2006
7. Meyer JS, Alvarez C, Milikowski C, Olson N, Russo I, Russo J, Glass A, Zehnbauer BA, Lister K, Parwaresch R: Breast carcinoma malignancy grading by Bloom–Richardson system vs proliferation index: Reproducibility of grade and advantages of proliferation index. Mod Pathol 18(8):1067–1078, 2005
8. Paradiso A, Ellis I, Zito F, Marubini E, Pizzamiglio S, Verderio P: Short-and long-term effects of a training session on pathologists' performance: The INQAT experience for histological grading in breast cancer. J Clin Pathol 62(3):279–281, 2009
9. Adams AL, Chhieng DC, Bell WC, Winokur T, Hameed O: Histologic grading of invasive lobular carcinoma: Does use of a 2-tiered nuclear grading system improve interobserver variability? Ann Diagn Pathol 13(4):223–225, 2009
10. Gandomkar Z, Brennan PC, Mello-Thoms C: Computer-based image analysis in breast pathology. J Pathol Inform 7:43, 2016
11. Cosatto E, Miller M, Graf HP, Meyer JS. Grading nuclear pleomorphism on histological micrographs. InPattern Recognition, 2008. ICPR 2008. 19th International Conference on 2008 Dec 8 (pp. 1-4). IEEE.
12. Khan AM, Sirinukunwattana K, Rajpoot N: A global covariance descriptor for nuclear atypia scoring in breast histopathology images. IEEE J Biomed Health Inform 19(5):1637–1647, 2015
13. Dunne B, Going J: Scoring nuclear pleomorphism in breast cancer. Histopathology 39(3):259–265, 2001
14. Zhang R, Chen H-j, Wei B, Zhang H-y, Pang Z-g, Zhu H, Zhang Z, Fu J, Bu H: Reproducibility of the Nottingham modification of the Scarff-Bloom-Richardson histological grading system and the complementary value of Ki-67 to this system. Chin Med J (Engl Ed) 123(15):1976, 2010
15. Racoceanu D, Capron F: Semantic integrative digital pathology: Insights into microsemiological semantics and image analysis scalability. Pathobiology 83(2–3):148–155, 2016
16. Saha K, Raychaudhuri G, Chattopadhyay BK, Das I: Comparative evaluation of six cytological grading systems in breast carcinoma. J Cytol 30(2):87–93, 2013
17. Abati A, McKee G: Grading of breast carcinoma in fine-needle aspiration cytology. Diagn Cytopathol 19(2):153–154, 1998
18. Robinson I, McKee G, Kissin M: Typing and grading breast carcinoma on fine-needle aspiration: Is this clinically useful information? Diagn Cytopathol 13(3):260–265, 1995
19. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, Schmitt C, and Thomas NE, A method for normalizing histology slides for quantitative analysis. pp. 1107–1110
20. Al-Kofahi Y, Lassoued W, Lee W, Roysam B: Improved automatic detection and segmentation of cell nuclei in histopathology images. IEEE Trans Biomed Eng 57(4):841–852, 2010
21. Irshad H: Automated mitosis detection in histopathology using morphological and multi-channel statistics features. J Pathol Inform 4:10, 2013
22. Gandomkar Z, Brennan PC, Mello-Thoms C: Determining image processing features describing the appearance of challenging mitotic figures and miscounted nonmitotic objects. J Pathol Inform 8:34, 2017
23. Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N: Taking the human out of the loop: A review of Bayesian optimization. Proc IEEE 104(1):148–175, 2016
24. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 16: 321–357, 2002
25. Viera AJ, Garrett JM: Understanding interobserver agreement: The kappa statistic. Fam Med 37(5):360–363, 2005
26. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1): 29–36, 1982
27. Mirza AN, Mirza NQ, Vlastos G, Singletary SE: Prognostic factors in node-negative breast cancer: A review of studies with sample size more than 200 and follow-up more than 5 years. Ann Surg 235(1):10–26, 2002
28. Gandomkar Z, Brennan PC, Mello-Thoms C. A framework for distinguishing benign from malignant breast histopathological images using deep residual networks. In14th International Workshop on Breast Imaging (IWBI 2018), International Society for Optics and Photonics, Vol. 10718, p. 107180U, 2018.
29. Cireșan DC, Giusti A, Gambardella LM, and Schmidhuber J, Mitosis detection in breast cancer histology images with deep neural networks. pp. 411–418
30. Gandomkar Z, Brennan PC, Mello-Thoms C: MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. Artif Intell Med 88:14–24, 2018
31. Gandomkar Z, Tay K, Brennan PC, and Mello-Thoms C, A model based on temporal dynamics of fixations for distinguishing expert radiologists' scanpaths. p. 1013606
32. Gandomkar Z, Tay K, Brennan PC, Mello-Thoms C: Recurrence quantification analysis of radiologists' scanpaths when interpreting mammograms. Med Phys 45:3052–3062, 2018
33. Gandomkar Z, Tay K, Ryder W, Brennan PC, and Mello-Thoms C, Predicting radiologists' true and false positive decisions in reading mammograms by using gaze parameters and image-based features. p. 978715
34. Gandomkar Z, Tay K, Ryder W, Brennan PC, Mello-Thoms C: iCAP: An individualized model combining gaze parameters and image-based features to predict radiologists' decisions while Reading mammograms. IEEE Trans Med Imaging 36(5):1066–1075, 2017