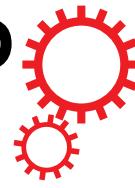


SCIENTIFIC REPORTS



OPEN

Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent

Received: 07 June 2016

Accepted: 20 March 2017

Published: 18 April 2017

Angel Cruz-Roa^{1,2}, Hannah Gilmore³, Ajay Basavanhally⁴, Michael Feldman⁵, Shridar Ganesan⁶, Natalie N.C. Shih⁵, John Tomaszewski⁷, Fabio A. González¹ & Anant Madabhushi⁸

With the increasing ability to routinely and rapidly digitize whole slide images with slide scanners, there has been interest in developing computerized image analysis algorithms for automated detection of disease extent from digital pathology images. The manual identification of presence and extent of breast cancer by a pathologist is critical for patient management for tumor staging and assessing treatment response. However, this process is tedious and subject to inter- and intra-reader variability. For computerized methods to be useful as decision support tools, they need to be resilient to data acquired from different sources, different staining and cutting protocols and different scanners. The objective of this study was to evaluate the accuracy and robustness of a deep learning-based method to automatically identify the extent of invasive tumor on digitized images. Here, we present a new method that employs a convolutional neural network for detecting presence of invasive tumor on whole slide images. Our approach involves training the classifier on nearly 400 exemplars from multiple different sites, and scanners, and then independently validating on almost 200 cases from The Cancer Genome Atlas. Our approach yielded a Dice coefficient of 75.86%, a positive predictive value of 71.62% and a negative predictive value of 96.77% in terms of pixel-by-pixel evaluation compared to manually annotated regions of invasive ductal carcinoma.

Detection of tumor cells in a histologic section is the first step for the pathologist when diagnosing breast cancer (BCa). In particular, tumor delineation from background uninvolved tissue is a necessary prerequisite for subsequent tumor staging, grading and margin assessment by the pathologist¹. However, precise tumor detection and delineation by experts is a tedious and time-consuming process, one associated with significant inter- and intra-pathologist variability in diagnosis and interpretation of breast specimens^{2–6}.

Invasive breast cancers are those that spread from the original site (either the milk ducts or the lobules) into the surrounding breast tissue. These comprise roughly 70% of all breast cancer cases^{7,8}, and they have poorer prognosis compared to the *in-situ* sub-types⁷. Isolation of invasive breast cancer allows for further analysis of tumor differentiation via the Bloom-Richardson and Nottingham grading schemes, which estimate cancer aggressiveness by evaluating histologic characteristics including: tubule formation, nuclear pleomorphism and mitotic count¹. Therefore, an automated and reproducible methodology for detection of invasive breast cancer on tissue slides could potentially reduce the total amount of time required to diagnose a breast case and reduce some of this inter- and intra-observer variability^{9,10}.

¹Universidad Nacional de Colombia, Bogota, Colombia. ²Universidad de los Llanos, Villavicencio, Colombia.

³University Hospitals Case Medical Center, Cleveland, OH, USA. ⁴Inspirata Inc., Tampa, FL, USA. ⁵Hospital of the University of Pennsylvania, Philadelphia, PA, USA. ⁶Cancer Institute of New Jersey, New Brunswick, NJ, USA.

⁷University at Buffalo, The State University of New York, Buffalo, NY USA. ⁸Case Western Reserve University, Cleveland, OH, USA. Correspondence and requests for materials should be addressed to A.C.-R. (email: aacruz@unillanos.edu.co)

Data set	Dice	PPV	NPV	TPR	TNR	FPR	FNR
TCGA	0.7586 ± 0.2006	0.7162 ± 0.2204	0.9677 ± 0.0511	0.8691 ± 0.1582	0.9218 ± 0.0764	0.0782 ± 0.0764	0.1309 ± 0.1582
NC	N/A	N/A	1 ± 0	N/A	0.9964 ± 0.0110	0.0036 ± 0.0110	N/A

Table 1. Performance measures for the ConvNet classifier on the TCGA (pathological, N = 195) and NC (normal, N = 21) data cohorts. The measures included Dice, PPV, NPV, TPR, TNR, FPR and FNR. Note that for the normal cases considered, not all the performance measures are shown because the NC data cohort did not have cancer annotations.

Digital pathology refers to the process of digitization of tissue slides. The process of slide digitization could enable more efficient storage, visualization, and pathologic analysis of tissue slides and could potentially improve overall efficiency of routine diagnostic pathology workflow¹¹.

Quantitative histomorphometry refers to the application of computational image analysis and machine learning algorithms to identify and characterize disease patterns on digitized tissue slides¹². In the context of breast cancer pathology, a number of computational imaging approaches have been recently applied for problems such as (i) detection of mitoses^{13–17}, tubules^{18,19}, nuclei^{19,20}, and lymphocytes²¹, (ii) cancer grading^{19,22}, (iii) correlation of quantitative histologic image features and molecular features of breast cancer aggressiveness²³, and (iv) identification of histologic image features that are predictive of breast cancer outcome and survival²⁴.

These previous approaches have typically limited their analysis to only small portions of tissue or tissue microarrays (TMAs) as opposed to larger whole slide images. Basavanhally *et al.*²², looked at the problem of computerized Bloom-Richardson grading of estrogen receptor positive breast cancers within manually defined regions of interest on whole slide images. While some approaches have looked at the problem of classifying images as either containing cancer or not^{25,26}, no approach that we are aware of has tackled the problem of automated delineation of invasive carcinoma on whole slide images.

Neural network learning refers to a class of machine learning methods that is gaining popularity in histopathology image analysis^{13,17,27–33}. A neural network is composed of artificial neurons that are arranged in layers and interchange information through connections. In recent years, neural network models comprising thousands of neurons arranged in several layers have been shown to perform exceptionally well in computer vision and pattern analysis tasks^{34–37}. Multi-level neural network learning approaches have recently acquired the name “deep learning” because of their multi-layer architecture. These networks are able to learn multiple levels of image representation to model complex non-linear relationships in the data, discovering more abstract and useful features that make it easier to extract useful information for high-level decision tasks such as segmentation, classification and prediction^{38–40}. Because of the large number of parameters involved, deep learning methods require a large number of labeled training exemplars in order to be optimally trained. In problems where large numbers of training exemplars are available, deep learning methods have shown impressive prediction results, often outperforming state-of-the-art classification methods^{36–38}. The advent of digitized whole pathology slides and the concomitant increase in the number of publicly available large histopathology image databases, such as The Cancer Genome Atlas, has made digital pathology a good candidate for the application of deep learning based classification models^{13,17,27–33}.

In this study, we present a classification approach for detecting presence and extent of invasive breast cancer on whole slide digitized pathology images using a ConvNet classifier^{38,41,42}. To ensure robustness of the classifier to variations in slide preparation, staining, and choice of scanning platform, we trained and validated the classifier with a large number of training exemplars drawn from three different institutions. Additionally the classifier was also independently evaluated on a large number of pathologic and normal cases drawn from The Cancer Genome Atlas (TCGA) and University Hospitals Case Medical Center. The goal of this study was to quantitatively evaluate the accuracy and robustness of a deep learning based machine classifier to automatically identify the extent of invasive breast cancer on digitized whole slide images.

Results

Quantitative evaluation for automatic invasive breast cancer detection. Table 1 shows the detection performance of the ConvNet classifier trained with data from Hospital of the University of Pennsylvania (HUP) and University Hospitals Case Medical Center/Case Western Reserve University (UHCMC/CWRU) in terms of mean and standard deviation of Dice coefficient, positive predictive value (PPV), negative predictive value (NPV), true positive rate (TPR), true negative rate (TNR), false positive rate (FPR) and false negative rate (FNR) for the validation data set, in turn comprised of the TCGA and the NC cohorts. Figure 1 shows some representative slide images from the validation data set. Figure 1A–C depict the ground truth annotations from the pathologists on three whole-slide images from the TCGA data cohort and Fig. 1D–F represent the automatic predictions of the fully-trained ConvNet classifier as a probability map of invasive breast cancer, with the color bar reflecting the probability values, high probability values reflected in red colors and low probability values in blue colors. Finally, three example slides without any malignant pathology and part of the NC cases are illustrated in Fig. 1G–I. As may be seen in Fig. 1G–I, the ConvNet classifier did not identify any regions as having invasive breast cancer.

Robustness and reproducibility analysis inside heterogeneous histopathology slides. A detailed analysis by subgroups of only a type of invasive breast cancer (i.e. IDC or ILC) and mixture of invasive and other types of *in situ* lesions (e.g. DCIS and LCIS) is presented in Table 2 for each of *ConvNet_{HUP}* and

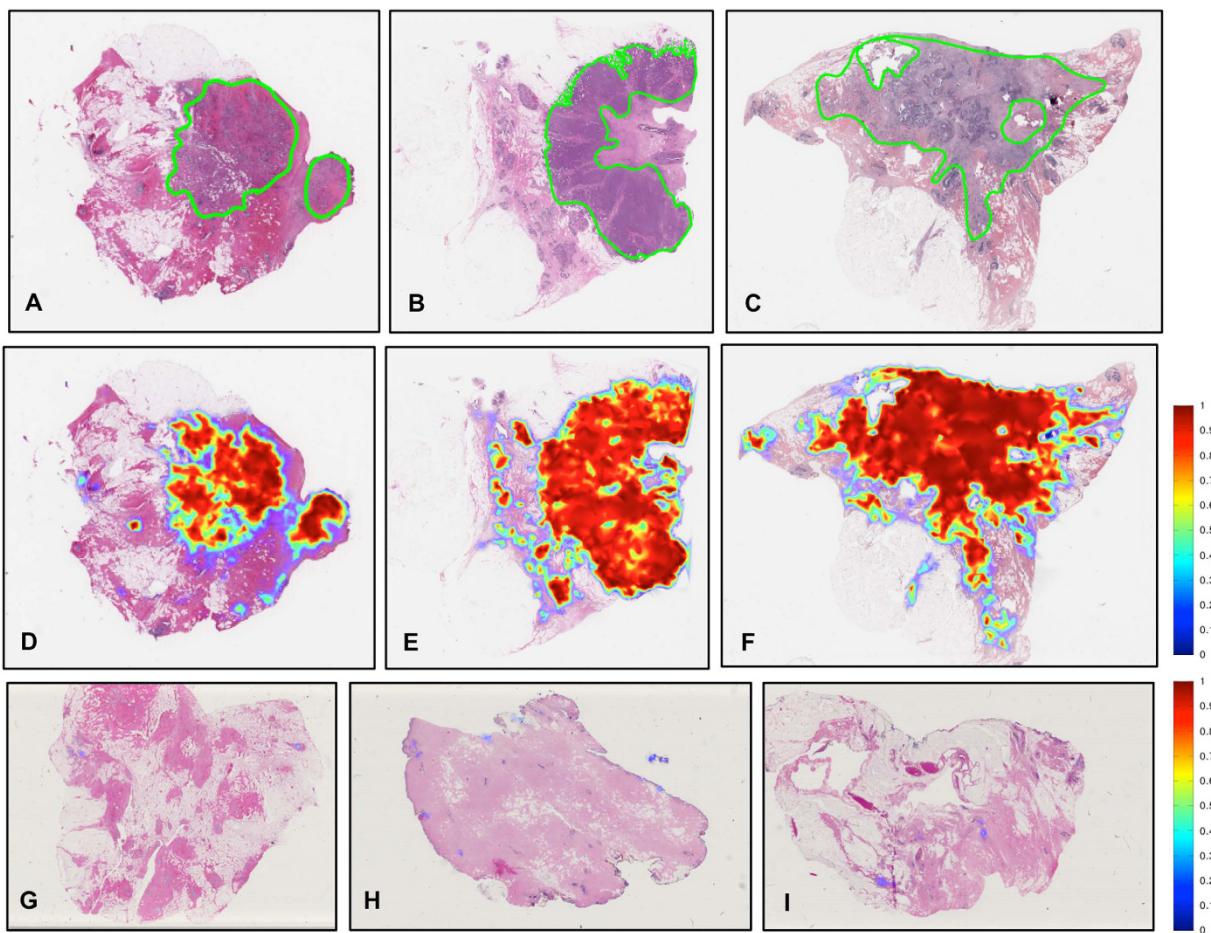


Figure 1. (A–C) Example whole-slide images from test TCGA data cohort with ground truth annotations from pathologists, (D–F) the corresponding region predictions produced by the ConvNet classifier and (G–I) region predictions for whole-slide images from the test NC data cohort of normal breast tissue without cancer.

Group	N	Dice	PPV	NPV
<i>ConvNet_{HUP}</i>				
All cases	40	0.6771 ± 0.2445	0.6464 ± 0.2870	0.9709 ± 0.0350
Only invasive	19	0.7578 ± 0.2166	0.7462 ± 0.2480	0.9654 ± 0.0355
Mixture	21	0.6041 ± 0.2501	0.5560 ± 0.2953	0.5560 ± 0.2953
<i>ConvNet_{UHCMC/CWRU}</i>				
All cases	40	0.6596 ± 0.2527	0.6370 ± 0.2941	0.9663 ± 0.0421
Only invasive	19	0.7596 ± 0.2074	0.7499 ± 0.2423	0.9614 ± 0.0440
Mixture	21	0.5691 ± 0.2602	0.5348 ± 0.3045	0.9708 ± 0.0409

Table 2. Performance of the *ConvNet_{HUP}* and *ConvNet_{UHCMC/CWRU}* classifiers on the CINJ data cohort in terms of means and standard deviation of Dice coefficient, PPV and NPV. The results in Table 2 are organized in terms of all cases in the CINJ cohort ($N = 40$), a subset of the CINJ cohort with invasive breast cancer alone ($N = 19$), and a mixture of invasive and other *in situ* subtypes of breast cancer ($N = 21$).

ConvNet_{UHCMC/CWRU} classifiers. Each of *ConvNet_{HUP}* and *ConvNet_{UHCMC/CWRU}* was trained with one of either the HUP or the UHCMC/CWRU cohorts. The quantitative performance results for both classifiers, *ConvNet_{HUP}* and *ConvNet_{UHCMC/CWRU}*, on the validation CINJ data cohort (*ConvNet_{HUP}*: Dice = 0.6771, PPV = 0.6464, NPV = 0.9709; *ConvNet_{UHCMC/CWRU}*: Dice = 0.6596, PPV = 0.6370, NPV = 0.9663) are similar. The results in Table 2 are also arranged according to the type of tumors in the sample (mixture or only invasive) and reveal that our method has better performance when the whole-slide images have only one type of invasive tumor (*ConvNet_{HUP}*: Dice = 0.7578, PPV = 0.7462, NPV = 0.9654; *ConvNet_{UHCMC/CWRU}*: Dice = 0.7596, PPV = 0.7462, NPV = 0.9614).

Figure 2 illustrates representative examples of whole slide images from the validation CINJ data cohort, involving only a single type of invasive tumor. The detection results obtained via *ConvNet_{HUP}* classifier were

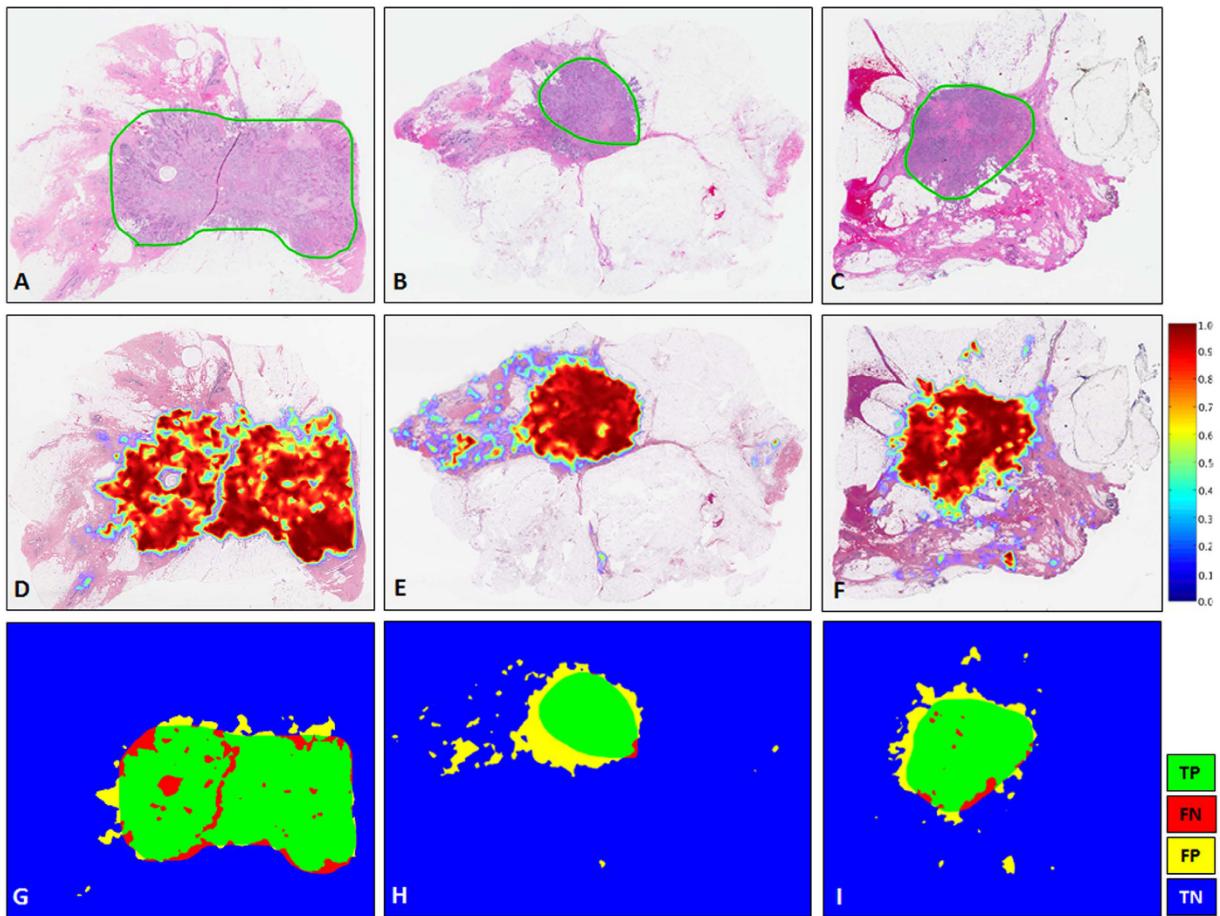


Figure 2. Example results for the $\text{ConvNet}_{\text{HUP}}$ classifier on the CINJ validation data cohort. The probability map predicted by the $\text{ConvNet}_{\text{HUP}}$ classifier (second row, (D–F)) was compared against ground truth annotations by a pathologist (first row (A–C)). The third row shows the evaluation results of the $\text{ConvNet}_{\text{HUP}}$ classifier in terms of TP (green), FN (red), FP (yellow), and TN (blue) regions.

compared against the ground truth annotations. Some cases from the CINJ validation data cohort where the $\text{ConvNet}_{\text{HUP}}$ classifier resulted in a poor detection performance are illustrated in Figs 3 and 4. The true-positives (TP), true-negatives (TN), false-positives (FP) and false-negatives (FN) regions, based on the predictions of the $\text{ConvNet}_{\text{HUP}}$ classifier, are illustrated in green, blue, yellow and red respectively. Figure 3 shows a case of mucinous (colloid) carcinoma, which is a rare type of invasive ductal carcinoma with a very low prevalence (2–3% of the total invasive breast cancer cases)⁴³. Figure 4 depicts a challenging case, which is composed of a mixture of invasive and *in situ* carcinoma elements.

Correspondence and reproducibility analysis among different classifiers and data cohorts. Table 3 illustrates the performance measures for the $\text{ConvNet}_{\text{HUP}}$ and $\text{ConvNet}_{\text{UHCMC/CWRU}}$ classifiers on the TCGA and NC testing sets. The consistency of the predictions of both models is estimated by calculating the correlation coefficient, r , between the performance measures obtained for each of $\text{ConvNet}_{\text{HUP}}$ and $\text{ConvNet}_{\text{UHCMC/CWRU}}$. On the TCGA cohort, the correlation coefficient in Dice coefficient for $\text{ConvNet}_{\text{HUP}}$ and $\text{ConvNet}_{\text{UHCMC/CWRU}}$ was $r = 0.8733$, reflecting a high degree of concordance. Figure 5 shows a scatter plot where the X axis corresponds to the Dice coefficient of the predictions generated by $\text{ConvNet}_{\text{HUP}}$ and the Y axis corresponds to the Dice coefficient of the predictions generated by $\text{ConvNet}_{\text{UHCMC/CWRU}}$ each dot corresponds to a slide sample from the TCGA data cohort. The scatter plot in Fig. 5 reveals a well-defined cluster with most cases aggregating in the upper-right corner. The scatter plot suggests that both the $\text{ConvNet}_{\text{HUP}}$ and $\text{ConvNet}_{\text{UHCMC/CWRU}}$ classifiers have a high degree of agreement in their predictions of the presence and extent of invasive tumor regions. Figure 5 also helps identify cases (red circles) where both the $\text{ConvNet}_{\text{HUP}}$ and $\text{ConvNet}_{\text{UHCMC/CWRU}}$ disagreed in their predictions. Figure 6 showcases the test images where the classifiers tended to disagree. A closer inspection of these cases suggested, suggests that the lack of concordance is primarily in those cases where the staining characteristics substantially deviate from the staining in the cases in the training cohorts. Figure 6A,B illustrate a couple of slides characterized by low levels of hematoxylin and high levels of eosin. The slide shown in Fig. 6C illustrates an example of a “black discoloration artifact” due to air bubbles on the slide, a common problem when the slide has been in storage for a long time. Usually, these cases are not appropriate for diagnosis and

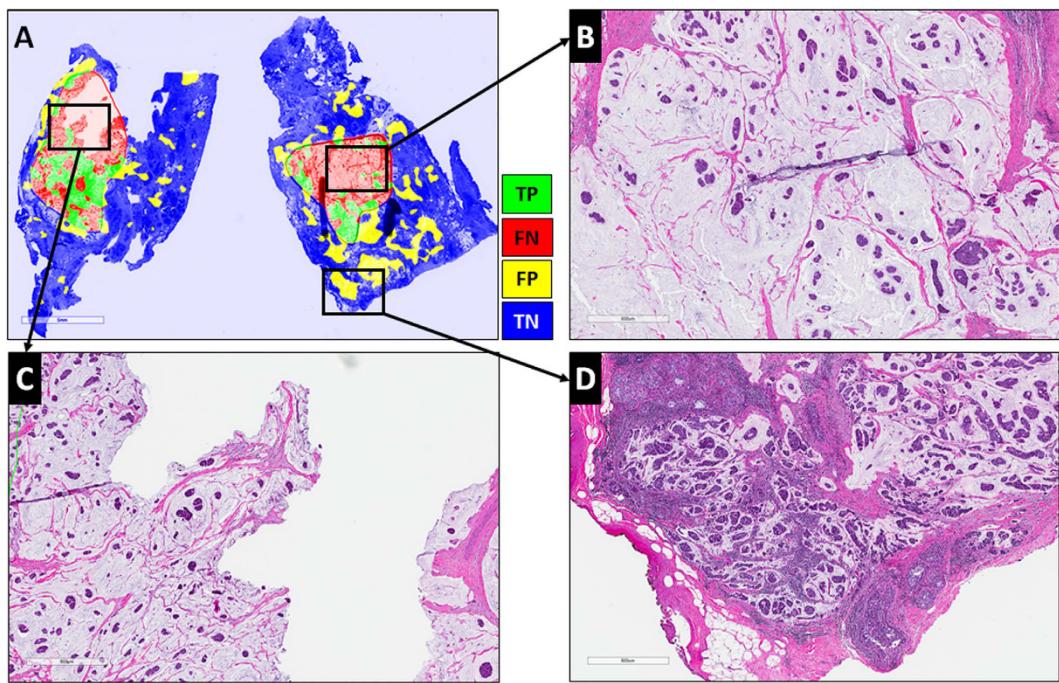


Figure 3. Whole-slide image from CINJ validation data cohort diagnosed with a rare type of IDC: mucinous carcinoma of the breast. (A) The comparison between the ground truth annotations and the predictions from the $\text{ConvNet}_{\text{HUP}}$ classifier reveal both FN (red) and FP (yellow) errors. (B,C) Most of the FN regions, i.e. tissues wrongly labeled as non-invasive tumor, correspond to mucinous carcinoma, whilst (D) most of FP regions, i.e. tissues wrongly predicted as invasive tumor, are actually invasive mucinous carcinoma that was not included in the annotations by the pathologist.

a pathologist would probably reject them in a quality control process ordering for another slide to be cut from the tissue sample.

Despite these special cases of disagreement caused by staining issues, both the $\text{ConvNet}_{\text{HUP}}$ and $\text{ConvNet}_{\text{UHCMC/CWRU}}$ classifiers yielded similar predictions and performance. However, the $\text{ConvNet}_{\text{HUP}}$ classifier appears to have a slightly higher confidence interval associated with the Dice and PPV performance measures. On the other hand, NPV and TNR from both classifiers show high mean values with very small standard deviation. Similarly on the NC data cohort, which is exclusively composed of normal breast samples, both the $\text{ConvNet}_{\text{HUP}}$ and $\text{ConvNet}_{\text{UHCMC/CWRU}}$ classifiers exhibited a very high mean TNR and a very low FPR, with very low associated standard deviation. This appears to suggest that both classifiers are able to confidently and consistently reject non-invasive tissue regions.

Example results of the predictions from the $\text{ConvNet}_{\text{HUP}}$ and $\text{ConvNet}_{\text{UHCMC/CWRU}}$ classifiers on the TCGA and NC test data sets are presented in Figs 7 and 8. While both the $\text{ConvNet}_{\text{HUP}}$ and $\text{ConvNet}_{\text{UHCMC/CWRU}}$ classifiers tend to produce consistent predictions, the ConvNet classifier, which was trained using the complete training data set, had the best overall performance (Fig. 1).

Discussion

The experimental results show that the method is able to detect invasive breast cancer regions on whole slide histopathology images with a high degree of precision, even when tested on cases from a cohort different to the one used for training. The most challenging cases for the method were slides where invasive breast cancer was mixed in with *in situ* disease (which is not surprising and could be reduced by training a more complex network that included examples of these precursor lesions).

An important part of the experimental setup was the analysis of the detection sensitivity of the method to the data used for training. The results show that the classifiers trained with two different data cohorts, HUP and UHCMC/CWRU, exhibit highly correlated performance measures ($r \geq 0.8$) over the independent TCGA test data cohort (see Table 3). Despite this, there are some differences in the prediction performance of the two classifiers, possibly suggesting “batch effects”⁴⁴, that originated from the process of ground truth annotation or slide digitization. This is illustrated in Figs 5 and 6, which show representative slides with artifacts due to problems in the histotechnique process. The method shows a very low false positive rate, as evidenced by the results in the NC cohort ($\text{ConvNet}_{\text{HUP}}$: FPR = 0.0284; $\text{ConvNet}_{\text{UHCMC/CWRU}}$: FPR = 0.0454), which comprised only normal breast sections. The performance of the ConvNet improved as the number of training samples increased, i.e. the ConvNet classifier trained with both the HUP and UHCMC/CWRU data cohorts yielded the best overall performance (Table 1 and Fig. 1).

The ConvNet was used as a patch-based classifier. We addressed the tissue classification task through a learned feature approach instead of a hand-crafted feature approach^{13,17,27,29,38,42}. However, any statistical or

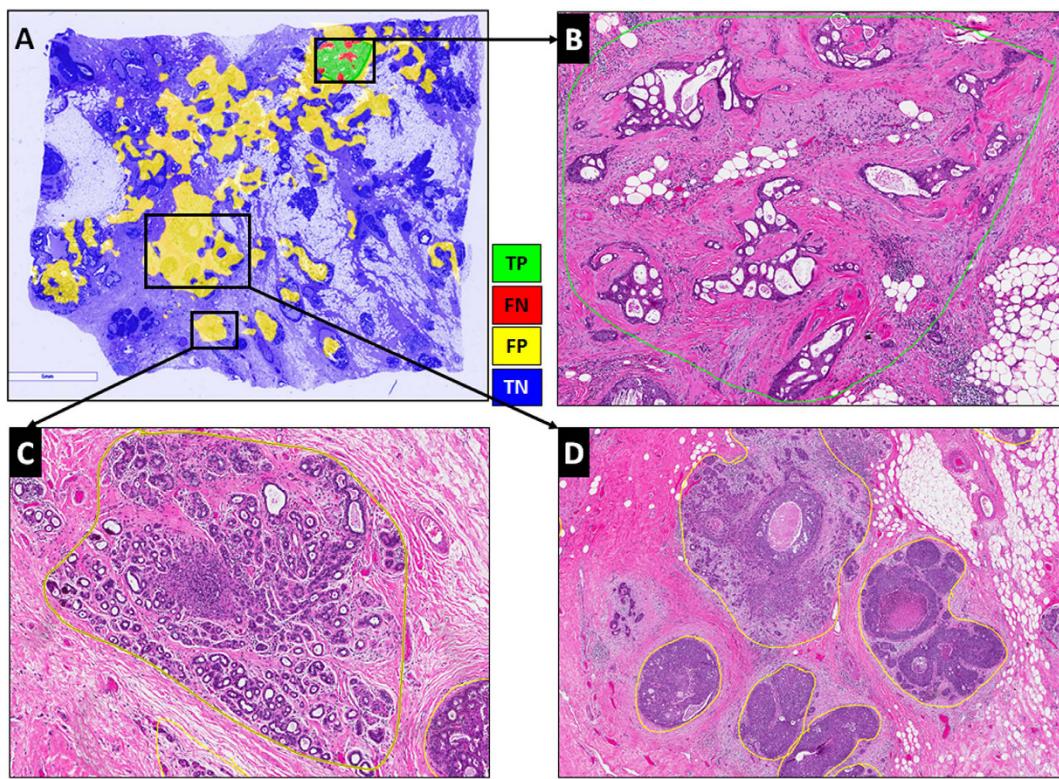


Figure 4. The most challenging whole-slide image in the CINJ validation cohort achieved the poorest performance via the $\text{ConvNet}_{\text{HUP}}$ classifier with (A) many FP regions and a Dice coefficient of 0.0745. (B) Some of the FN errors are due to the confounding morphologic attributes of the tumor, arising due to a mixing of IDC with fat cells and irregular, infiltrating looking cribriform glands with DCIS. The FP regions appear to be primarily be due to (C) sclerosing adenosis, and (D) DCIS surrounded by IDC.

machine learning classifier could be used in combination with a set of hand-crafted features for tissue classification. For instance, in addition to successful deep learning methods (i.e. ConvNets and Autoencoders) applied in histopathology image analysis^{13,17,27–29}, a set of hand-crafted features (color/intensity features, texture features, graph-based features, etc.) and machine learning methods (random forests and support vector machines) could and have been applied to histopathology image analysis^{33,45–48}. We did a comparative analysis with some of these visual features used in histopathology image analysis against three different ConvNet architectures. The ConvNet classifiers showed better performance in our patch-based image classification task. These results are presented in subsection: Invasive Breast Cancer Tissue Detection in Whole-Slide Images.

Our study did, however, have its limitations. There are some subtypes of invasive breast cancers that our method is not able to detect in a precise way such as the rare special histologic subtype mucinous carcinoma that comprises around 3% of the invasive breast cancers. In fact, in the test data set there are two cases similar to Fig. 3, with mucinous carcinoma that were not detected. Another limitation is that some *in situ* breast cancer regions were incorrectly classified as invasive breast cancer, *in situ* disease is different from invasive cancer. However, the reporting of the presence of both invasive and *in situ* carcinoma is a critical part of a diagnostic pathology workup. It is worth noting though that our approach was able to achieve a very high level of accuracy in terms of rejecting non-invasive tissue regions (normal controls) as not being cancer. Exemplars of DCIS and LCIS could, in future work, be included as part of an expanded learning set, as it would not doubt improve the classification performance and generalizability of the model. Additionally and as part of future work, the learning set could be expanded to include other rare variants of invasive ductal carcinoma, such as mucinous invasive carcinomas.

Batch effects are one of the main sources of variation in evaluating the performance of automated machine learning approaches. These batch effects include stain variability due to different histology protocols from different pathology labs and variations in the digitization process on account of the use of different slide scanners⁴⁴. Our results suggest a slight batch effect with two different data cohorts ($\text{ConvNet}_{\text{HUP}}$ and $\text{ConvNet}_{\text{UHCMC/CWRU}}$). Results of Table 2 appears to suggest that the differences between both classifiers is related more to the number of samples employed for training each of the classifiers (HUP, $N = 239$, and UHCMC/CWRU, $N = 110$) and possibly less related to the constitution of the different histologic subtypes within the training cohorts. However, the use of all available training data (HUP and UHCMC/CWRU) results in a more confident, accurate and robust ConvNet classifier. Clearly, increasing the training data set size and diversity results in a better and more robust algorithm. ConvNet also performs better when a case has only a single morphologic pattern of invasive breast cancer in the whole slide images. Cases with a mixture of invasive and *in situ* breast cancer resulted in a reduction in the overall accuracy of the ConvNet classifier (*in situ* tumors may be incorrectly classified as invasive carcinoma). One way

	Dice	PPV	NPV	TPR	TNR	FPR	FNR
TCGA							
<i>ConvNet_{HUP}</i>	0.7494 ± 0.2071	0.7071 ± 0.2254	0.9658 ± 0.0514	0.8600 ± 0.1705	0.9188 ± 0.0805	0.0812 ± 0.0805	0.1400 ± 0.1705
<i>ConvNet_{UHCMC/CWRU}</i>	0.7068 ± 0.2061	0.6464 ± 0.2188	0.9629 ± 0.0584	0.8676 ± 0.1706	0.8880 ± 0.0824	0.1120 ± 0.0824	0.1324 ± 0.1706
<i>r</i>	0.8733	0.9258	0.8109	0.6345	0.8055	0.8055	0.6345
NC							
<i>ConvNet_{HUP}</i>	N/A	N/A	1 ± 0	N/A	0.9716 ± 0.0693	0.0284 ± 0.0693	N/A
<i>ConvNet_{UHCMC/CWRU}</i>	N/A	N/A	1 ± 0	N/A	0.9546 ± 0.0816	0.0454 ± 0.0816	N/A
<i>r</i>	N/A	N/A	N/A	N/A	0.6876	0.6876	N/A

Table 3. Comparison and correlation of the *ConvNet_{UHCMC/CWRU}* and *ConvNet_{HUP}* classifiers in terms of Dice, PPV, NPV, TPR, TNR, FPR and FNR. Note that for the normal cases considered, not all the performance measures are shown because the NC data cohort did not have cancer annotations.

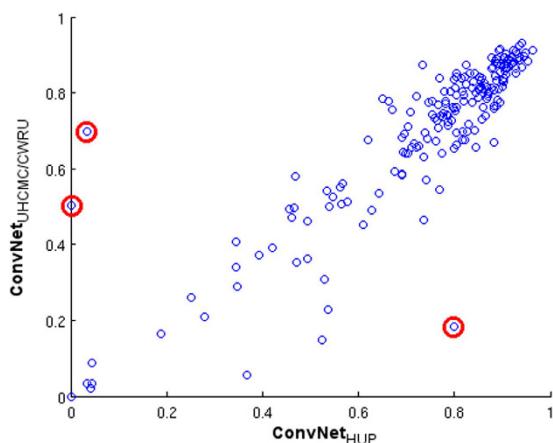


Figure 5. Agreement plot of the Dice coefficient for the *ConvNet_{HUP}* (X-axis) and *ConvNet_{UHCMC/CWRU}* (Y-axis) classifiers for each slide (blue circles) in the TCGA cohort. The slides with higher disagreement are identified with red circles (see Fig. 6).

of potentially reducing batch effects is to apply color normalization on the digitized images prior to training or application of the ConvNet classifier. To reduce false positive classification errors we are exploring the expansion of the current two class ConvNet classifier into a multiclass predictor. This will allow for the ConvNet classifier to explicitly deal with the detection of additional subtypes of invasive and *in situ* breast cancers.

One interesting aspect of our work is that the trained ConvNet classifier can be easily integrated into other computational frameworks such as automated tumor grading of ER+ breast cancer subtypes in histopathology images²². Our automated invasive cancer detection algorithm could thus pave the way for creation of decision support tools for breast cancer diagnosis, prognosis and theragnosis for use by the pathology community. Future studies will address these opportunities. Additionally follow on work will need to systematically compare the approach presented in this paper with state of the art visual features and machine learning approaches that have been previously applied to the problem of histopathology image analysis.

In conclusion, we presented an automatic invasive breast cancer detection method for whole slide histopathology images. Our study is unique in that it involved several hundred studies from multiple different sites for training the model. Independent testing of the model on multi-site data revealed that the model was both accurate and robust. This method can be applied to large, digitized whole slide images to detect invasive tissue regions, which could be integrated with other computerized solutions in digital pathology such as tumor grading.

Methods

Ethics Statement. Data analysis was waived review and consent by the IRB board, as all data was being analyzed retrospectively, after de- identification. All experimental protocols were approved under the IRB protocol No. 02-13-42C with the University Hospitals of Cleveland Institutional Review Board, and all experiments were carried out in accordance with approved guidelines.

Patients and Data Collection. This study involved images from five different cohorts from different institutions/pathology labs in the United States of America and TCGA^{49,50}. The five cohorts were used for training, validation and independent testing of our method. The training data set had 349 estrogen receptor-positive (ER+) invasive breast cancer patients, of which 239 were from Hospital of the University of Pennsylvania (HUP), and 110 from University Hospitals Case Medical Center/Case Western Reserve University (UHCMC/CWRU). Patients from the HUP cohort ranged in age between 20 and 79 (average age 55 ± 10). In the UHCMC/CWRU

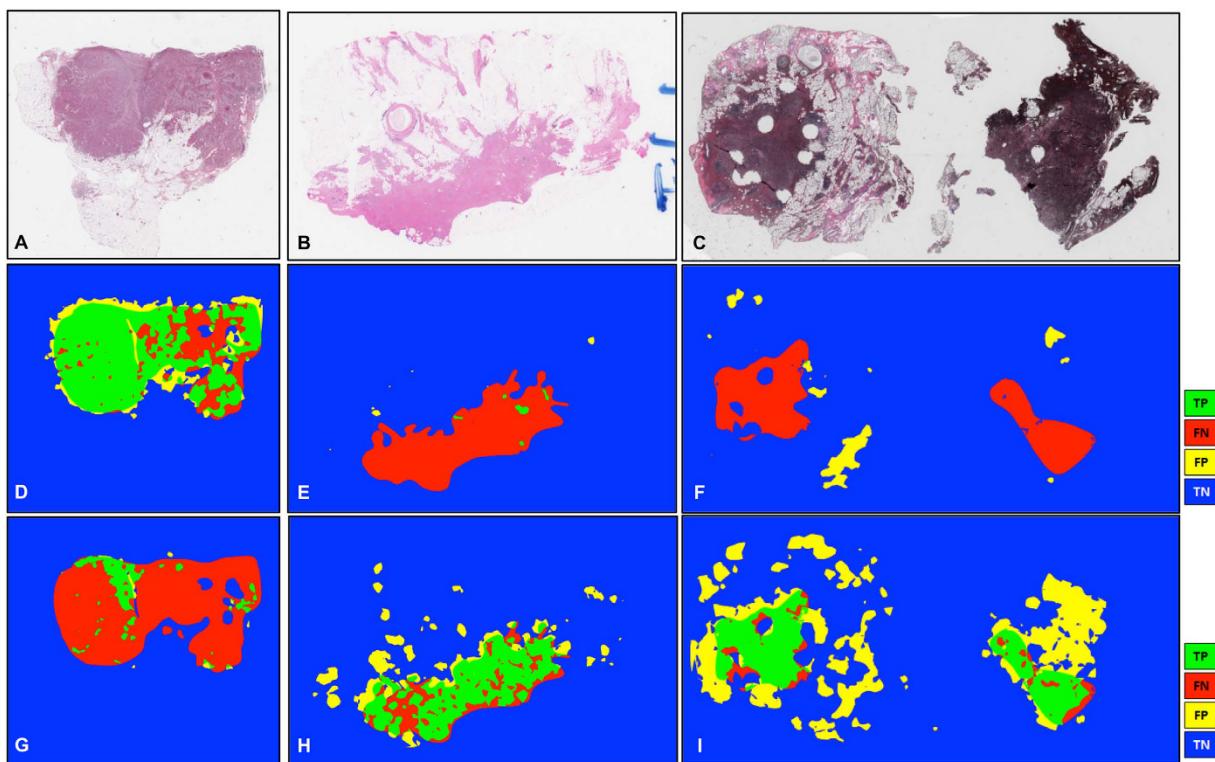


Figure 6. (A–C) Slides from the TCGA cohort which revealed disagreement between the predictions of the $\text{ConvNet}_{\text{HUP}}$ and $\text{ConvNet}_{\text{UHCMC/CWRU}}$ classifiers. The predictions of the (D–F) $\text{ConvNet}_{\text{HUP}}$ and (G–I) $\text{ConvNet}_{\text{UHCMC/CWRU}}$ classifiers were compared against the ground truth annotations in terms of TP (green), FN (red), FP (yellow) and TN (blue) regions.

cohort, the patient ages ranged from 25 to 81 (average age 58 ± 10). The validation data set contained 40 ER+ invasive breast cancer patients from the Cancer Institute of New Jersey (CINJ). The test data set was composed of two distinct subsets of positive and negative controls. For the test data set, we accrued a set of 195 ER+ invasive breast cancer cases from TCGA, ages ranging from 26 to 90 (average age 57 ± 13). For the negative controls (NC) in the test data set, we used normal breast tissue sections taken from uninvolved adjacent tissue from 21 patients diagnosed with invasive ductal carcinoma from UHCMC/CWRU, Cleveland, OH. Patient specific information pertaining to race, tumor grade, and outcome were not explicitly recorded for this study.

Hematoxylin and eosin (H&E) slides from all the various training, validation and testing cohorts (HUP, CINJ, UHCMC/CWRU, TCGA) were independently reviewed by four expert pathologists (NS, JT, MF, HG) to confirm the presence of at least one type of invasive breast cancer tumor. The normal control H&E slides were reviewed by one pathologist (HG). Tumors were categorized into one of the following histological types: invasive carcinoma were categorized as either invasive ductal carcinoma (IDC) or invasive lobular carcinoma (ILC), while pre-invasive carcinoma was categorized as ductal carcinoma *in situ* (DCIS) or lobular carcinoma *in situ* (LCIS). Only those cases were considered in our study where at least two pathologists concurred on the diagnosis.

Slide Digitization and Pathologists Ground Truth. H&E stained histopathology slides were digitized via a whole-slide scanner at 40x magnification for this study. An Aperio Scanscope CS scanner was used to digitize cases from the HUP, CINJ and TCGA cohorts. The Ventana iCoreo scanner was used for scanning the UHCMC/CWRU and NC data cohorts. 40x magnification corresponds to Aperio's slides at $0.25 \mu\text{m}/\text{pixel}$ resolution and to Ventana's slides at $0.23 \mu\text{m}/\text{pixel}$.

Expert pathologists provided the ground truth annotations of invasive breast cancer regions for all the data cohorts (HUP, CINJ, UHCMC/CWRU, TCGA). The region annotations were obtained via manual delineation of invasive breast cancer regions by expert pathologists using the ImageScope v11.2 program from Aperio and the Ventana Image Viewer v3.1.4 from Ventana. To alleviate the time and effort required to create the ground truth annotations for extent of invasive breast cancer, the pathologists were asked to perform their annotations at 2x magnification or less. All whole-slide images previously sampled at 40x were thus subsequently downsampled (by a factor of 16:1) to a resolution of $4 \mu\text{m}/\text{pixel}$.

In order to analyze the agreement between expert pathologists, the Dice coefficient and Cohen's Kappa coefficient were calculated between NS + MF and HG manual delineations. The Cohen's Kappa coefficient was determined to be $\kappa = 0.748^{51}$, in turn reflecting good agreement between the experts⁵². In addition, the Dice coefficient was calculated to measure the overlap between the cancer annotations between NS + MF and HG delineations and was determined to be $DSC = 0.6685^{53}$. Figure 9 below depicts the Dice coefficient dispersion between expert pathologists. Figure 9 shows that the DSC measure is not a Gaussian distribution and has a median value equal

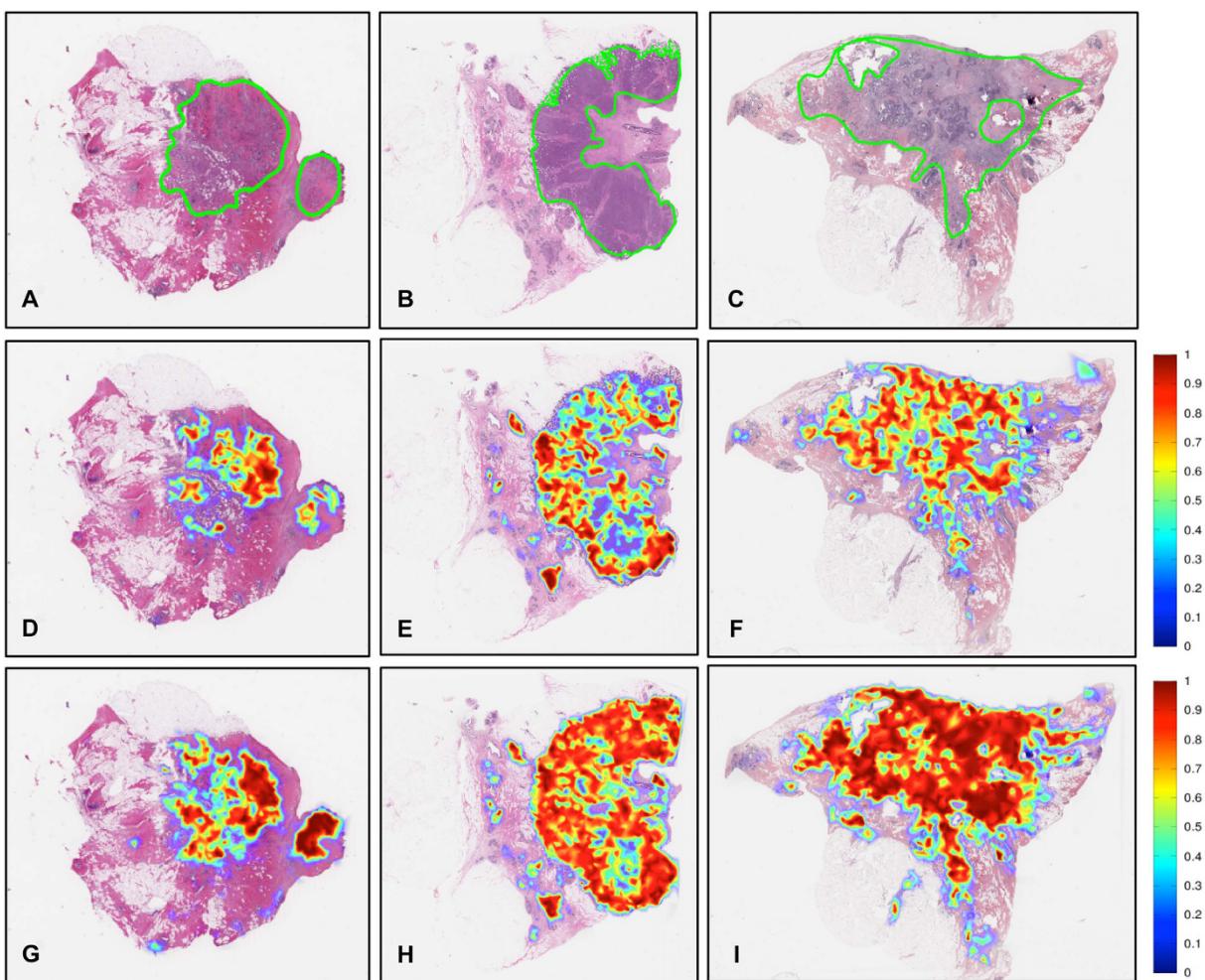


Figure 7. (A–C) Example whole-slide images from the TCGA data cohort with corresponding ground truth annotations. The probability maps generated by the $\text{ConvNet}_{\text{UHCMC/CWRU}}$ and $\text{ConvNet}_{\text{HUP}}$ classifiers are shown in panels (D–F,G–I) respectively.

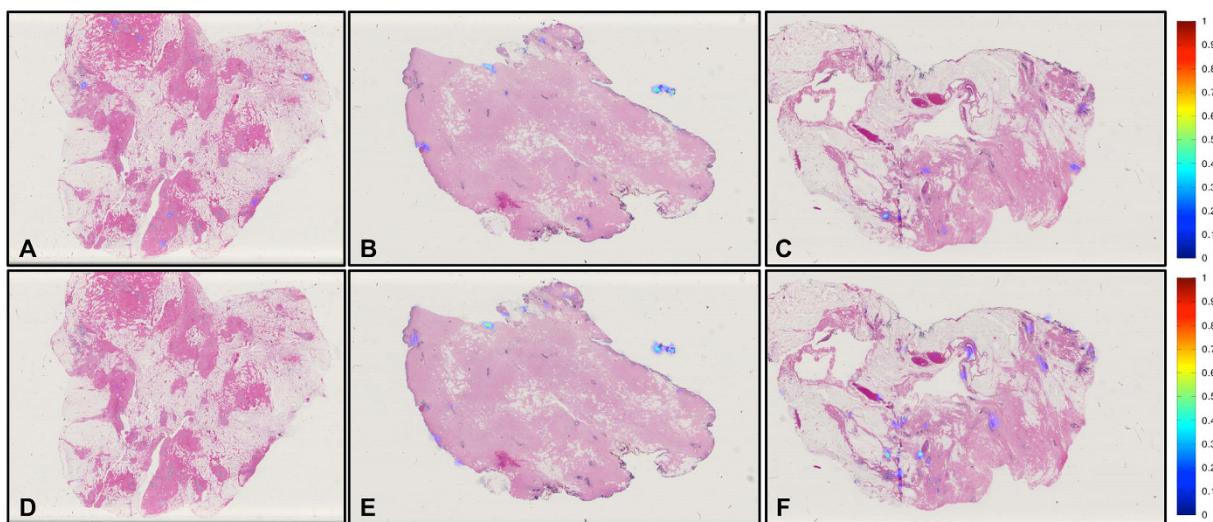


Figure 8. The probability maps obtained via the $\text{ConvNet}_{\text{UHCMC/CWRU}}$ and $\text{ConvNet}_{\text{HUP}}$ classifiers on whole-slide images of normal breast sections from the UHCMC/CWRU and NC data cohorts are shown in panels (A–C,D–F) respectively.

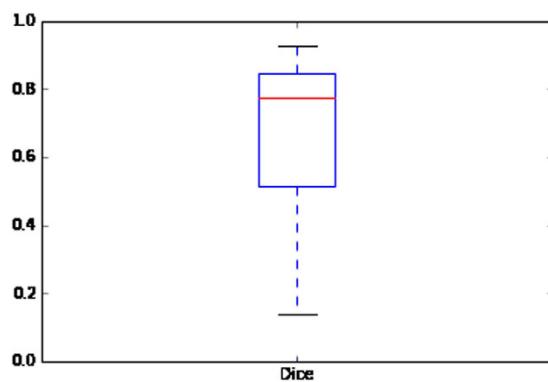


Figure 9. Dice coefficient between pathologist annotations for the CINJ data cohort ($N = 40$).

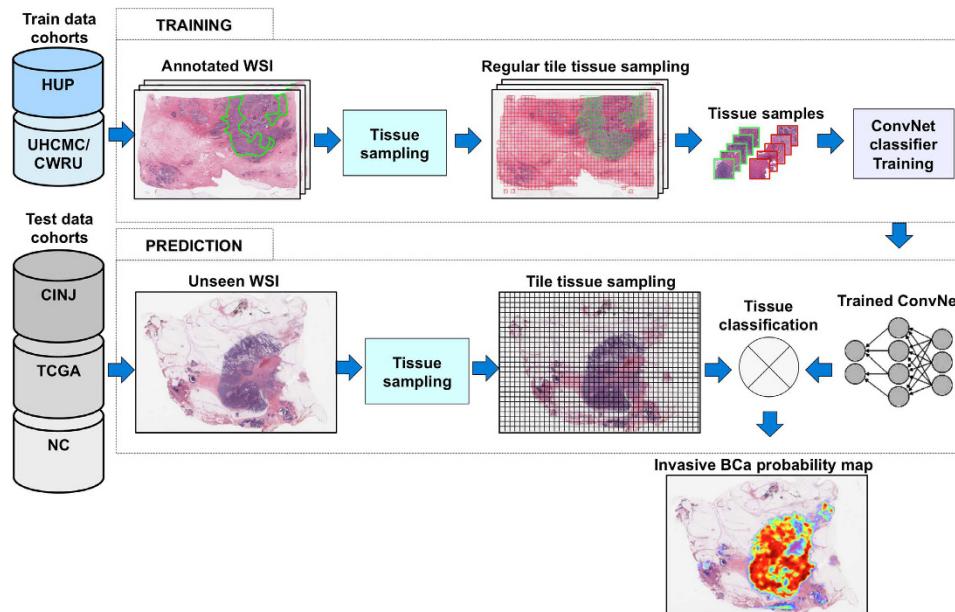


Figure 10. Overview of the process of training and testing of the deep learning classifiers for invasive breast cancer detection on whole-slide images. The training data set had 349 ER+ invasive breast cancer patients (HUP $N = 239$, UHCMC/CWRU $N = 110$). The validation data set contained 40 ER+ invasive breast cancer patients from the Cancer Institute of New Jersey (CINJ). The test data set was composed of 195 ER+ invasive breast cancer cases from TCGA and 21 negative controls (NC).

to 0.7764. The DSC agreement was found to be greater than 0.7 for a majority of the images studied, where good agreement is typically defined as when agreement is greater than 60%.

Invasive Breast Cancer Tissue Detection in Whole-Slide Images. Our deep-learning based approach for detection of invasive breast cancer on whole-slide images is illustrated in Fig. 10. The approach comprises three main steps: (i) tile tissue sampling, (ii) tile pre-processing, and (iii) convolutional neural network (ConvNet) based classification. In this work, a tile is a square tissue region with a size of $200 \times 200 \mu\text{m}$. The tile tissue sampling process involves extraction of square regions of the same size ($200 \times 200 \mu\text{m}$), on a rectangular grid for each whole-slide image. Only tissue regions are invoked during the sampling process and any regions corresponding to non-tissue within the background of the slide are ignored. The first part of the tile pre-processing procedure involves a color transformation from the original Red-Green-Blue color space representation to a YUV color space representation. A color normalization step is then applied to the digitized slide image to get zero mean and unit variance of the image intensities, and to remove correlations among the pixel intensity values. Tiles extracted from new whole-slide images, different from the ones used for training, are preprocessed using the same mean and standard deviation values in the YUV color space learned during training. The ConvNet classifier^{41,42}, was trained using a set of image tiles extracted from invasive (positive examples) and non-invasive (negative examples) tissue regions, annotated on whole slide digitized images by expert pathologists. Positive examples were identified as those in which the detected cancer regions had a minimum of 80% overlap with the

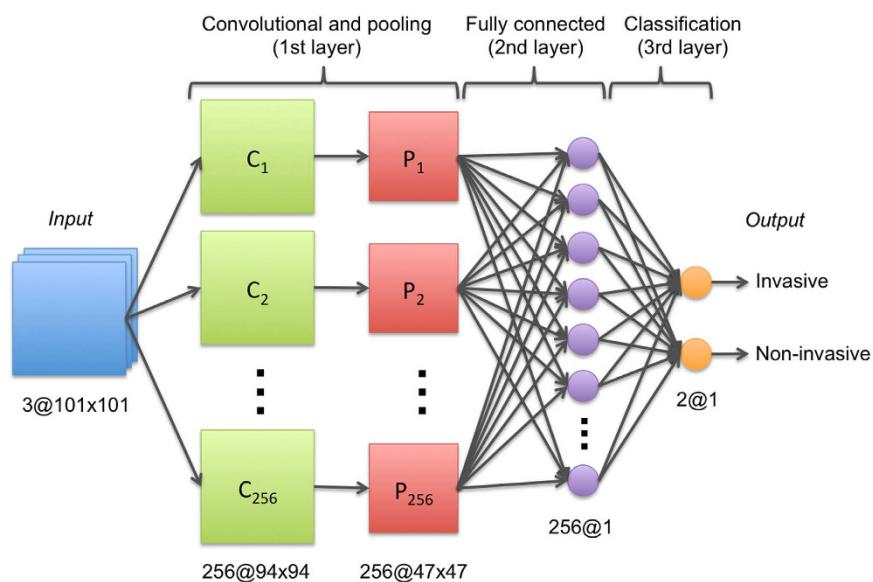


Figure 11. 3-layer ConvNet architecture.

manual annotations of the expert pathologists. Three different ConvNet architectures were evaluated using the training data: 1) a simple 3-layer ConvNet architecture, 2) a typical 4-layer ConvNet architecture, and 3) a deeper 6-layer ConvNet architecture. The 3-layer ConvNet architecture is constituted as follows, the first layer is the convolutional and pooling layer and the second is a fully connected layer, where each layer has 256 units (or neurons). The third is the classification layer with two units as outputs, one for each class (invasive and non-invasive), corresponding to a value between zero and one. The 4-layer ConvNet architecture is comprised of an initial convolutional and pooling layer with 16 units, followed by a second convolutional and pooling layer with 32 units, the third layer is a fully connected layer with 128 units, and the final classification layer comprises two units as class outputs (invasive and non-invasive). The 6-layer ConvNet architecture comprises four convolutional and pooling layers with 16 units, a fully connected layer with 128 units, and a final classification layer with two units as class outputs (invasive and non-invasive). The 3-layer ConvNet resulted in the best performance and hence was selected as the model of choice for all subsequent experiments (Fig. 11). The implementation of the ConvNets classifier was performed using Torch 7, a scientific computing framework for machine learning⁵⁴.

The ConvNet classifier was trained with images from HUP and UHCMC/CWRU. The training set comprised a large number of cases manually annotated by pathologists, i.e. 349 cases (239 from HUP and 110 from UHCMC/CWRU). The validation data cohort was the smaller data set with manual annotations from pathologists of invasive tumors (CINJ, $N = 40$), and the testing data sets were: a publicly available data set with invasive tumors (TCGA, $N = 195$) and normal control cases without breast cancer (NC, $N = 21$). Our training set comprised a total of 344,662 patches, of which 91,952 were from the positive class (invasive) and 252,710 were from the negative class (non-invasive). We applied data augmentation only to the positive class, the positive class being the minority class in terms of number of samples. The data augmentation process for the positive class comprised of duplicating the number of patches with artificial rotations and mirroring of patches. The weights were randomly initialized and updated during the training stage by using the stochastic gradient descent algorithm. This strategy was used to “learn” the weights (features) of the network from the training set. The number of epochs to train the ConvNets classifiers was 25. The mini-batch size was 32. The remaining parameters for the ConvNet classifier were tuned during the training process. These parameters included the learning rate, learning rate decay, non-linear function and pooling function. The optimal parameter configuration was determined to be $1e^{-3}$, $1e^{-7}$, ReLU and L2-norm, respectively. The best parameter configuration of the classifier was identified using the average area under the ROC curve (AUC) calculated over all slides in the CINJ data cohort, $N = 40$. The CINJ data cohort was used as the validation data set because it is the smaller pathological data set with manual independent annotations from 3 different pathologists of invasive tumors. The AUC is a non-biased classification measure that allows for the evaluation of classification performance independent of a fixed threshold. In this work classification performance was evaluated over all the image tiles extracted from all the whole-slide images in the CINJ data cohort, tiles that correspond to either invasive or non-invasive tissue classes. Table 4 presents a comparison between the ConvNet classifiers and state of the art handcrafted visual features (color, shape, texture and topography) used in histopathology image analysis. The classification results associated with these handcrafted features is lower compared to the ConvNet classifier and also results in more variability. The comparative evaluation helped identify the ConvNet classifier with the best classification performance and simplest configuration (Avg. AUC = 0.9018 ± 0.0093) for the subsequent experiments involving the independent test set.

Method evaluation. We evaluated the accuracy of the ConvNet classifier in whole slide images by comparing the predictions of invasive regions in the test data set against the corresponding ground-truth regions

Method	AUC
6-layer ConvNet	0.9021 +/- 0.0097
3-layer ConvNet	0.9018 +/- 0.0093
4-layer ConvNet	0.8915 +/- 0.0093
Color and intensity features ^{47,55,56}	0.8711 +/- 0.0947
Color histograms ^{47,56}	0.8448 +/- 0.1047
Shape Index histogram ⁵⁷	0.8444 +/- 0.1065
Haralick features ^{47,56,58}	0.8385 +/- 0.0942
Topography and Graph-based features ^{46,55,56}	0.7998 +/- 0.1068

Table 4. Comparison of ConvNet classifiers and visual features (color, shape, texture and topography) in terms of AUC.

annotated by expert pathologists. The test data sets included the slides in the TCGA and NC cohorts. A quantitative evaluation was performed by measuring the Dice coefficient, positive predictive value (PPV), negative predictive value (NPV), true positive rate (TPR), true negative rate (TNR), false positive rate (FPR) and false negative rate (FNR) across all the test slides. These measures were evaluated for each whole-slide image and the mean and standard deviation in performance measures were calculated for each test data cohort.

In addition to training the ConvNet classifier with the full training data set (HUP and UHCMC/CWRU), two additional classifiers were trained using, in each case, one of the training cohorts: *ConvNet_{HUP}* trained with the HUP cohort and *ConvNet_{UHCMC/CWRU}* trained with the UHCMC/CWRU cohort. The motivation was to analyze the sensitivity of the classifier to the training data sets. Both *ConvNet_{HUP}* and *ConvNet_{UHCMC/CWRU}* were evaluated on both the validation (CINJ cohort) and test data sets (TCGA and NC cohorts) to analyze how and where their predictions diverged. Specifically we measured the correlation coefficient r between the prediction performance measures for *ConvNet_{HUP}* and *ConvNet_{UHCMC/CWRU}* across all slides in each test cohort.

References

- Genestie, C. *et al.* Comparison of the prognostic value of Scarff-Bloom-Richardson and nottingham histological grades in a series of 825 cases of breast cancer: major importance of the mitotic count as a component of both grading systems. *Anticancer Research* **18**, 571–576 (1998).
- Elston, C. W. & Ellis, I. O. 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**, 403–410 (1991).
- Frierson, H. F. *et al.* Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma. *American journal of clinical pathology* **103**, 195–8 (1995).
- Gomes, D. S., Porto, S. S., Balabram, D. & Gobbi, H. Inter-observer variability between general pathologists and a specialist in breast pathology in the diagnosis of lobular neoplasia, columnar cell lesions, atypical ductal hyperplasia and ductal carcinoma *in situ* of the breast. *Diagnostic pathology* **9**, 121 (2014).
- Longacre, T. A. *et al.* Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an NCI breast cancer family registry study. *Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc* **19**, 195–207 (2006).
- Perez, A. A., Balabram, D., Salles, M. d. A. & Gobbi, H. Consultation in breast surgical pathology: interobserver diagnostic variability of atypical intraductal proliferative lesions. *Revista brasileira de ginecologia e obstetrícia: revista da Federação Brasileira das Sociedades de Ginecologia e Obstetrícia* **35**, 164–70 (2013).
- Dillon, D. A., Guidi, A. J. & Schnitt, S. J. Pathology of invasive breast cancer. In Harris, J. R., Lippman, M. E., Morrow, M. & Osborne, C. K. (eds) *Diseases of the Breast* chap. Chapter 28, 374–407 4th edition edn (Lippincott Williams & Wilkins, 2010).
- DeSantis, C., Siegel, R., Bandi, P. & Jemal, A. Breast cancer statistics, 2011. *CA: A Cancer Journal for Clinicians* **61**, 408–418 (2011).
- van Baardwijk, A. *et al.* PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and involved nodal volumes. *International Journal of Radiation Oncology Biology Physics* **68**, 771–778 (2007).
- Weaver, D. L. *et al.* Comparison of pathologist-detected and automated computer-assisted image analysis detected sentinel lymph node micrometastases in breast cancer. *Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc* **16**, 1159–63 (2003).
- Madabhushi, A. Digital pathology image analysis: opportunities and challenges. *Imaging In Medicine* **1**, 7–10 (2009).
- Lewis, J. S., Ali, S., Luo, J., Thorstad, W. L. & Madabhushi, A. A Quantitative Histomorphometric Classifier (QuHbIC) Oropharyngeal Squamous Cell Carcinoma. *American Journal of Surgical Pathology* **38**, 128–137 (2014).
- Ciresan, D., Giusti, A., Gambardella, L. & Schmidhuber, J. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013*, vol. 8150 of *Lecture Notes in Computer Science* 411–418 (Springer Berlin Heidelberg, 2013).
- Irshad, H. Automated mitosis detection in histopathology using morphological and multi-channel statistics features. *Journal of Pathology Informatics* **4**, 10 (2013).
- Ludovic, R. *et al.* Mitosis detection in breast cancer histological images An ICPR 2012 contest. *Journal of Pathology Informatics* **4**, 8 (2013).
- Veta, M. *et al.* Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical Image Analysis* **20**, 237–248 (2015).
- Wang, H. *et al.* Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging* **1**, 34003 (2014).
- Basavanhally, A. *et al.* Incorporating domain knowledge for tubule detection in breast histopathology using O'Callaghan neighborhoods. In *SPIE Medical Imaging*, vol. 7963 of *Computer-Aided Diagnosis* 796310. SPIE (SPIE, 2011).
- Dalle, J. R., Leow, W. K., Racocceanu, D., Tutac, A. E. & Putti, T. C. Automatic breast cancer grading of histopathological images. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 3052–5 (2008).
- Monaco, J. P. *et al.* Image segmentation with implicit color standardization using spatially constrained expectation maximization: detection of nuclei. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* **15**, 365–372 (2012).

21. Basavanhally, A. *et al.* Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology. *IEEE transactions on bio-medical engineering* **57**, 642–653 (2010).
22. Basavanhally, A. *et al.* Multi-Field-of-View Framework for Distinguishing Tumor Grade in ER+ Breast Cancer From Entire Histopathology Slides. *IEEE transactions on biomedical engineering* **60**, 2089–2099 (2013).
23. Basavanhally, A. *et al.* Multi-field-of-view strategy for image-based outcome prediction of multi-parametric estrogen receptor-positive breast cancer histopathology: Comparison to Oncotype DX. *Journal of pathology informatics* **2**, S1 (2011).
24. Beck, A. H. *et al.* Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science translational medicine* **3**, 108ra113 (2011).
25. Dundar, M. M. *et al.* Computerized classification of intraductal breast lesions using histopathological images. *IEEE Transactions on Biomedical Engineering* **58**, 1977–1984 (2011).
26. Naik, S. *et al.* Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 284–287 (IEEE, 2008).
27. Arevalo, J., Cruz-Roa, A., Arias, V., Romero, E. & Gonzalez, F. A. An unsupervised feature learning framework for basal cell carcinoma image analysis. *Artificial Intelligence in Medicine* **64**, 131–145 (2015).
28. Cruz-Roa, A., Arévalo, J., Basavanhally, A., Madabhushi, A. & Gonzalez, F. A comparative evaluation of supervised and unsupervised representation learning approaches for anaplastic medulloblastoma differentiation. In *Tenth International Symposium on Medical Information Processing and Analysis (SIPAIM 2014)* (2014).
29. Cruz-Roa, A., Arevalo, J., Madabhushi, A. & González, F. A Deep Learning Architecture for Image Representation, Visual Interpretability and Automated Basal-Cell Carcinoma Cancer Detection. In Mori, K., Sakuma, I., Sato, Y., Barillot, C. & Navab, N. (eds) *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013*, vol. 8150 of *Lecture Notes in Computer Science* 403–410 (Springer Berlin Heidelberg, 2013).
30. Cruz-Roa, A. *et al.* Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *Proc. SPIE* **9041**, 904103–904115 (2014).
31. Le, Q. V. *et al.* Learning invariant features of tumor signatures. In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on* 302–305 (IEEE, 2012).
32. Malon, C. & Cosatto, E. Classification of mitotic figures with convolutional neural networks and seeded blob features. *Journal of Pathology Informatics* **4**, 9 (2013).
33. Otálora, S. *et al.* *Combining Unsupervised Feature Learning and Riesz Wavelets for Histopathology Image Representation: Application to Identifying Anaplastic Medulloblastoma*, chap. Combining 581–588 (Springer International Publishing, Cham, 2015).
34. Donahue, J. *et al.* DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *International Conference in Machine Learning (ICML)* (2014).
35. Hinton, G. & Srivastava, N. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).
36. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems* **25**, 1106–1114 (2012).
37. Le, Q. *et al.* Building high-level features using large scale unsupervised learning. In *International Conference in Machine Learning* (2012).
38. Bengio, Y., Courville, A. & Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**, 1798–1828 (2013).
39. Bengio, Y. Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning* **2**, 1–127 (2009).
40. Deng, L. & Yu, D. Deep Learning: Methods and Applications. *Foundations and Trends in Signal Processing* **7**, 197–387 (2014).
41. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324 (1998).
42. LeCun, Y. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems* 253–256 (2010).
43. Lacroix-Triki, M. *et al.* Mucinous carcinoma of the breast is genetically distinct from invasive ductal carcinomas of no special type. *The Journal of Pathology* **222**, 282–298 (2010).
44. Kothari, S. *et al.* Removing Batch Effects From Histopathological Images for Enhanced Cancer Diagnosis. *Biomedical and Health Informatics, IEEE Journal of* **18**, 765–772 (2014).
45. Fuchs, T. J. & Buhmann, J. M. Computational pathology: challenges and promises for tissue analysis. *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society* **35**, 515–30 (2011).
46. Gurcan, M. N. *et al.* Histopathological Image Analysis: A Review. *IEEE reviews in biomedical engineering* **2**, 147–171 (2009).
47. He, L., Long, R., Antani, S. & Thoma, G. Histology image analysis for carcinoma detection and grading. *Computer Methods and Programs in Biomedicine* **107**, 538–556 (2012).
48. Veta, M., Pluim, J. P. W., van Diest, P. J. & Viergever, M. A. Breast Cancer Histopathology Image Analysis: A Review. *Biomedical Engineering, IEEE Transactions on* **61**, 1400–1411 (2014).
49. McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
50. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* **45**, 1113–1120 (2013).
51. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* **20**, 37–46 (1960).
52. Viera, A. & Garrett, J. Understanding interobserver agreement: The kappa statistic. *Family Medicine* **37**, 360–363 (2005).
53. Dice, L. R. Measures of the amount of ecological association between species. *Ecology* **26**, 297–302 (1945).
54. Collobert, R., Kavukcuoglu, K. & Farabet, C. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop* (2011).
55. Doyle, S., Agner, S., Madabhushi, A., Feldman, M. & Tomaszewski, J. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *5th IEEE ISBI: From Nano to Macro* 496–499 (IEEE, 2008).
56. Arévalo, J., Cruz-Roa, A. & González, F. A. Histopathology image representation for automatic analysis: A state-of-the-art review. *Revista Med* **22**, 79–91 (2014).
57. Dorai, C. & Jain, A. Cosmos-a representation scheme for 3d free-form objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**, 1115–1130 (1997).
58. Doyle, S., Madabhushi, A., Feldman, M. & Tomaszewski, J. A Boosting Cascade for Automated Detection of Prostate Cancer from Digitized Histology. In Larsen, R., Nielsen, M. & Sporrings, J. (eds) *International Conference on Medical Image Computing and Computer-Assisted Intervention LNCS*, 504–511 (Springer, Berlin, Heidelberg, 2006).

Acknowledgements

This work was partially funded by project numbers 1225-569-34920 through Colciencias contract number 0213-2013, and C03-F02-35-2015 by DGI-Unillanoss. Dr. Cruz-Roa was supported via a doctoral fellowship grant from the Administrative Department of Science, Technology and Innovation of Colombia (Colciencias) 528/2011. Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award numbers 1U24CA199374-01, R01CA202752-01A1 R01CA208236-01A1 R21CA179327-01; R21CA195152-01 the National Institute of Diabetes and Digestive and Kidney Diseases under award number

R01DK098503-02, National Center for Research Resources under award number 1 C06 RR12463-01 the DOD Prostate Cancer Synergistic Idea Development Award (PC120857); the DOD Lung Cancer Idea Development New Investigator Award (LC130463), the DOD Prostate Cancer Idea Development Award; the DOD Peer Reviewed Cancer Research Program W81XWH-16-1-0329 the Case Comprehensive Cancer Center Pilot Grant VelaSano Grant from the Cleveland Clinic the Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering at Case Western Reserve University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors also thank for K40 Tesla GPU donated by NVIDIA, High Performance Computing Resources in the Core Facility for Advanced Research Computing at Case Western Reserve University and GridUIS at Industrial University of Santander, which were used for training process of ConvNets classifiers. The results presented here are in part based upon data generated by the TCGA Research Network: <http://cancergenome.nih.gov/>.

Author Contributions

A.M. conceived of the study. A.M., A.C.R., F.G. designed the study. A.C.R., F.G designed the deep learning method. A.C.R. did the majority of the writing of the paper, implemented the deep learning method and also ran all the experiments. A.B. contributed to obtain the selected cases and annotations in the appropriate format for the study from pathologists. A.B. and S.G. contributed to delimitate the scope of the study. All co-authors contributed to the editing of the manuscript. H.G., N.S., J.E.T., M.D. provided annotations of the cancer extent (ground truth) for the region of cancer on all the digitized slide images employed in this study.

Additional Information

Competing Interests: Drs Madabhushi, Feldman, Ganesan, and Tomaszewski are scientific consultants for the digital pathology company Inspirata Inc. Drs Madabhushi, Feldman, Ganesan, and Tomaszewski also serve on the scientific advisory board for the digital pathology company Inspirata Inc. Dr. Madabhushi also has an equity stake in Inspirata Inc. and Elucid Bioimaging Inc.

How to cite this article: Cruz-Roa, A. *et al.* Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Sci. Rep.* 7, 46450; doi: 10.1038/srep46450 (2017).

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



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017