Artificial Intelligence Improves the Accuracy in Histologic Classification of Breast Lesions

António Polónia, MD, PhD,[1,2] Sofia Campelos, MD,[1,2] Ana Ribeiro, MD,[3] Ierece Aymore, MD,[1,2] Daniel Pinto, MD,[4] Magdalena Biskup-Fruzynska, MD,[5] Ricardo Santana Veiga, MD,[6] Rita Canas-Marques, MD,[7] Guilherme Aresta, MEng,[8,9] Teresa Araújo, MEng,[8,9] Aurélio Campilho, PhD,[8,9] Scotty Kwok, MSc,[10] Paulo Aguiar, PhD,[2,11] and Catarina Eloy, MD, PhD[1,2,12]

From the [1]Department of Pathology, Ipatimup Diagnostics, Institute of Molecular Pathology and Immunology, University of Porto, Porto, Portugal; [2]I3S – Instituto de Investigação e Inovação em Saúde, University of Porto, Porto, Portugal; [3]Department of Pathology, Centro Hospitalar de Vila Nova de Gaia / Espinho, EPE, Vila Nova de Gaia, Portugal; [4]Department of Pathology, Centro Hospitalar de Lisboa Ocidental, EPE, Lisboa, Portugal; [5]Department of Tumor Pathology, Maria Sklodowska-Curie National Research Institute of Oncology (MSCNRIO), Gliwice, Poland; [6]Department of Surgical Pathology, Hospital da Luz Lisboa, Lisboa, Portugal; [7]Pathology Service, Champalimaud Clinical Center, Lisboa, Portugal; [8]INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal; [9]Faculty of Engineering, University of Porto, Porto, Portugal; [10]Sebit Company Limited, Sha Tin, Hong Kong; [11]Instituto Nacional de Engenharia Biomédica (INEB), Universidade do Porto, Porto, Portugal; and [12]Faculty of Medicine, University of Porto, Porto, Portugal.

Key Words: Artificial intelligence; Histology; Breast cancer; Computational pathology; Machine learning; Convolutional neural networks; Deep learning

Am J Clin Pathol April 2021;155:527-536

DOI: 10.1093/AJCP/AQAA151

ABSTRACT

Objectives: *This study evaluated the usefulness of artificial intelligence (AI) algorithms as tools in improving the accuracy of histologic classification of breast tissue.*

Methods: *Overall, 100 microscopic photographs (test A) and 152 regions of interest in whole-slide images (test B) of breast tissue were classified into 4 classes: normal, benign, carcinoma in situ (CIS), and invasive carcinoma. The accuracy of 4 pathologists and 3 pathology residents were evaluated without and with the assistance of algorithms.*

Results: *In test A, algorithm A had accuracy of 0.87, with the lowest accuracy in the benign class (0.72). The observers had average accuracy of 0.80, and most clinically relevant discordances occurred in distinguishing benign from CIS (7.1% of classifications). With the assistance of algorithm A, the observers significantly increased their average accuracy to 0.88. In test B, algorithm B had accuracy of 0.49, with the lowest accuracy in the CIS class (0.06). The observers had average accuracy of 0.86, and most clinically relevant discordances occurred in distinguishing benign from CIS (6.3% of classifications). With the assistance of algorithm B, the observers maintained their average accuracy.*

Conclusions: *AI tools can increase the classification accuracy of pathologists in the setting of breast lesions.*

Key Points

- Artificial intelligence algorithms can support evaluation as computer-aided diagnosis tools in the histologic classification of breast tissue
- The algorithm used in microscopic photographs achieved higher classification accuracy (0.87) than the average of pathologists (0.83)
- With the assistance of the algorithm, pathologists significantly increased their average accuracy to 0.90.

Digital pathology (DP) has been implemented in several pathology departments around the world.[1-6] One of the main advantages of using whole-slide images (WSI) is the potential for implementing computer-aided diagnosis (CAD) tools that may improve the evaluation of tissue morphology, both quantitatively and qualitatively, adding robustness to image diagnosis.[7] The subjectivity in the appreciation of the histologic features of breast pathology sometimes leads to lower than desired interobserver concordance rates, with borderline cases usually as the reasons for disagreements.[8-10] The misclassification of breast diseases may result in under- or overtreatment with important consequences to patients' health.

CAD tools can potentially provide a complementary and objective assessment of histologic features, improving both sensitivity and specificity of the pathologic diagnosis without increasing the workload of pathologists. In recent years, an explosion of studies have been published

reporting high accuracy levels of automatic classification of histology images by machine learning algorithms in several disease models.[11-18] In fact, the use of artificial intelligence in DP is seen as the third revolution of pathology, following the introduction of immunohistochemistry (IHC) and molecular pathology.[19]

Our group recently organized an image analysis challenge (Breast Cancer Histology [BACH]) as part of the 15th International Conference on Image Analysis and Recognition (ICIAR 2018) that aimed at the automatic classification of breast tissue histology using H&E-stained microscopy photographs and WSIs.[20] Remarkably, the best performing methods achieved performances similar to those of human experts.

The purpose of the current work is to compare the classification accuracy of the best algorithms of the BACH challenge with the accuracy of a larger group of human observers (including pathologists and pathology residents). In addition, we also aimed to assess whether the output from the algorithms could be used by the observers to improve their classification accuracy.

Materials and Methods

Characteristics of the Tested Images and Algorithms

The test data set of the BACH challenge was composed of 2 independent parts (A and B).[20] Test A consisted on 100 H&E-stained breast tissue microscopic photographs (from 38 patients) classified into 4 classes: normal, benign, carcinoma in situ (CIS), and invasive carcinoma (IC). Test B consisted of 152 regions of interest (ROI) in 10 H&E-stained WSIs of breast tissue (from 8 patients) classified into the same 4 classes described above. The cases included formalin-fixed, paraffin-embedded needle core biopsies and surgical excision specimens diagnosed between 2013 and 2017 originating from 2 histology laboratories (Ipatimup Diagnostics and Centro Hospitalar Universitário Cova da Beira). The ground truth (GT) was established by 2 pathologists (A.P. and C.E.) with glass slides and cases, with disagreements resolved through common microscopy sessions. IHC analysis was performed in all IC and CIS and some benign lesions (ductal hyperplasia, intraductal papilloma, sclerosing lesions, and fat necrosis). At the end of the study, no observer disagreed with the GT classifications. Each photograph and ROI included only 1 of the 4 classes, except for normal tissue that could be present with any other class. The characterization of the 4 classes in both tests is summarized in Table 1.

Photographs from test A were acquired with a Leica DM 2000 LED microscope and a Leica ICC50 HD camera with a ×20 objective and 0.4 numerical aperture originating RGB images in a TIFF format without compression, a size of 2,048 × 1,536 pixels (0.56 mm^2), and a pixel scale of 0.42 µm/pixel, without color normalization. WSIs from test B were acquired with a Leica SCN400 scanner with a ×20 objective (pixel scale of 0.47 µm/pixel). The irregularly shaped (freehand) ROIs had different sizes, varying from 0.04 to 171.19 mm^2 and a median of 0.49 mm^2.

The algorithms used for assessing the images of tests A and B were the ones that achieved the best performance on the BACH challenge's independent test set. Namely, for test A, we selected the method with the highest overall accuracy (measured as the ratio between the correct answers and the total number of photographs) and better accuracy on distinguishing between normal and nonnormal samples (algorithm A). Likewise, for test B, we selected the method with the highest classification performance (algorithm B).[20,21] Both algorithms rely on deep learning and were developed by the same participant using the training images of the BACH challenge. Algorithm A is based on a convolutional neural network that classifies patches of 299 × 299 pixels resized from patches of 1,495 × 1,495 pixels collected from the original image. For each case, overlapping patches are collected at a fixed distance interval, and the final label is produced by averaging

Table 1
Characterization of Tests A and B

Classes	Test A	Test B
Normal, No. (%)	25 (25)	31 (20.4)
Benign, No. (%)	25 (25)	65 (42.8)
Fibrocystic change	1	31
Inflammation	0	15
Columnar cell change	4	8
Microcalcification	0	5
Fibroadenoma	2	2
Ductal hyperplasia	1	3
Apocrine metaplasia	3	0
Fat necrosis	3	0
Atrophy	2	0
Intraductal papilloma	4	0
Sclerosing lesion	2	1
Adenosis	2	0
Secretory changes	1	0
Carcinoma in situ, No. (%)	25 (25)	33 (21.7)
Ductal	24	7
Lobular	1	26
Invasive carcinoma, No. (%)	25 (25)	23 (15.1)
Ductal	23	11
Lobular	1	2
Tubular	1	0
Mucinous	0	10
Total	100	152

DOI: 10.1093/ajcp/aqaa151

the patch-wise predictions. Algorithm B uses the same convolutional neural network to predict patch-wise classifications on the WSIs. Instead of averaging all predictions as in algorithm A, each pixel of the image is labeled as the average of the overlapping patch-wise predictions, creating a pixel-wise abnormality classification map (for additional details, see supplemental data; all supplemental data can be found at *American Journal of Clinical Pathology* online). For this study, a classification of the ROI was obtained if more than 95% of the classification map pixels shared the same label; if not, the 2 most frequent pixel classifications were used as the favorite and alternative classifications, respectively.

Evaluation Criteria of the Observers' Accuracy

The accuracy of 4 pathologists (P1 to P4) and 3 pathology residents (R1 to R3) were evaluated in different phases. P1 to P3 are generalist pathologists with 6, 6, and 42 years of practice, respectively. P4 is a subspecialist breast pathologist with 4 years of practice. In phase 1, the observers classified photographs from test A and ROIs overlaid in the entire WSI from test B into 4 classes, without exceptions. In cases of doubt between classes, observers could choose their favorite classification and provide the respective alternative.

In phase 2, the observers had the opportunity to reclassify the photographs and ROIs, knowing their initial classification and the one performed by the algorithms, without being aware of its accuracy or their own. Some rules of engagement were established (summarized in Table 2 and Supplemental Figure S1): if the observer classification matched the algorithm classification, it could not be changed; in this case, if there was an alternative classification (by the observer, the algorithm, or both), there was the possibility to keep or discard the alternative classification. If the participant classification did not match the algorithm classification, with or without the presence of an alternative classification, the observers could reclassify the photograph or ROI.

In test A, before phase 3, both confusion matrices of the observers and the algorithm were revealed, showing their global accuracy and the types of errors between different categories, without specifying the correct answer of each photograph. Then, observers performed the same task as in phase 2. Test B did not have a phase 3. The observers had no time constraint applied during the classification, and no washout period existed between the evaluation of different phases. All photographs and ROIs were classified before the next phase started. Each phase was performed in less than a week, and all phases were performed in less than a month. Photographs were reviewed with Windows Photo Viewer (Microsoft) and WSIs with Aperio ImageScope v12.3.2 (Leica Biosystems). The classification was recorded manually in a prefilled Excel sheet (Microsoft). None of the observers were involved in the establishment of the GT or received training by the pathologists responsible by the GT. Moreover, IHC information was not available for the classification of images in both tests in all phases, which was based on morphology alone. In addition, P4 did not participate in test B.

Ethics approval and informed consent were not required for this study, given the anonymized images of the samples.

Statistical Analysis

Statistical analyses were performed using SPSS version 25.0 for Windows (IBM). The Pearson χ^2 test (or the Fisher exact test, if appropriate) was used for comparison of qualitative variables, and the Mann–Whitney *U* test (MW), the Wilcoxon (WC) test, and the Kruskal-Wallis test were used for quantitative variables. The level of significance was set at $P < .05$. Accuracy was defined as the ratio between the correct answers and the total number of photographs or ROIs. Concordance rates were evaluated with quadratic weighted κ statistics to penalize discordances with higher clinical impact. The Landis and Koch classification was used to interpret the values: no agreement to slight agreement (<0.20), fair agreement (0.21-0.40), moderate agreement (0.41-0.60), substantial agreement (0.61-0.80), and excellent agreement (>0.81).[22]

Table 2
Summary of Tasks Developed in Different Tests and Phases

Tests and Phases	Tasks
Test A	100 H&E photographs
Test B	152 ROIs in 10 H&E WSIs
Phase 1[a]	Classification of tests A and B with 4 classes: normal, benign, carcinoma in situ and invasive carcinoma
Phase 2[b]	Same as in phase 1, knowing the classification of the algorithms
Phase 3 (test A only)	Same as in phase 2, knowing the accuracy of the algorithm and the observers and the types of errors between different classes

ROI, region of interest; WSI, whole-slide image.
[a]In cases of doubt between classes, observers should choose their favorite classification and provide the respective alternative.
[b]If the observer classification matched the algorithm classification, it could not be changed; if there was an alternative classification, there was the possibility of keeping or discarding the alternative classification. If the observer classification did not match the algorithm classification, observers could reclassify the photograph or ROI.

DOI: 10.1093/ajcp/aqaa151

Results

Test A

Accuracy of Algorithm A

Algorithm A had an accuracy of 0.87 (**Table 3**, Supplemental Table S1, and **Figure 1A**) and a concordance rate with the GT of 0.88 (Supplemental Figures S2A and S2B). The benign class had lower accuracy (0.72; 18/25) in comparison with the remaining classes (0.96 [24/25], 0.88 [22/25], and 0.92 [23/25] for normal, CIS, and IC, respectively; Fisher exact test, P = .02). Most discordances with GT occurred in distinguishing normal from benign (4%) and benign from IC (4%) (Supplemental Table S2). Fat necrosis was the benign lesion confused with IC **Image 1A** and **Image 1B**. The accuracy of the algorithm was 0.71 (10/14) in photographs correctly classified by less than 50% of the observers, increasing to 0.92 (60/65) in photographs correctly classified by more than 75% of the observers (χ^2, P = .03) **Figure 1C**.

Accuracy of the Observers

Phase 1.—The observers had an average accuracy of 0.80; only 1 pathologist had accuracy higher (P2, 0.94) than that obtained by the algorithm A (Table 3, Supplemental Table S1, and Figure 1A). The mean concordance rate between the observer's classification and the GT was 0.86 (range, 0.80-0.93), with 2 pathologists having concordance rates higher than algorithm A (P1 and P2: 0.93). In this phase, the mean interobserver concordance rate was 0.83 (range, 0.75-0.90) (Supplemental Figure S2A). IC was the class with higher accuracy (average, 0.95) in comparison to the other classes (average of 0.73, 0.78 and 0.76 for normal, benign and CIS, respectively; MW, P < .001). Most discordances with GT occurred in distinguishing normal from benign (7.4%) and benign from CIS (7.1%) (**Image 1C** and **Image 1D** and Supplemental Table S2).

The observers proposed an alternative classification in 14% of the photographs, with pathologists proposing more alternative classifications than residents (18.3% and 8.3%, respectively; MW, P = .002) (Supplemental Table S3). The most frequent alternative classifications were those between CIS and IC (4.7%), benign and IC (4.7%), and benign and CIS (4.3%).

Phases 2 and 3.—In phase 2, the observers increased their average accuracy from 0.80 to 0.85 (WC, P < .001), with 3 pathologists obtaining accuracies equal to or higher than algorithm A (Table 3, Supplemental Table S1, and Figure 1A). In phase 3, the observers had an additional increase in their average accuracy from 0.85 to 0.88 (WC, P = .001), with 3 pathologists and 2 residents with accuracies higher than algorithm A. In this last phase, the mean concordance rate between the observer´s classification and the GT increased from 0.86 to 0.91 (range, 0.85-0.99), with only 2 observers having concordance rates lower than algorithm A. In addition, the mean interobserver concordance rate increased from 0.83 to 0.90 (range, 0.83-0.95) (Supplemental Figure S2B). The accuracy increased in all classes (average of 0.86, 0.82, 0.89, and 0.97 for normal, benign, CIS, and IC, respectively). Most discordances with GT decreased and occurred in distinguishing normal from benign (5.4% and 4.7%) and benign from CIS (5.7% and 3.4%) in phases 2 and 3, respectively (Supplemental Table S2).

A similar proportion of alternative classifications was proposed by the observers in phase 2 compared with phase 1 (13.1% vs 14%, respectively; WC, P = .96), increasing in phase 3 in comparison with phase 2 (16.6% vs 13.1%, respectively; WC, P = .002) (Supplemental Table S3). The most frequent alternative classifications were those between benign and CIS (5.7%), benign and IC (4.6%), and normal and benign (4.1%).

The favorite classification was modified, on average, in 6.3% of the photographs in phase 2, increasing to 10.9% in phase 3 (WC, P < .001), with only 2 observers with less than 5% modifications in both phases (P3 and R1) (Supplemental Table S4). In addition, pathologists and residents had similar frequencies of modifications on their favorite classification (6.8% and 5.7% [MW, P = .24] for phase 2 and 9.5% and 12.7% [MW, P = .36] for phase 3). The alternative classification of the photographs was modified, on average, in 15.7% in phase 2, increasing to 19.3% in phase 3 (WC, P < .001) (Supplemental Table S4). In addition, pathologists had more frequent modifications than residents on their alternative classification (21.8% and 7.7% [MW, P < .001] for phase 2 and 23.8% and 13.3%

Table 3
Diagnostic Accuracy in Test A and B

	Phase	Test A	Test B
Algorithm A	1	0.87	NA
Algorithm B	1	NA	0.49
Pathologists (average)	1	0.83	0.83
	2	0.87	0.82
	3	0.90	NA
Residents (average)	1	0.77	0.89
	2	0.82	0.88
	3	0.87	NA
All observers (average)	1	0.80[a]	0.86
	2	0.85[a,b]	0.85
	3	0.88[b]	NA

NA, not applicable.
[a]Wilcoxon, P < .001
[b]Wilcoxon, P = .001.

DOI: 10.1093/ajcp/aqaa151

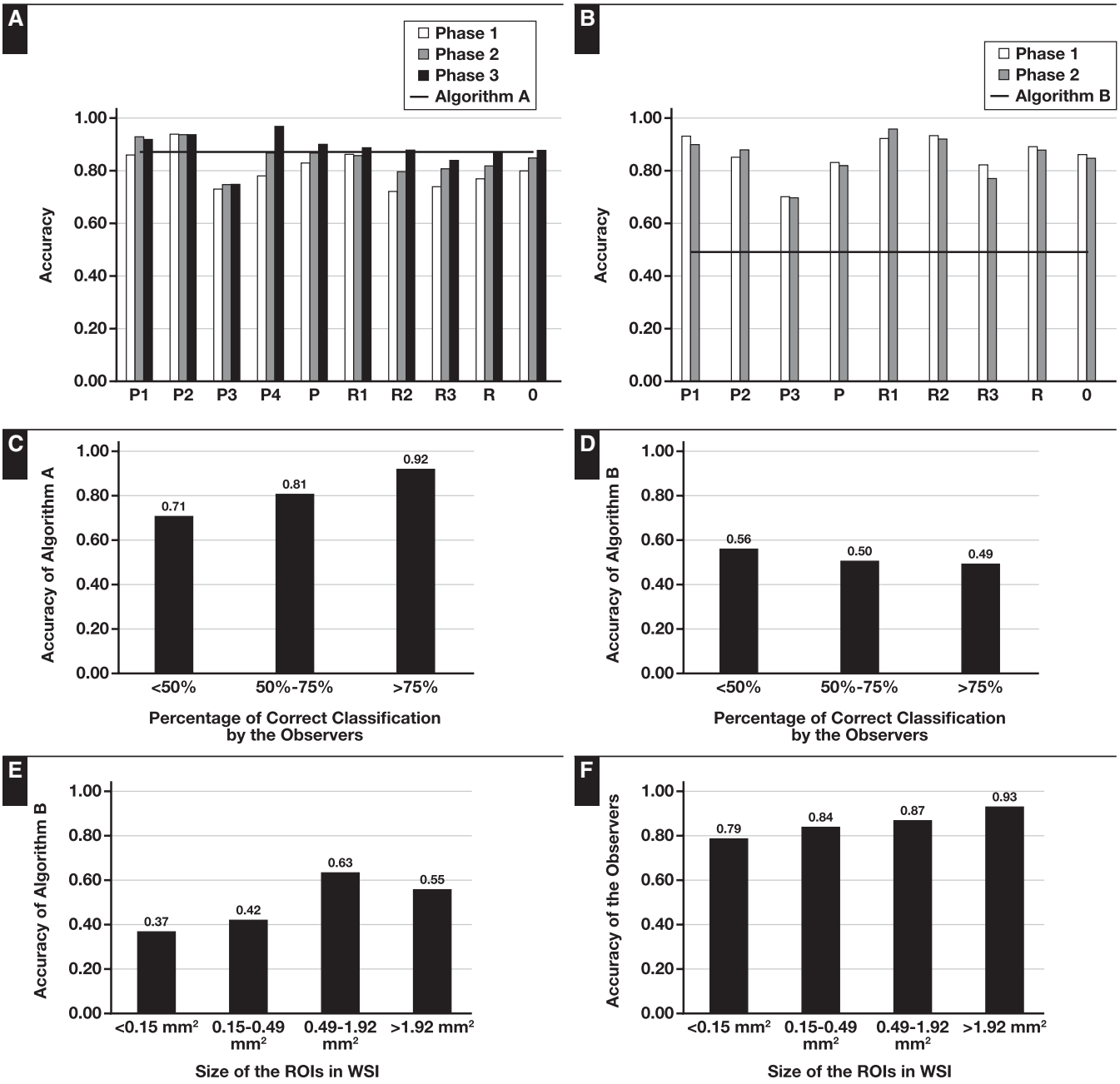

Figure 1 Classification accuracy of test A (**A**) and test B (**B**) in all phases. **C**, Accuracy of algorithm A in photographs correctly classified by <50% (10/14), 50%-75% (17/21) and >75% of the observers (60/65). **D**, Accuracy of algorithm B in ROIs correctly classified by <50% (5/9), 50%-75% (13/26), and >75% of the observers (57/117). **E**, Accuracy of algorithm B in ROIs <0.15 mm^2 (14/38), 0.15-0.49 mm^2 (16/38), 0.49-1.92 mm^2 (24/38), and >1.92 mm^2 (21/38). **F**, Average accuracy of the observers in ROIs <0.15, 0.15-0.49, 0.49-1.92, and >1.92 mm^2. The cutoffs used correspond to the 25th, 50th, and 75th percentiles of the size of the ROIs. O, average of all observers; P, average of pathologists; P1-P4, pathologists 1-4; R, average of residents; R1-R3, residents 1-3; ROI, region of interest; WSI, whole-slide image.

[MW, $P = .003$] for phase 3). The alternative classification had more frequent modifications than the favorite classification in both phases (15.7% and 6.3% in phase 2, and 19.3% and 10.9% in phase 3; WC, $P < .001$ for both phases).

Test B

Accuracy of Algorithm B

Algorithm B had accuracy of 0.49 (Table 3, Supplemental Table S5, and **Figure 1B**) and a concordance rate with the GT of 0.37 (Supplemental Figures

DOI: 10.1093/ajcp/aqaa151

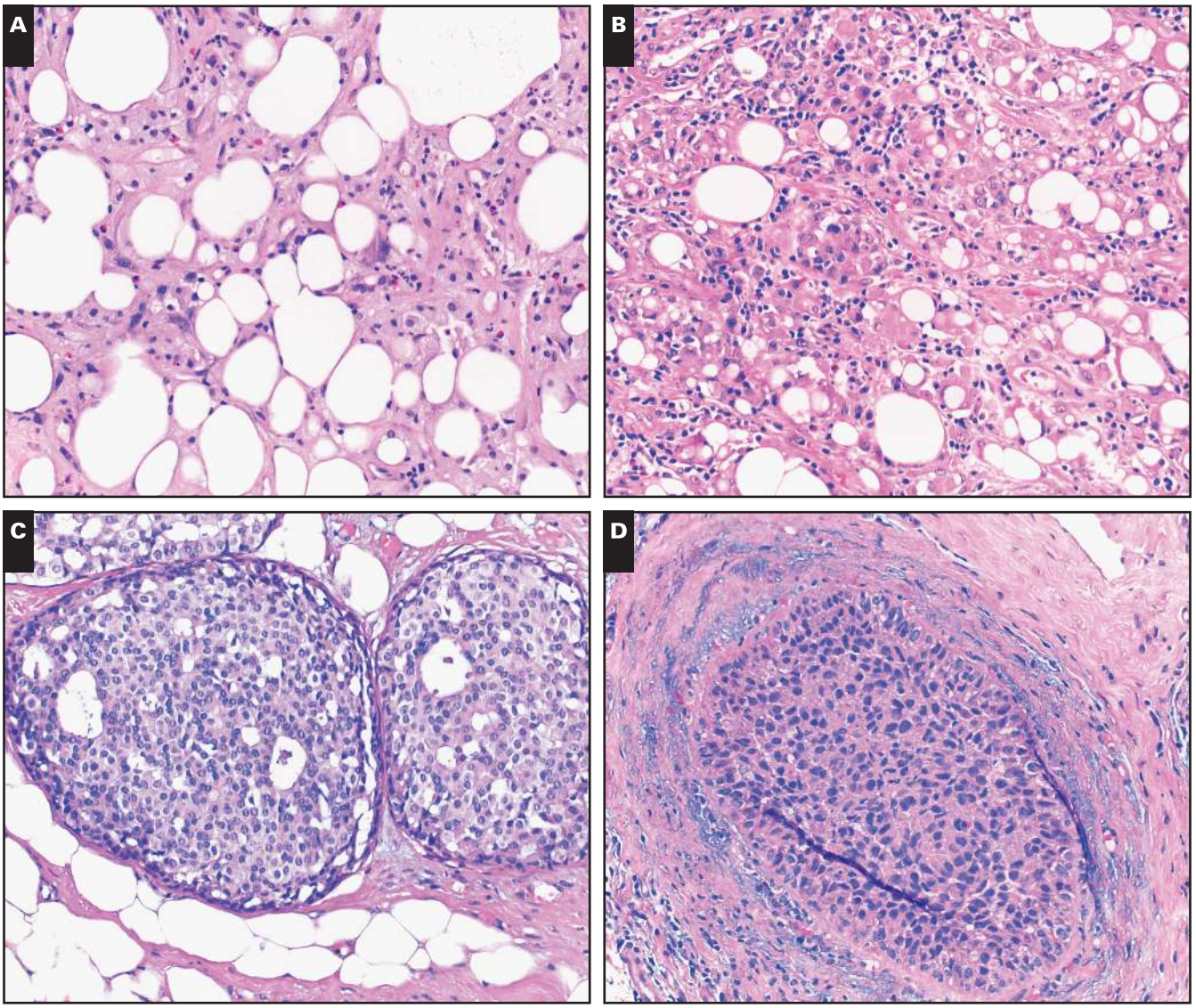

Image 1 Benign (fat necrosis) (H&E, ×200), correctly classified in phase 1 by 6 of 7 observers (**A**) and 3 of 7 observers (**B**), and as IC by algorithm A. **C**, DCIS (H&E, ×200), classified in phase 1 as benign by 4 of 7 observers, as DCIS by 3 of 7 observers, and as DCIS by algorithm A. **D**, DCIS (H&E, ×200), classified in phase 1 as DCIS by 4 of 7 observers, as benign by 3 of 7 observers, and as DCIS by algorithm A.

S2C and S2D). CIS was the class with lower accuracy (0.06; 2/33) in comparison to the remaining classes (0.84 [26/31], 0.58 [38/65], and 0.39 [9/23] for normal, benign, and IC, respectively; χ^2, $P < .001$). Most discordances with GT occurred in distinguishing normal from benign (15.1%), benign from IC (14.5%), and benign from CIS (11.8%) (Supplemental Table S6). Inflammation was the benign lesion confused with IC **Image 1E** and **Image 1F**. The accuracy of the algorithm was similar in ROIs correctly classified by more than 75% of the observers (0.49; 57/117) compared with ROIs correctly classified by less than 50% of the observers (0.56; 5/9; Fisher exact test, $P = .74$) **Figure 1D**. Moreover, the accuracy of the algorithm was lower in ROIs smaller than 0.49 mm^2 (0.39; 30/76) in comparison to larger ROIs (0.59; 45/76; χ^2, $P = .02$) **Figure 1E**.

Algorithm B proposed an alternative classification in 53.9% of the ROIs. The most frequent alternative classifications were those between normal and benign (37.5%), and between benign and IC (7.9%) (Supplemental Table S7).

Accuracy of the Observers

Phase 1.—The observers had average accuracy of 0.86, with all observers with accuracies higher than those obtained by the algorithm B (Table 3, Supplemental

DOI: 10.1093/ajcp/aqaa151

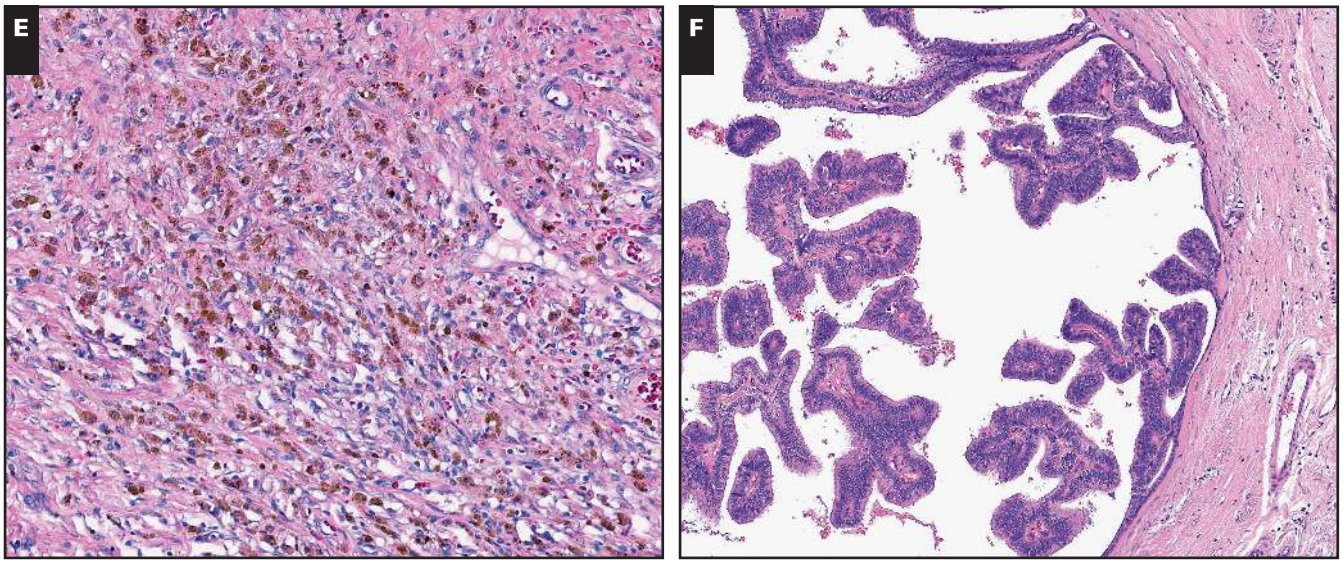

Image 1 (cont) **E**, Benign (inflammation) (H&E, ×100), correctly classified in phase 1 by 6 of 6 observers and as IC by algorithm B. **F**, DCIS (H&E, ×100), correctly classified in phase 1 by 4 of 6 observers and as benign by algorithm B. DCIS, ductal carcinoma in situ; IC, invasive carcinoma.

Table S5, and Figure 1B). The mean concordance rate between the observer´s classification and the GT was 0.91 (range, 0.84-0.95), with all observers having concordance rates higher than algorithm B. In this phase, the mean interobserver concordance rate was 0.87 (range, 0.79-0.96) (Supplemental Figure S2C). IC was the class with higher accuracy (average, 0.99) in comparison to the other classes (average of 0.84, 0.85, and 0.81 for normal, benign, and CIS, respectively; MW, $P < .001$). Most discordances with GT occurred in distinguishing normal from benign (7.4%) and benign from CIS (6.3%) (Image 1F and Supplemental Table S6). The average accuracy of the observers increased from 0.79 in ROIs smaller than 0.15 mm^2 to 0.93 in ROIs larger than 1.92 mm^2 (Kruskal-Wallis, $P = .001$) **Figure 1F**.

The observers proposed an alternative classification in 5% of the ROIs, with pathologists and residents proposing similar alternative classifications (4.4% and 5.7%, respectively; MW, $P = .34$) (Supplemental Table S7). The most frequent alternative classifications were those between benign and CIS (2.3%).

Phase 2.—The observers had similar accuracy (average of 0.85) in comparison to phase 1 (WC, $P = .96$) (Table 3 and Supplemental Table S5, and Figure 1B). The mean concordance rate between the observer´s classification and the GT maintained at 0.91 (range, 0.84-0.98), and the mean interobserver concordance rate was 0.87 (range, 0.79-0.94) (Supplemental Figure S2D). All classes maintained their accuracy (average of 0.86, 0.85, 0.76, and 0.99 for normal, benign, CIS, and IC, respectively). Most discordances with GT occurred in distinguishing normal from benign (7.7%) and benign from CIS (6.5%) (Supplemental Table S6). The favorite classification was modified, on average, in 4.6% of the ROIs, with pathologists and residents showing similar frequencies of modifications (3.7% and 5.5%, respectively; MW, $P = .18$) (Supplemental Table S8).

The observers increased the alternative classification from 5% to 15.6% of the ROIs (phase 1 vs 2; WC, $P < .001$), with residents proposing more alternative classifications than pathologists (22.2% and 9.0%, respectively; MW, $P < .001$) (Supplemental Table S7). The most frequent alternative classifications were those between benign and CIS (8.8%). The alternative classification of the ROIs was modified, on average, in 13.7%, with pathologists showing a lower frequency of modifications than residents (8.7% and 18.8%, respectively; MW, $P < .001$) (Supplemental Table S8). The alternative classification had more frequent modifications than the favorite classification (13.7% and 4.6%, respectively; WC, $P < .001$).

Discussion

Image fidelity in the computer display has been a major concern regarding digital diagnosis, an issue previously addressed by digital radiologists.[23,24] Systematic reviews have been performed to evaluate the concordance of pathologic diagnoses by WSIs in comparison to

DOI: 10.1093/ajcp/aqaa151

traditional light microscope (LM), revealing mean diagnostic concordance higher than 90%. This result demonstrates that DP can be used for primary diagnosis, provided that current best practice recommendations are followed.[25,26] In this work, the images used in both tests had resolutions near 0.5 µm/pixel, comparable to an LM, and excellent mean interobserver concordance rates were achieved in both tests.[27] We recognize that microscopic photographs are not the method for pathology diagnosis, as shown by the higher classification accuracies after the observation of ROIs in WSIs. Nevertheless, the similar accuracy achieved by the pathologists in both tests reveals that the photographs contain enough information to simulate clinical practice. In future studies, we would like to measure the role of AI algorithm outputs in the classification of WSIs without the use of ROIs.

The accuracy of algorithm A (photographs) was higher than the average accuracy of the observers, including the average accuracy of the pathologists, with excellent agreement with the GT. This result indicates that it is possible to develop an algorithm with the ability to perform a complex task, such as medical image interpretation or diagnosis, at an expert level. However, it had lower accuracy in classifying the photographs that observers classified correctly less frequently, indicating its limitation in assisting pathologists in classifying difficult cases. In the future, the accuracy for diagnosis of difficult cases may eventually be increased if the training sets of these types of algorithms are enriched in such cases. In real life, extraordinary cases without established GT will almost always need the intervention of an expert pathologist. This reinforces the idea that CAD tools will not replace pathologists in the future but probably will originate a trend of superspecialization to solve those difficult cases.

In contrast, the accuracy of algorithm B (WSIs) was lower than that of all observers for all classes, with only fair agreement with the GT. In addition, the algorithm had a large performance drop for ROIs smaller than 0.49 mm^2, given that it was trained to predict patches of approximately 0.50 mm^2 with consideration of the classification of the neighboring patch. When predicting a patch smaller than the training size, the nonrelevant classification of neighboring patches will have a greater effect on the classification of the patch, lowering the performance of the algorithm. A possible approach to improve the sensitivity of the algorithm would be to change the decision rule for the overlapping patches (eg, from average to local maximum) to increase the importance of these small regions in the final WSI labeling. Smaller lesions will probably continue to be a challenge for both pathologists and image analysis algorithms. The use of ROIs allowed direct comparison of the accuracy of the observers and the algorithm in precise regions, even small ones, given that the observers were forced to classify all ROIs, without exception.

Both algorithms had problems in the classification of benign lesions, usually showing difficulties in distinguishing benign from CIS (a known pitfall in LM diagnosis) and benign from IC, demonstrated by the recurrent misclassification of fat necrosis and inflammation as IC. Benign lesions have higher morphologic variability, making discriminant features more difficult to learn and lowering accuracy. These algorithms are probably learning that inflammation associated with some ICs is a typical characteristic of IC; this learning could give rise to a false-positive diagnosis, suggesting that these tools must be human supervised. We also recognize that a limitation of this study was the use of only 4 classes when performing the classification task. These classes do not cover all categories of breast lesions or the low number of patients who do not represent the wide morphologic pattern variation observed in real practice. However, we wanted to establish a proof of concept that artificial intelligence could be useful in DP diagnosis using the most common classes in breast pathology.

In our study, the observers had average accuracy higher in WSIs than in the photographs for all classes. This fact could be explained by the larger size of the ROIs, with more morphologic features to reach the correct classification, and the presence of adjacent context outside the ROIs in the WSIs. As expected, IC was the class more often correctly classified by the observers in both tests, which reflects the training and ability in detecting this relevant clinical lesion. The absence of a washout period between the evaluation of different phases had the objective of removing the intraobserver variability in the following phases and measuring only the impact of the algorithms in the change of the classification by the observers. The rules of engagement, which prevented changes when the classification of the observers matched the classification of the algorithms, had the purpose of simulating the future situation of the pathologist having access to the output of the algorithm and confirm or exclude their own classifications. Although the impact of revisiting the cases without AI assistance was not measured in this work, we estimate it to be low, given that the observers in test B did not improved their accuracy in phase 2.

The assistance provided by algorithm A significantly increased the average accuracy of the observers (in all classes) and the mean interobserver concordance rate, suggesting that CAD tools may be used to increase classification accuracy and homogeneity in pathology, even in important differential diagnostic problems, such as those between benign and CIS.

DOI: 10.1093/ajcp/aqaa151

In this work, we show that the recognition of CIS by the observers was suboptimal in both tests, even when shown directly to the observers in either microscopic photographs or ROIs in WSIs. The identification of CIS has been shown to be underdiagnosed, despite being clinically relevant and identifying patients who usually need surgical treatment and close follow-up due to increase risk of developing IC.[8-10] Importantly, the CIS classification accuracy was substantially increased with the support of algorithm A, showing that CAD tools can close the gap of false-negative results and ultimately contribute to increased patient health.

In test B, the excellent mean concordance rate between observers and the GT was maintained throughout the phases, meaning that an algorithm with a lower accuracy than that of the observers did not jeopardize their accuracy. In this case, the algorithm was not providing a credible alternative classification but rather keeping the observers faithful to their initial classification. However, in the last phase of test B, more alternative classifications were proposed, specially by residents with less experience than the pathologist, letting algorithm B work as a confusion generator. These results suggest that only CAD tools with the high accuracy should be implemented for clinical use.

Awareness of the accuracy and types of errors of algorithm A in phase 3 allowed measurement of its effect in the observers. This awareness translated into a higher proportion of changes in the classification of the photographs and in more alternative classifications, particularly from pathologists who took more advantage from the algorithm, even surpassing the accuracy of the algorithm. This effect points to the concept that better classification accuracy is achieved when both algorithm and observers work together rather than alone, producing a synergic effect. Phase 3 in test B was not performed because the observers would never consider an algorithm output with lower accuracy than their own.

We are aware that the use of IHC, as part of the daily practice in pathology, could have a positive impact on the observer´s accuracy. IHC was not available to the observers, representing a limitation of this work. Nevertheless, one of the goals was to test whether CAD tools could improve the observer´s accuracy on H&E.

Interestingly, in test A, there were two observers making less than 5% modifications of the favorite classification in both phases. These "nonbelievers" were the ones with concordance rates with GT lower than algorithm A in the last phase, indicating that CAD tools may have different impacts on different types of observers. Moreover, pathologists more often changed the alternative classification with algorithm A than with algorithm B, and residents more often changed the alternative classification with algorithm B than with algorithm A, suggesting that the higher experience of pathologists may have a role in determining how far they let the use of a CAD tool influence their final classification.

Conclusions

To our knowledge, this study represents the first time that machine learning algorithms in DP have been used to measure their impact in the classification accuracy of pathologists and pathology residents. We demonstrate that such CAD tools can increase the classification accuracy in the setting of breast lesions, providing the basis for its future clinical implementation with supervision.

Corresponding author: António Polónia, MD, PhD; antoniopolonia@yahoo.com.

Guilherme Aresta is funded by the Fundação para a Ciência e a Tecnologia (FCT) grant contract SFRH/BD/120435/2016. Teresa Araujo is funded by the FCT grant contract SFRH/BD/122365/2016. The research of Aurélio Campilho is financed by National Funds through the Portuguese funding agency, FCT, as part of project UID/EEA/50014/2013.

References

1. Stathonikos N, Veta M, Huisman A, et al. Going fully digital: perspective of a Dutch academic pathology lab. *J Pathol Inform.* 2013;4:15.
2. Thorstenson S, Molin J, Lundström C. Implementation of large-scale routine diagnostics using whole slide imaging in Sweden: digital pathology experiences 2006-2013. *J Pathol Inform.* 2014;5:14.
3. Snead DR, Tsang YW, Meskiri A, et al. Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology.* 2016;68:1063-1072.
4. Cheng CL, Azhar R, Sng SH, et al. Enabling digital pathology in the diagnostic setting: navigating through the implementation journey in an academic medical centre. *J Clin Pathol.* 2016;69:784-792.
5. Hartman DJ, Pantanowitz L, McHugh JS, et al. Enterprise implementation of digital pathology: feasibility, challenges, and opportunities. *J Digit Imaging.* 2017;30:555-560.
6. Araújo ALD, Arboleda LPA, Palmier NR, et al. The performance of digital microscopy for primary diagnosis in human pathology: a systematic review. *Virchows Arch.* 2019;474:269-287.
7. Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep.* 2016;6:26286.
8. Palli D, Galli M, Bianchi S, et al. Reproducibility of histological diagnosis of breast lesions: results of a panel in Italy. *Eur J Cancer.* 1996;32A:603-607.

DOI: 10.1093/ajcp/aqaa151

9. Wells WA, Carney PA, Eliassen MS, et al. Statewide study of diagnostic agreement in breast pathology. *J Natl Cancer Inst.* 1998;90:142-145.
10. Elmore JG, Longton GM, Carney PA, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA.* 2015;313:1122-1132.
11. Arevalo J, Cruz-Roa A, Arias V, et al. An unsupervised feature learning framework for basal cell carcinoma image analysis. *Artif Intell Med.* 2015;64:131-145.
12. Araújo T, Aresta G, Castro E, et al. Classification of breast cancer histology images using convolutional neural networks. *PLoS One.* 2017;12:e0177544.
13. Sharma H, Zerbe N, Klempert I, et al. Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. *Comput Med Imaging Graph.* 2017;61:2-13.
14. Fondón I, Sarmiento A, García AI, et al. Automatic classification of tissue malignancy for breast carcinoma diagnosis. *Comput Biol Med.* 2018;96:41-51.
15. Arvaniti E, Fricker KS, Moret M, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci Rep.* 2018;8:12054.
16. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med.* 2018;24:1559-1567.
17. Nir G, Hor S, Karimi D, et al. Automatic grading of prostate cancer in digitized histopathology images: learning from multiple experts. *Med Image Anal.* 2018;50:167-180.
18. Downing MJ, Papke DJ Jr, Tyekucheva S, et al. A new classification of benign, premalignant, and malignant endometrial tissues using machine learning applied to 1413 candidate variables. *Int J Gynecol Pathol.* 2020;39:333-343.
19. Salto-Tellez M, Maxwell P, Hamilton P. Artificial intelligence—the third revolution in pathology. *Histopathology.* 2019;74:372-376.
20. Aresta G, Araújo T, Kwok S, et al. BACH: grand challenge on breast cancer histology images. *Med Image Anal.* 2019;56:122-139.
21. Kwok S. Multiclass classification of breast cancer in whole-slide images. In: Campilho A, Karray F, ter Haar Romeny B, eds. *Image Analysis and Recognition.* Cham, Switzerland: Springer; 2018.
22. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159-174.
23. Brettle DS, Bacon SE. Short communication: a method for verified access when using soft copy display. *Br J Radiol.* 2005;78:749-751.
24. Krupinski EA, Kallergi M. Choosing a radiology workstation: technical and clinical considerations. *Radiology.* 2007;242:671-682.
25. Goacher E, Randell R, Williams B, et al. The diagnostic concordance of whole slide imaging and light microscopy: a systematic review. *Arch Pathol Lab Med.* 2017;141:151-161.
26. Pantanowitz L, Sinard JH, Henricks WH, et al; College of American Pathologists Pathology and Laboratory Quality Center. Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Arch Pathol Lab Med.* 2013;137:1710-1722.
27. Ho J, Parwani AV, Jukic DM, et al. Use of whole slide imaging in surgical pathology quality assurance: design and pilot validation studies. *Hum Pathol.* 2006;37:322-331.

DOI: 10.1093/ajcp/aqaa151