Automatic detection of invasive ductal carcinoma in whole slide images with Convolutional Neural Networks

Angel Cruz-Roa[a], Ajay Basavanhally[b], Fabio González[a], Hannah Gilmore[c], Michael Feldman[d], Shridar Ganesan[e], Natalie Shih[d], John Tomaszewski[f] and Anant Madabhushi[g]

[a]Universidad Nacional de Colombia, Bogotá, Colombia
[b]Rutgers University, Piscataway, NJ, USA
[c]University Hospitals, Cleveland, OH, USA
[d]Hospital of the University of Pennsylvania, Philadelphia, PA, USA
[e]Cancer Institute of New Jersey, New Brunswick, NJ, USA
[f]University at Buffalo, The State University of New York, Buffalo, NY USA
[g]Case Western Reserve University, Cleveland, OH, USA

ABSTRACT

This paper presents a deep learning approach for automatic detection and visual analysis of invasive ductal carcinoma (IDC) tissue regions in whole slide images (WSI) of breast cancer (BCa). Deep learning approaches are learn-from-data methods involving computational modeling of the learning process. This approach is similar to how human brain works using different interpretation levels or layers of most representative and useful features resulting into a hierarchical learned representation. These methods have been shown to outpace traditional approaches of most challenging problems in several areas such as speech recognition and object detection. Invasive breast cancer detection is a time consuming and challenging task primarily because it involves a pathologist scanning large swathes of benign regions to ultimately identify the areas of malignancy. Precise delineation of IDC in WSI is crucial to the subsequent estimation of grading tumor aggressiveness and predicting patient outcome. DL approaches are particularly adept at handling these types of problems, especially if a large number of samples are available for training, which would also ensure the generalizability of the learned features and classifier. The DL framework in this paper extends a number of convolutional neural networks (CNN) for visual semantic analysis of tumor regions for diagnosis support. The CNN is trained over a large amount of image patches (tissue regions) from WSI to learn a hierarchical part-based representation. The method was evaluated over a WSI dataset from 162 patients diagnosed with IDC. 113 slides were selected for training and 49 slides were held out for independent testing. Ground truth for quantitative evaluation was provided via expert delineation of the region of cancer by an expert pathologist on the digitized slides. The experimental evaluation was designed to measure classifier accuracy in detecting IDC tissue regions in WSI. Our method yielded the best quantitative results for automatic detection of IDC regions in WSI in terms of F-measure and balanced accuracy (71.80%, 84.23%), in comparison with an approach using handcrafted image features (color, texture and edges, nuclear textural and architecture), and a machine learning classifier for invasive tumor classification using a Random Forest. The best performing handcrafted features were fuzzy color histogram (67.53%, 78.74%) and RGB histogram (66.64%, 77.24%). Our results also suggest that at least some of the tissue classification mistakes (false positives and false negatives) were less due to any fundamental problems associated with the approach, than the inherent limitations in obtaining a very highly granular annotation of the diseased area of interest by an expert pathologist.

Keywords: Breast cancer, convolutional neural networks, deep learning, digital pathology, whole-slide imaging, invasive ductal carcinoma, handcrafted features

Further author information: (Send correspondence to Anant Madabhushi)
Anant Madabhushi: E-mail: anant.madabhushi@case.edu, Telephone: 1 (216) 368-8619

1. INTRODUCTION

Invasive ductal carcinoma (IDC) is the most common phenotypic subtype of all Breast cancers (BCa) comprising nearly 80% of them.[1] This is routinely identified by pathologists through visual analysis of tissue slides stained with hematoxylin and eosin (H&E). Assessment of disease aggressiveness (i.e. tumor grading) is usually limited to regions containing invasive cancer.[2] Hence, the first step in the histopathological characterization of resected breast tissue is to distinguish between tissue regions corresponding to invasive tumor and non-invasive or healthy tissues. Isolating IDC allows for further analysis of tumor differentiation via the well-known Bloom-Richardson and Nottingham grading schemes.[3] Invasive breast cancer detection is a time consuming and challenging task primarily because it involves a pathologist scanning large swathes of benign regions to ultimately identify the areas of malignancy. Precise delineation of IDC in WSI is crucial to the subsequent estimation of grading tumor aggressiveness and predicting patient outcome.

Breast cancer detection in histopathology images has been previously explored by other researchers.[4–9] Most of these approaches involve segmentation of histologic primitives (e.g. nuclei) and then extracting features characterizing the appearance or arrangement of these primitives to distinguish malignant from benign regions. Petushi et al.[4] presented a method for tissue microtexture classification of BCa by segmenting nuclei and extracting two textural features (surface density of nuclei and spatial position). The approach was used to distinguish between different grades of tumoral cells and stroma or adipose tissues. Naik et al.,[5] presented a methodology for automated detection and segmentation of glands and nuclei in BCa histopathology images. A total of 51 graph-based features from Voronoi diagrams, minimum spanning tree, and Delaunay triangulation. The nuclear centroids was employed to distinguish between regions of cancer from benign areas on whole slide images, yielding an overall accuracy of 80% in conjunction with a support vector machine classifier. Doyle et al.[6] used the graph-based features, 15 statistics gray level features, 16 second order statistical features, 64 Gabor filter features and 24 nuclear features per each of three color channels in HSI color space using three different windows sizes. Dundar et al.[7] presented a prototype system for automatically classifying breast microscopic tissues to distinguish between usual ductal hyperplasia and actionable subtypes (atypical ductal hyperplasia and ductal carcinoma in situ). Niwas et al.[8] evaluated color texture features of segmented nucleus in the core needle biopsy images by using Log-gabor complex wavelet bases. This approach results in different convolution maps of log-gabor filters for different scales and orientations from which two set of features are extracted first order statistical and second order statistical features.

Most of the previous approaches described above involve combining a large number of different types and complex handcrafted features to represent the visual content of BCa histopathology images. These approaches typically involve a series of pre-processing steps including detection, segmentation with the result that the final classification result is dependent on the accuracy of the preceding steps. Recently, generalized representation and feature learning approaches that do not incorporate domain knowledge have been employed in complex learning tasks.[10, 11] These methods address learning tasks like a fully learn-from-data approach avoiding segmentation and manual selection of handcrafted features.[12]

In recent years, deep learning (DL) architectures have gained popularity thanks to their success in different computer vision and pattern recognition tasks. These methods are an evolution of conventional neural networks.[10, 12] These approaches typically involve multiple non-linear transformations of the data, with the goal of yielding more abstract and ultimately more useful representations.[11] These methods are becoming increasingly popular, often outperforming traditional approaches that include handcrafted features for data representation and machine learning methods for learning task.[13–18] This has happened in the context of an unprecedented growth of available digital data (i.e. big data) and powerful computational resources.[19]

Digital pathology is one of the most recent instances of big electronic data and comes from the digitalization process of histopathology glass slides by digital scanners.[20] The digitized whole slide pathology images are often several gigabytes in size.[21] While digital pathology is being routinely used for clinical diagnosis in some parts of Europe,[22] in the US there is a growing interest in asking research questions anchored in digital pathology. Deep feature learning of very large digital pathology images provides a very unique opportunity to learn hidden patterns that may not be discernible by visual or human examination. In fact, deep learning architectures have been shown to be successful in the automated segmentation and classification of disease extent on digitized histopathology images. Cruz-Roa et al. in[23] presented a deep learning model based on sparse autoencoders for automated tumor

detection and visual analysis of basal-cell carcinoma. This deep learning model comprised a visual interpretable step to highlight relevant cancerous regions on histopathology images such as a digital staining. Other types of deep learning models include convolutional neural networks (CNN), which are a family of multi-layer neural networks particularly designed for use on two-dimensional data, such as images and videos.[12] CNN are feed-forward neural network whose architectures are tailored for minimizing the sensitivity to translations, local rotations and distortions of the input image.[24] Le et. al.[25] applied a two-layer neural network to quantitatively interrogate tissue sections for characterizing necrosis, apoptotic, and viable regions of Glioblastoma Multifrome (GBM). Recently, CNN models had been also applied to challenging task of automatic mitosis detection in BCa histopathology images.[26–28]

The main advantage of deep learning models is that they learn the most appropriate representation in a hierarchical manner as part of the training process. This is akin to the way pathologists analyze a histology slide across different resolutions and fields of view. However, most deep learning strategies involving histopathology involve applied a pixel level classification over relative small images. Our approach on the other hand employs a DL based classification on square tissue regions from WSI obtained by regular sampling, thereby enabling the application of the classifier over the entire canvas of the WSI. The dataset selected comprises 162 cases of patients diagnosed with IDC of BCa which 113 slides were selected for training and 49 slides for stand alone testing. We compared our approach against a traditional classification approach comprising handcrafted descriptors, global features (color and texture) and histopathology image features (nuclear textural and architectural). Our goal was to compare handcrafted feature approaches against a feature learning approach of deep learning.

The overall framework described in Figure 1 comprises the following steps: 1) grid sampling of image patches is performed over all tissue-containing regions in the WSI; 2) a Convolutional Neural Network is trained from sampled patches to predict the probability of a patch belonging to IDC tissue; and 3) finally a probability map built over the WSI highlighting the predicted IDC regions.

The rest of this paper is organized as follows: Section 2 describes the methodology; Section 3 presents the experimental evaluation process. Section 4 presents experimental results and discussion. Section 5 concludes with the main findings and future directions.

2. METHODOLOGY

Step 1. Image patch sampling

Each WSI is split into non-overlapping image patches of 100×100 pixels via grid sampling. Patches with mostly fatty tissue or slide background are discarded. For training purposes, regions containing IDC are manually annotated by a pathologist and used to generate a binary annotation mask. An image patch is considered to be a positive sample if at least 80% of the patch falls within the annotation mask, otherwise it used as a negative sample. Figure 2 illustrates an example of image patch sampling involving the WSI and pathologists annotations.
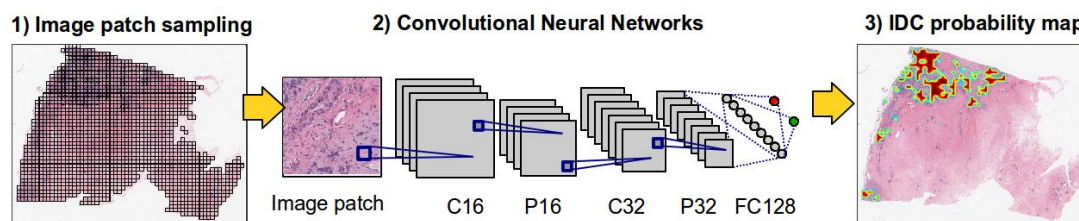

Figure 1. Overall framework for automated detection of IDC in WSI using CNN comprises: 1) Image patch sampling from WSI, 2) Convolutional Neural Networks to classify each image patch, and 3) complete probability map showing patches predicted by CNN to contain IDC with probability > 0.29.

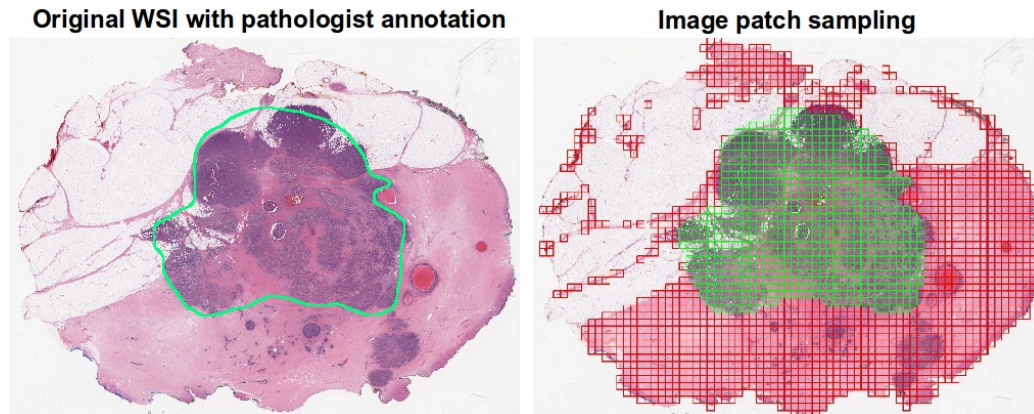**Original WSI with pathologist annotation**   **Image patch sampling**

Figure 2. Image patch sampling process. The original WSI with manual annotations from pathologists (left) is split into non-overlapping image patches via grid sampling only taking the tissue regions (red and green). Red patches correspond to positive examples of IDC and green patches correspond to negative examples.
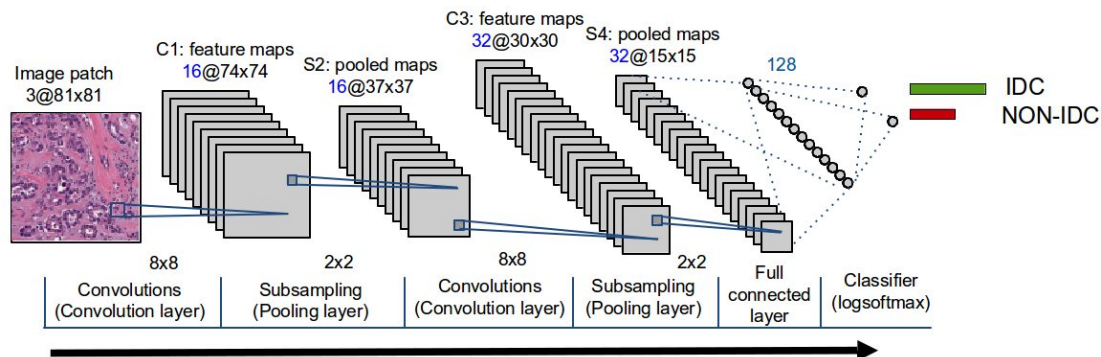

Figure 3. 3-layer CNN architecture composed by two layers of convolutional and pooling layers, a full-connected layer and a logistic regression classifier to predict if an image patch belongs to a IDC tissue or not.

Step 2. Convolutional Neural Networks

CNN[24] involve application of local feature detectors or filters over the whole image to measure the correspondence between individual image patches and signature patterns within the training set. Then, an aggregation or pooling function is applied to reduce the dimensionality of the feature space. The image patches collected in Step 1 are then used as inputs to a 3-layer CNN architecture (Figure 3) in which two layers are used for convolution and pooling while the remaining layer is fully-connected. The detailed steps are:

1. *Patch preprocessing:* Each image patch is converted from RGB to YUV color space and normalized to a mean of zero and variance of one. This step removes correlations of raw pixels, allowing to focus on properties that are not dependent on covariances, such as sparseness.[29] The motivation to do that are to accentuate differences between input features and accelerate gradient-based learning.[30]

2. *Convolutional layer:* The convolution process is an strategy to represent larger images (i.e. more than 64×64 pixels) since a set of learned features working as feature detectors. To do that we apply a 2D convolution of the input feature maps (i.e. image channels for the first layer) with a squared convolution kernel (i.e. filters or features), is defined as $y_j = \tanh(\sum_i k_{ij} * x_i)$, where x_i corresponds to the ith input feature map, k_{ij} is the convolution kernel, and y_j corresponds to j^{th} output feature map. The feature map looks like a salient map where learned features was detected. The $tanh(\cdot)$ function is used to rectify non-linearities in the convolution output, a common problem in the modeling of biological processes.[31] A

contrast normalization step is applied to each y_j independently to help reduce overfitting and generalize overall performance.[32]

3. *Pooling layer:* This stage allows to reduce original large dimension of image representation through a subsampling strategy which support local space invariances. Then, this layer applies a L2 pooling function over feature maps by a spatial windows without overlapping (i.e. pooling kernel). L2 pooling is used to allow for optimal learning of invariant features in each window.[33]

4. *Fully-connected layer:* This layer is typically applied in the top layer of a CNN architecture in order to capture complex relationships between high-level features. In this stage spatial information is ignored to learn correlation between different locations. Thus, the output of pooling layer is the input of a fully-connected layer which mixes them into a feature vector. This is like a well known perceptron neural network.

5. *Classification layer:* This final layer is a fully-connected layer with one neuron per each of the two classes (invasive or not) activated by a logistic regression model (i.e. softmax classifier). The whole CNN model is trained using Stochastic Gradient Descent to minimize the following loss function: $L(x) = -log\left[\frac{e^{x_i}}{\sum_j e^{x_j}}\right]$, where x_i corresponds to outputs of full-connected layer multiplied by logistic model parameters. Thus the outputs of CNN are the log likelihoods of class membership.

Step 3. IDC probability map

The exponential function is applied to each output value obtained from positive class neuron of classified patches by CNN to get values between 0 and 1, so that they could be interpreted as probabilities. For each patch we have the original coordinates on WSI and its corresponding probabilities, thus we stitch them to get the IDC probability map over WSI.

3. EXPERIMENTAL EVALUATION

3.1 Description of dataset

The data cohort comprises digitized BCa histopathology slides from 162 women diagnosed with IDC at the Hospital of the University of Pennsylvania and The Cancer Institute of New Jersey. All slides were digitized via a whole-slide scanner at 40x magnification (0.25 μm/pixel resolution). Operating on entire whole-slide histopathology images is intractable due to their extremely large size (on the order of 10^{10} pixels). In this work, each WSI was downsampled (by a factor of 16:1) to a resolution of 4 μm/pixel.

3.2 Ground truth annotation

The ground truth of IDC regions over digitized histopathology slides was obtained via manual delineation of the cancer region by an expert pathologist. These annotations were made using the freely available viewing software ImageScope from Aperio. Due to the large amount of time and effort needed to produce expert ground truth annotations, this study relies on relatively coarse annotations performed at 2x magnification or less. This led to the inclusion of some stromal and non-invasive tissue near the desired IDC tissue regions.

3.3 Image patch based dataset construction

The original data cohort of 162 slides was randomly split into three different subsets comprising 84 training (D1) and 29 validation (D2) cases for parameter exploration, and 49 test cases for final evaluation (D3). The patch-based datasets comprise 82,883 and 31,352 instances for training (D1) and validation (D2), i.e. 114,235 instances for full training (D1+D2), and 50,963 instances for testing (D3). Examples of positive (IDC) and negative (non-IDC) tissue regions from the training and test sets are shown in Figure 4.
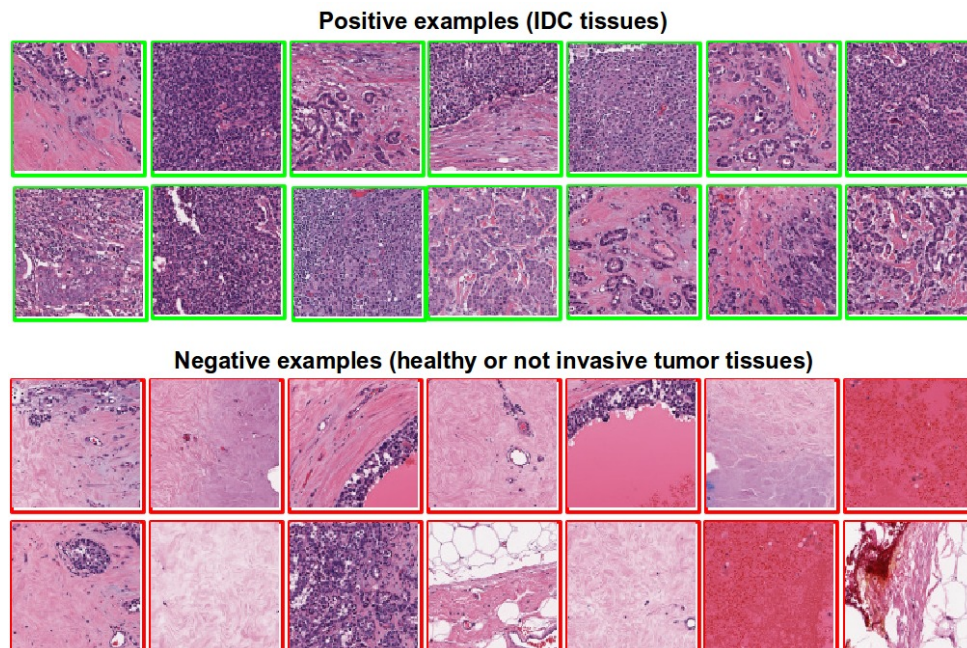
Figure 4. Example of image patches for positive (green) and negative (red) categories of the training set. As can be appreciated from the panels in Figure 4, substantial variability exists in the appearance of the image patches in both categories.

Table 1. Set of state-of-the-art handcrafted features used in this work to compare against the deep learning approach.

Handcrafted feature	Visual property captured
Gray Histogram (GH)[34]	Luminance
Fuzzy Color Histogram (FCH)[35]	Color
HSV Color Histogram (HSVCH)[36]	Color
RGB Histogram (RGBH)[34,37]	Color
JPEG Coefficient Histogram (JPEGCH)[36]	Color and Texture
Local Binary Partition Histogram (LBP)[38]	Texture
MPEG7 Edge Histogram (M7Edge)[39]	Texture
Haralick features[9]	Nuclear textural (NT)
Graph-based features[9]	Nuclear Architectural (NA)

3.4 Experimental setup

In order to evaluate our CNN based system, we compare it against a suite of state-of-the-art handcrafted features (global features or histopathology features). The two different approaches are described below.

CNN based framework: Our system adapts a 3-layers CNN architecture employing 16, 32, and 128 neurons, for the first and second convolutional-pooling layers and the fully-connected layer respectively. For all experiments, a fixed convolutional kernel of size 8×8 and pool kernel of size 2×2 were used.

Handcrafted features: Variety of handcrafted feature sets were selected in order to evaluate different visual properties of histopathology images such as staining, tissue appearance, morphological and architectural arrangement of cells by using color, texture and graph-based descriptors. The first set comprises of global features used in computer vision to represent color and texture information of whole images, having been previously successfully used in content-based image retrieval of natural[35–37] and histopathology images.[34] The second set comprises features used to describe nuclear arrangement and morphology via using Haralick and graph-based descriptors.[9,40] Table 1 details the complete set of handcrafted features, global features and histopathology image features, with the corresponding visual property captured by each one of them.

Table 2. Quantitative performance for classification of IDC and healthy tissues. The performance measures showed are Precision (Pr), Recall (Rc) or Sensitivity (Sen), Specificity (Spc), F-measure (F1) and Balanced Accuracy (BAC).

	Pr	**Rc/Sen**	**Spc**	**F1**	**BAC**
CNN	0.6540	0.7960	0.8886	**0.7180**	**0.8423**
FCH	0.7086	0.6450	0.9298	0.6753	0.7874
RGBH	0.7564	0.5956	0.9493	0.6664	0.7724
GH	0.7102	0.5240	0.9434	0.6031	0.7337
JPEGCH	0.7570	0.4646	0.9605	0.5758	0.7126
M7Edge	0.7360	0.4372	0.9585	0.5485	0.6979
NT	0.6246	0.2851	0.9547	0.3915	0.6199
LBP	0.7575	0.2291	0.9806	0.3518	0.6048
NA	0.6184	0.2413	0.9606	0.3472	0.6009
HSVCH	0.7662	0.2223	0.9821	0.3446	0.6022

Random Forest Classifier: The classifier employs a set of decision trees from a given dataset of labeled samples where each tree is developed from a bootstrap sample in a training set. The individual decision trees are constructed by selecting a random subset of attributes from which the best attribute is selected for the split. The random forest classifier is constructed by voting of the majority of individual trees classifiers in the forest. The main advantages of this method are that it is resilient to overfitting, requires few parameters to choose from (i.e. number of trees) and guaranteed to converge.[41]

3.5 Performance measures

For each WSI, all patches are classified as IDC or Non-IDC with a class label, of 1 or 0 respectively. Classification of image patches allows evaluation of IDC detection over an entire WSI. Hence, classification results are evaluated over a validation (D2) dataset for parameters selection, and over test dataset (D3) comparing the prediction of image patches and ground truth. By calculating true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN) we calculate a set of performance measures for IDC detection. Precision (Pr) allows estimate the proportion of IDC detected from total actual IDC regions. Recall (Rc), Sensitivity (Sen) or true positive rate allows estimate the proportion of IDC correctly predicted from whole IDC automatically predicted. Specificity (Spc), or true negative rate, is defined as the proportion of Non-IDC regions correctly predicted from total actual Non-IDC regions. The trade-off that occurs when simultaneously minimizing FP and FN can be assessed using the advanced classification performance measures F-measure (F1) and balanced accuracy (BAC), which are defined in equations 1 and 2 respectively.

$$F1 = \frac{2 \cdot Pr \cdot Rc}{Pr + Rc} \tag{1}$$

$$BAC = \frac{Sen + Spc}{2} \tag{2}$$

3.6 Parameter exploration

For both CNN and handcrafted features, a parameter exploration was performed using D1 and D2. For each parameter combination, training and validation was performed over D1 and D2 respectively. The parameter selection was done according to F1 measure. Optimal CNN parameter values for number of epochs, learning rate, learning rate decay and classification threshold of stochastic gradient descent algorithm were found to be 25, 1e-2, 1e-7, and 0.29 respectively. For all handcrafted features, the optimal number of trees for Random Forests (RF) classifier was found to be 200. The rest of parameters such as image patch size, step size for patch sampling (it affects number of patches sampled) and percentage of ground truth area for labeling positive examples, were empirically selected.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Evaluation deep learning

Our totally learn-from-data approach for automated IDC tissue region detection was quantitatively evaluated on D3, yielding a classification performance of 71.80% (F-measure) and 84.23% (BAC). This performance is detailed in Row 1 of Table 2 including precision, recall or sensitivity and specificity with 65.40%, 79.60% and 88.86% respectively.

The first experiment addresses the question of how similar is the ground truth manual annotation compared to the automated estimation of IDC. Figure 5 shows three WSI examples from D3. The images on the left side draw the manual annotations provided from pathologists of the IDC tissue regions. The images on the right side correspond to IDC probability map produced by our approach. Note that only regions greater than the optimal threshold of 0.29 (found during parameter exploration) are shown on each image.

Figure 9 shows the qualitative results of our CNN DL classifier, with green patches corresponding to true positives (TP), red patches to false negatives (FN), yellow patches corresponding to false positives (FP) and blue patches to true negatives (TN). A closer examination of the qualitative segmentation results suggests that some misclassifications (both FP and FN errors) are a result of imperfect manual annotation. For example, the proximity of many FP errors to TP regions suggests that some FP errors may be a result of coarse manual annotations used in this study. Other FP errors appear to correspond to confounding tissue classes (e.g. ductal carcinoma in situ), which can potentially be resolved by a multi-class approach. Similarly, many FN errors could also be the result of a coarse manual annotation (e.g. Figure 9; bottom) in which image patches containing stroma and adipose tissues are included in the annotated regions.

Because WSI was split in homogeneous tissue regions and each of them was classified as IDC with a probability value assigned by the CNN classifier, we can quantify the number of image patches assigned similar probabilities. This is shown in figure 6 by quantizing the range of probabilities between 0.3 and 1, all probabilities being assigned to 10 bins. Hence, these histograms shows the distribution of confidence of IDC tissue regions in WSI. The first column shows a small tissue sample where most of regions (squared image patches) were classified as IDC with a probability larger than 0.8. This means that our CNN model was confident about the class belongingness of those patches assigned a red color. The second column depicts a large tissue sample with a large tumor region with tissue regions classified as IDC with different degrees of confidence by our CNN classifier. Again, a large number of tissue regions were classified as IDC with a probability larger than 0.8. The third column reveals that for the image being displayed, a large number of tissue regions were classified as IDC. These were all assigned a large range of probability values, reflecting different degrees of confidence by the CNN classifier. Examples of image patches classified as IDC by the CNN classifier are shown in Figure 8.

4.2 Comparison against state-of-the-art handcrafted features

In order to evaluate the efficiency of our approach, we compared it against a set of handcrafted features intended to capture different visual properties in histology images (see Section 3.4). Table 2 shows quantitative classification performance on D3. CNN yields the best overall performance in terms of both F-measure and BAC (71.80%, 84.23%). By comparison, the best handcrafted features are Fuzzy Color Histogram (67.53%, 78.74%) and RGB Histogram (66.64%, 77.24%). However our approach yields improved F-measure and BAC results (by 4% and 6%, respectively) over the best handcrafted features. The remaining global features as well as the nuclear texture and architectural features do not yield competitive results. These results are not surprising since color-based global features are more representative of how a pathologist visually identifies IDC at lower resolutions. By contrast, features that describe nuclear texture and nuclear architecture characterize more detailed aspects of tissue morphology that have been shown to be important in distinguishing between different grades of IDC.[9]

In this work we are also interested in understanding the performance of the best hand-crafted features. Figure 7 illustrates IDC regions (dark red color) on a single WSI as identified by several different types of handcrafted features. Note that the subimages in Figure 7 are listed in the order of decreasing classification performance with respect to their corresponding handcrafted features as shown in Table 2. These results reveal that Fuzzy Color Histogram and RGB Histogram are the best performing handcrafted features for IDC detection. By contrast, textural and architectural features perform poorly and are unable to consistently detect the IDC regions.

5. CONCLUDING REMARKS

Automated detection of invasive ductal carcinoma is a challenging and relevant problem for breast cancer diagnosis. Accurate and reproducible IDC detection is an important task because it is often the first step in the diagnosis and treatment of BCa. The majority of previous works in histopathology tumor detection address this problem by combining different types of handcrafted features and machine learning algorithms. By contrast, we presented a novel deep learning framework for automated detection of IDC regions in WSI of BCa histopathology.

This work represents one of the first applications of deep learning models to WSI histopathology analysis. Furthermore, we present a novel application of CNN in digital pathology for visual image analysis and compare it against well-known handcrafted features. In addition to CNN, we show that only two color-based global features are able to accurately identify IDC on WSI BCa images. An important characteristic of our approach is that it is a truly learn-from-data approach for IDC detection. Our approach makes no assumptions in advance about the relevant visual features to represent the image content associated with IDC tissue. Additionally, since identifying regions of IDC does not perhaps require transparency or interpretability of the discriminating features in the same way that might be needed for a classifier performing breast cancer grading. In this problem, the reliability, confidence, speed and reproducibility of the classifier were of greater importance compared to feature interpretability, criterion successfully achieved by our CNN classifier.

One of the more interesting findings was that the misclassified tissue regions are due mainly for not detailed annotations from pathologists more than mistakes of our proposed method. The most remarkable characteristic of our approach is its reproducibility in different unseen WSI data which is very close to subjective binary manual annotations from a trained pathologist providing quantitative support of its decisions.

Future work will explore the effects of CNN models with deeper architectures (i.e. more layers and neurons) and validation on larger cohorts. However, the most interesting future directions are how automatically learn and identify visual features related to histopathology architecture and morphology to obtain quantitative semantic feature, such as percentage of necrotic tissue or tumor aggressiveness highlighting the related areas in WSI.

Acknowledgments

REFERENCES

[1] DeSantis, C., Siegel, R., Bandi, P., and Jemal, A., "Breast cancer statistics, 2011," *CA: A Cancer Journal for Clinicians* **61**(6), 408–418 (2011).

[2] Elston, C. and Ellis, I., "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," *Histopath.* **19**(5), 403–410 (1991).

[3] Genestie, C., Zafrani, B., Asselain, B., Fourquet, A., Rozan, S., Validire, P., Vincent-Salomon, A., and Sastre-Garau, X., "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 Res* **18**(1B), 571–6 (1998).

[4] Petushi, S., Garcia, F. U., Haber, M. M., Katsinis, C., and Tozeren, A., "Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer.," *BMC medical imaging* **6**, 14 (Jan. 2006).

[5] Naik, S., Doyle, S., Agner, S., Madabhushi, A., Feldman, M., and Tomaszewski, J., "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 (May 2008).

[6] Doyle, S., Agner, S., Madabhushi, A., Feldman, M., and Tomaszewski, J., "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in [*2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*], 496–499, IEEE (May 2008).

[7] Dundar, M. M., Badve, S., Bilgin, G., Raykar, V., Jain, R., Sertel, O., and Gurcan, M. N., "Computerized classification of intraductal breast lesions using histopathological images," *IEEE Transactions on Biomedical Engineering* **58**, 1977–1984 (July 2011).

[8] Niwas, S. I., Palanisamy, P., Zhang, W., Mat Isa, N. A., and Chibbar, R., "Log-gabor wavelets based breast carcinoma classification using least square support vector machine," in [*2011 IEEE International Conference on Imaging Systems and Techniques*], 219–223, IEEE (May 2011).

[9] Basavanhally, A., Ganesan, S., Feldman, M. D., Shih, N., Mies, C., Tomaszewski, J. E., and Madabhushi, A., "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 (Aug 2013).

[10] Bengio, Y., "Learning deep architectures for ai," *Foundations and Trends in Machine Learning* **2**, 1–127 (Jan. 2009).

[11] Bengio, Y., Courville, A., and Vincent, P., "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(8), 1798–1828 (2013).

[12] Arel, I., Rose, D. C., and Karnowski, T. P., "Research frontier: Deep machine learning–a new frontier in artificial intelligence research," *Comp. Intell. Mag.* **5**, 13–18 (Nov. 2010).

[13] Weston, J., Bengio, S., and Usunier, N., "Large scale image annotation: Learning to rank with joint word-image embeddings," *Machine Learning* **81**, 21–35 (Oct. 2010).

[14] Seide, F., Li, G., and Yu, D., "Conversational speech transcription using context-dependent deep neural networks," in [*Proc Interspeech*], 437–440 (2011).

[15] Glorot, X., Bordes, A., and Bengio, Y., "Domain adaptation for large-scale sentiment classification: A deep learning approach," in [*Proceedings of the Twenty-eight International Conference on Machine Learning (ICML'11)*], **27**, 97–110 (June 2011).

[16] Boulanger-Lewandowski, N., Bengio, Y., and Vincent, P., "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription," in [*Proceedings of the Twenty-nine International Conference on Machine Learning (ICML'12)*], ACM (2012).

[17] Ciresan, D. C., Meier, U., and Schmidhuber, J., "Multi-column deep neural networks for image classification," in [*Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], CVPR '12, 3642–3649, IEEE Computer Society, Washington, DC, USA (2012).

[18] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in [*Advances in Neural Information Processing Systems 25*], 1106–1114 (2012).

[19] Hey, T. and Trefethen, A., [*The Data Deluge: An e-Science Perspective*], 809–824, John Wiley & Sons, Ltd (2003).

[20] Ghaznavi, F., Evans, A., Madabhushi, A., and Feldman, M., "Digital imaging in pathology: Whole-slide imaging and beyond," *Annual Review of Pathology: Mechanisms of Disease* **8**(1), 331–359 (2013). PMID: 23157334.

[21] Madabhushi, A., "Digital pathology image analysis: opportunities and challenges (editorial)," *Imaging In Medicine* **1**, 7–10 (October 2009).

[22] Rojo, M. G., Daniel, C., and Schrader, T., "Standardization efforts of digital pathology in europe," *Analytical Cellular Pathology* **35**, 19–23 (Jan. 2012).

[23] Cruz-Roa, A., Arevalo, J., Madabhushi, A., and Gonzalez, F., "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection," in [*Medical Image Computing and Computer-Assisted Intervention MICCAI 2013*], *Lecture Notes in Computer Science* **8150**, 403–410, Springer Berlin Heidelberg (2013).

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

[25] Le, Q., Han, J., Gray, J., Spellman, P., Borowsky, A., and Parvin, B., "Learning invariant features of tumor signatures," in [*Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*], 302–305 (May 2012).

[26] Malon, C., Miller, M., Burger, H. C., Cosatto, E., and Graf, H. P., "Identifying histological elements with convolutional neural networks," in [*Proceedings of the 5th International Conference on Soft Computing As Transdisciplinary Science and Technology*], CSTST '08, 450–456, ACM, New York, NY, USA (2008).

[27] Malon, C. and Cosatto, E., "Classification of mitotic figures with convolutional neural networks and seeded blob features," *Journal of Pathology Informatics* **4**(1), 9 (2013).

[28] Ciresan, D., Giusti, A., Gambardella, L., and Schmidhuber, J., "Mitosis detection in breast cancer histology images with deep neural networks," in [*Medical Image Computing and Computer-Assisted Intervention MICCAI 2013*], Lecture Notes in Computer Science **8150**, 411–418, Springer Berlin Heidelberg (2013).

[29] Hyvärinen, A., Hurri, J., and Hoyer, P. O., [*Natural image statistics*], vol. 39, Springer (2009).

[30] LeCun, Y., "Learning invariant feature hierarchies," in [*Computer Vision–ECCV 2012. Workshops and Demonstrations*], 496–505, Springer (2012).

[31] Pinto, N., Cox, D. D., and DiCarlo, J. J., "Why is real-world visual object recognition hard?," *PLoS computational biology* **4**, e27 (Jan. 2008).

[32] Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y., "What is the best multi-stage architecture for object recognition?," in [*Computer Vision, 2009 IEEE 12th International Conference on*], 2146–2153 (Sept 2009).

[33] Le, Q., Ngiam, J., Chen, Z., hao Chia, D. J., Koh, P. W., and Ng, A., "Tiled convolutional neural networks," in [*Advances in Neural Information Processing Systems 23*], 1279–1287 (2010).

[34] Caicedo, J. C., *A Prototype System to Archive and Retrieve Histopathology Images by Content*, Master's thesis, Universidad Nacional de Colombia (2008).

[35] Han, J. and Ma, K.-K., "Fuzzy color histogram and its use in color image retrieval," *Image Processing, IEEE Transactions on* **11**, 944 – 952 (aug 2002).

[36] Lux, M. and Chatzichristofis, S. A., "Lire: lucene image retrieval: an extensible java cbir library," in [*Proceedings of the 16th ACM international conference on Multimedia*], MM '08, 1085–1088, ACM, New York, NY, USA (2008).

[37] Deselaers, T., *Features for Image Retrieval*, PhD thesis, RWTH Aachen University (2003).

[38] Ahonen, T., Hadid, A., and Pietikinen, M., "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**, 2037–2041 (2006).

[39] Messing, D., van Beek, P., and Errico, J., "The mpeg-7 colour structure descriptor: image description using colour and local spatial information," in [*Image Processing, 2001. Proceedings. 2001 International Conference on*], **1**, 670 –673 vol.1 (2001).

[40] Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., and Yener, B., "Histopathological image analysis: A review," *Biomedical Engineering, IEEE Reviews in* **2**, 147–171 (2009).

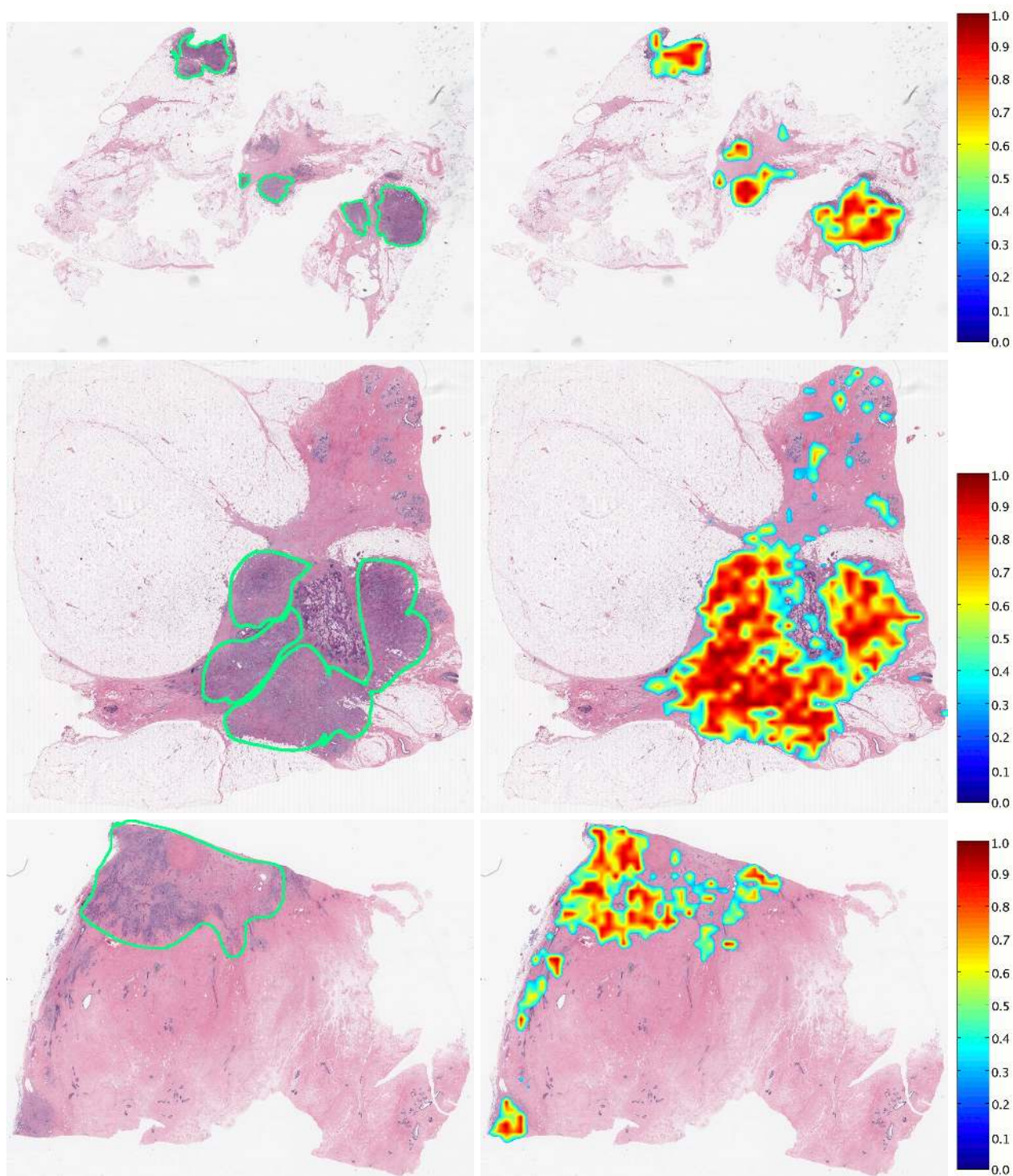[41] Breiman, L., "Random forests," *Machine Learning* **45**, 5–32 (Oct. 2001).

Figure 5. Three different BCa WSI from test dataset. The original manual annotations by pathologist (left) are very close to the probability map prediction from our model (right). The probability map shows the highest probabilities of IDC tissue regions in warm colors (red and orange) whereas the lowest probabilities of IDC regions in cold colors (green and blue).
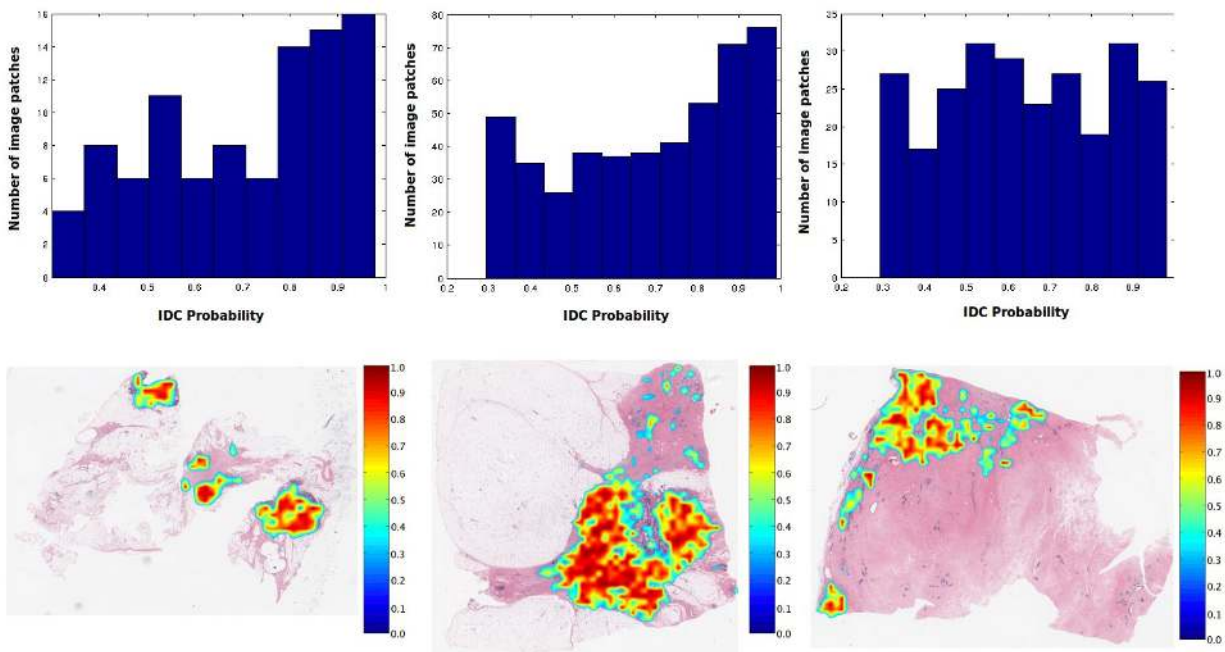
Figure 6. Histogram distribution of IDC probabilities (top) of their corresponding IDC probability map of WSI (bottom). These histograms depict how many tissue regions of each type of IDC tissue according to confidence to be invasive. WSIs a) and b) had more number of IDC regions with high probability (red areas) in proportion to whole tumor area, whereas WSI c) had a more homogenous distribution of IDC regions with different probability values into tumor.
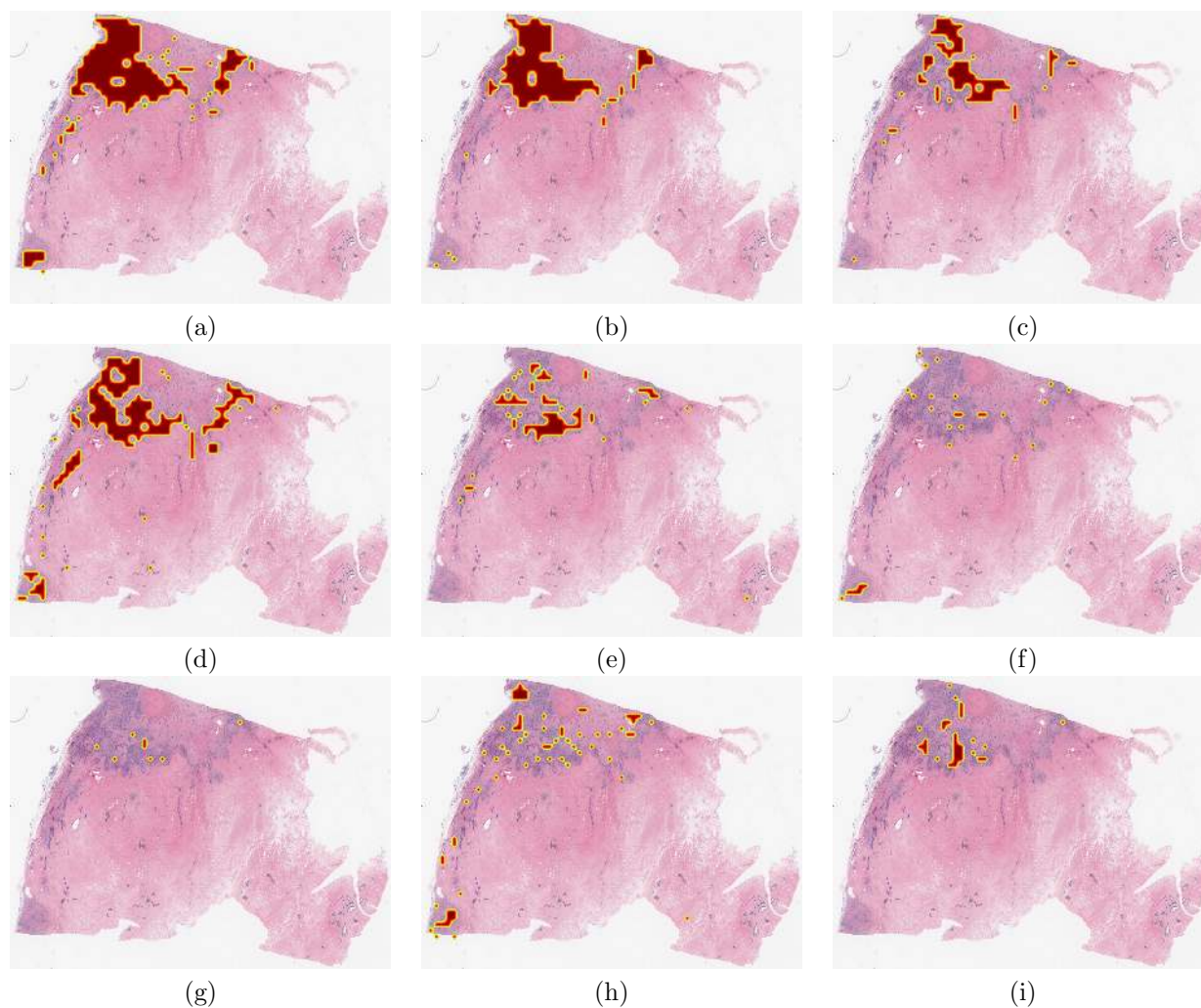
Figure 7. IDC tissue regions prediction (dark-red) for a WSI in descent order according to classification performance in testing for each of handcrafted features: a) Fuzzy Color Histogram, b) RGB Histogram, c) Gray Histogram, d) JPEG Coefficient Histogram, e) MPEG7 Edge Histogram, f) Nuclear Textural, g) Local Binary Partition Histogram , h) Nuclear Architectural, and i) HSV Color Histogram.
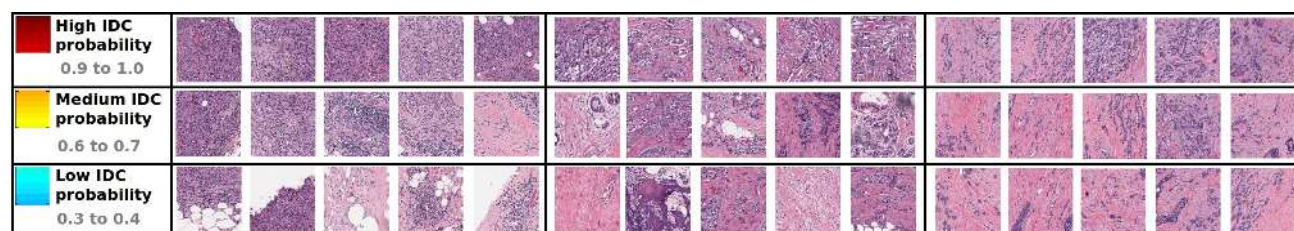

Figure 8. Example of image patches classified as IDC with different values of probability from three WSI of the Figure 6. High IDC probability 0.9 to 1.0 (top), Medium IDC probability 0.6 to 0.7 (middle), and Low IDC probability 0.3 to 0.4 (bottom).
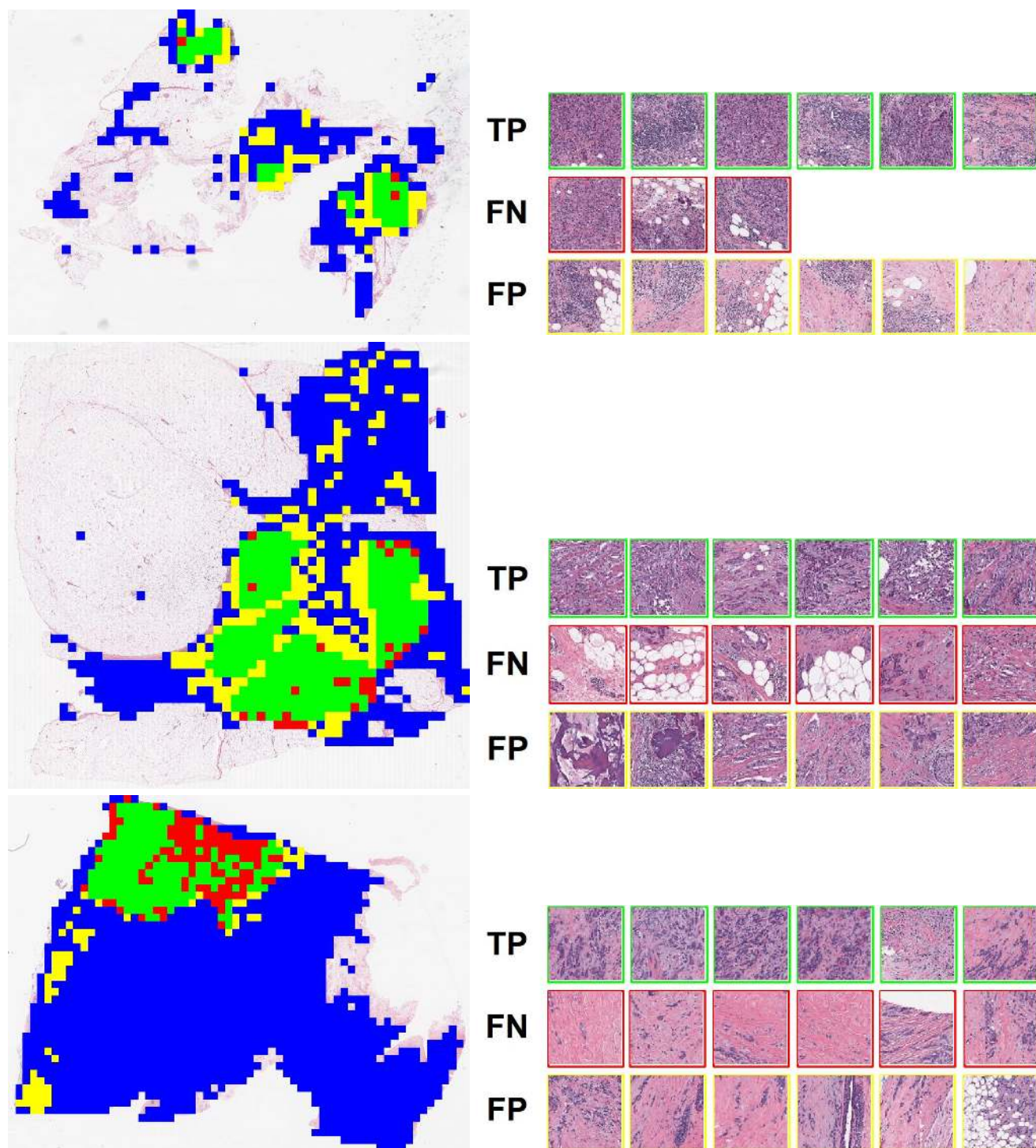
Figure 9. Detailed analysis of image patch classification by CNN model. The WSI on the left side shows the image patches sampled with their corresponding classification into true-positives (TP, green), false-negatives (FN, red) and false-positives (FP, yellow). The image patches on the right side show some examples for TP, FN and FP per each WSI.