Imitating Pathologist Based Assessment With Interpretable and Context Based Neural Network Modeling of Histology Images

Arunima Srivastava[1], Chaitanya Kulkarni[1], Kun Huang[2], Anil Parwani[3], Parag Mallick[4] and Raghu Machiraju[1]

[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA. [2]School of Medicine, Indiana University, Indianapolis, IN, USA. [3]Department of Pathology, The Ohio State University, Columbus, OH, USA. [4]Canary Center at Stanford for Cancer Early Detection, Stanford University, Palo Alto, CA, USA.

ABSTRACT: Convolutional neural networks (CNNs) have gained steady popularity as a tool to perform automatic classification of whole slide histology images. While CNNs have proven to be powerful classifiers in this context, they fail to explain this classification, as the network engineered features used for modeling and classification are ONLY interpretable by the CNNs themselves. This work aims at enhancing a traditional neural network model to perform histology image modeling, patient classification, and interpretation of the distinctive features identified by the network within the histology whole slide images (WSIs). We synthesize a workflow which (a) intelligently samples the training data by automatically selecting only image areas that display visible disease-relevant tissue state and (b) isolates regions most pertinent to the trained CNN prediction and translates them to observable and qualitative features such as color, intensity, cell and tissue morphology and texture. We use the Cancer Genome Atlas's Breast Invasive Carcinoma (TCGA-BRCA) histology dataset to build a model predicting patient attributes (disease stage and node status) and the tumor proliferation challenge (TUPAC 2016) breast cancer histology image repository to help identify disease-relevant tissue state (mitotic activity). We find that our enhanced CNN based workflow both increased patient attribute predictive accuracy (~2% increase for disease stage and ~10% increase for node status) and experimentally proved that a data-driven CNN histology model predicting breast invasive carcinoma stages is highly sensitive to features such as color, cell size, and shape, granularity, and uniformity. This work summarizes the need for understanding the widely trusted models built using deep learning and adds a layer of biological context to a technique that functioned as a classification only approach till now.

**KEYWORDS:** CNN, neural networks, modeling, cancer histology

Introduction

Automated classification of histology (hematoxylin and eosin stained) whole slide images (WSI) has been the subject of detailed research.[1,2] As a result, convolutional neural networks (CNNs) have recently gained steady popularity as the central technique to model and classify these images in the context of cancer.[3,4] They have shown to be precise and efficient classifiers in many different experiments pertaining to a variety of cancer subtypes. Our previous work assessing the predictive power of CNNs in the context of multiple patient attributes shows that CNN-based image models characterize disease staging more effectively than other trans-omics indicators.[5] Unfortunately, as is typical of sophisticated machine learning modeling techniques, one of the drawbacks the research community faces while utilizing CNNs for this purpose is the lack of interpretability of CNN-based features. Traditional CNNs learn incrementally from training data, generating an abstract set of features used by the network layers to classify regions in the image. These features do not have a precise translation to tissue structure, morphology or nuclei/cell organization, hence are not interpretable to clinicians or researchers who rely on the

use of these indicators to characterize disease. While observing activation of features in each network layer illuminates relationships between CNN features and pathology driven features,[4] there still is a dearth of an attempt to conclusively find interpretable signatures from CNN features. This work makes an effort to extend our foundational exploration of deep histology image models by aiming to optimize the power of CNNs while mimicking the output produced by pathologists who classify histology WSIs using qualitative and observable disease-specific indicators (Figure 1).

The interpretability of CNNs is a daunting task. It essentially involves de-convolving sophisticated learning operations to identify and map features in existing decipherable space. Thus, our workflow, which aims to perform effective and interpretable CNN modeling, focuses on two main tasks. First, it aims to reduce noise and irrelevant variance in the training data by utilizing only disease-relevant regions within the whole slide images to perform modeling. This step is motivated by pathologists' protocol where they demarcate regions of interest (ROIs) before analyzing tissue specimens.[6] As we model and analyze breast invasive carcinoma for the purposes of this work,

Manual feature assessment and use by pathologists
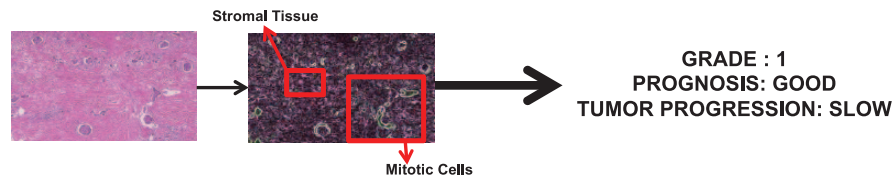
Automatic feature identification and patient/ tumor stratification by CNN modeling
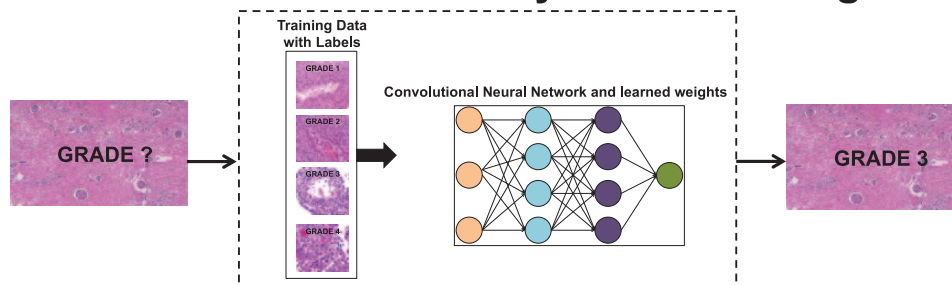
Automatic and interpretive feature identification and patient/tumor stratification by semantic CNN modeling

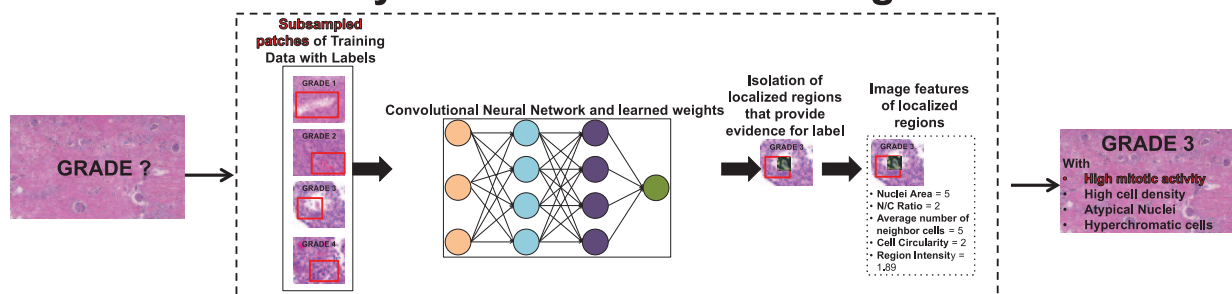**Figure 1.** Histology slide assessment and use by (1) pathologists, whose protocol dictates honing in on a region of interest, classifying different structures in the tissue, analyzing the state and structure of cells and finally performing prediction and delivering prognosis; (2) traditional CNN modeling, which trains from input images with a corresponding label, and is used to predict the probability of a new image belonging to these labels by performing high-dimensional modeling with self engineered features; and (3) our approach to interpretable context based CNN modeling, which intelligently selects only disease-relevant input images for training and modeling, finally resulting in label prediction using CNNs as well as an observable qualitative label signature for each new image.

we chose mitotic activity as a viable indicator for marking disease-relevant tissue region.[7,8] Second, we identify regions within the whole slide images that were most valued by the CNN when performing prediction of patient attributes (eg, American Joint Committee on Cancer [AJCC] stage). These "CNN relevant regions of interest (CNN-ROIs)" are further assessed in accordance with known image and morphology features. Both these tasks are achieved by utilizing a combination of specialized CNN architecture (AlexNet[9] with modulated parameters), methodology to visualize CNN learned weights (Class Activation Mapping—CAM[10]), and tools to extract image and morphology-specific features (pixel-wise k-means for dominant color extraction, Ilastik[11] and

CellProfiler[12] for shape, size and texture assessment). The dataset used to train a model to identify mitosis is the tumor proliferation challenge (TUPAC 2016) WSI repository and patient attribute prediction is performed on and using the Cancer Genome Atlas's Breast Invasive Carcinoma (TCGA-BRCA) histology dataset.[13]

Intelligent, mitotic activity based, sampling of training data resulted in significant improvement in the automatic prediction of patient staging and node status. Extracting an interpretable signature relates three specific types of qualitative features to the regions that were most informative to the CNN predicting staging of breast invasive carcinomas. CNN relevant ROI consistently presented unique dominant (pixel-wise most

frequently occurring) hues, specific shape, and size of cells and distinct morphological texture in the form of granularity and uniformity. While color/hue is an inherent property of staining different components of the tissue, aberrant cell size and shape is known to be a marker for tumor cells, and texture measures have proven to be highly distinct for histology images containing tumors and presenting specific disease subtypes.[14,15] This semantic context adds a layer of understandable characterization to CNN based models and helps identify critical components of histology images most relevant to this successful modeling.

Methods and Materials

The workflow to perform interpretable CNN modeling on breast invasive carcinoma is divided into four (4) distinct steps as visualized in Figure 2. Namely, the steps are (1) building a deep learning model to identify areas of high mitotic activity, (2) building a deep learning model to predict patients attributes (stages and node status) using intelligent sampling of training data where we retain areas of high mitotic activity only, (3) implementing a method for performing visualization and identification of "CNN relevant regions of interest (CNN-ROIs)" that are regions which were most informative to the CNN while predicting patient attributes and lastly (4) using CNN-ROIs to perform qualitative feature extraction which relates patient attributes to observable and interpretable features. Each of these steps is described in detail below in Figure 2.

Predicting areas of high mitotic activity

Data and pre-processing. The data used to train a neural network to identify regions of high mitotic activity are the auxiliary dataset provided by the Tumor Proliferation Assessment Challenge 2016 (TUPAC 2016), which was one of Medical Image Computing and Computer Assisted Intervention (MICCAI) 2016 grand challenges. It consists of images from 73-breast cancer cases aggregated from three pathology centers. All cases are represented by a number of image regions stored as TIFF images, with the mitotic regions (as classified by two pathologists) annotated. The WSIs are produced at 40x magnification and at the spatial resolution of 0.25 μm/pixel.

All whole slide images are tiled for parallel and faster processing. We used tile dimensions 224px x 224px to be cognizant of the structures we needed to identify while dispelling noise and artifacts. Since the average cell size in these tissues is ~40 to 100 pixels, (observationally), we concluded that 224px x 224px tile size would be apt as it would allow a minimum of ~2 to 3+ cells per tile taking into account varying distances between cells.

Normalization was performed employing the widely used Macenko[16] technique of normalizing histology images for quantitative analysis. This technique has been utilized successfully during histopathology assessment.[4,17] It uses the highest varying optical density (a transformation of the RGB vector describing absorbance) in the image to create a color transform applied to
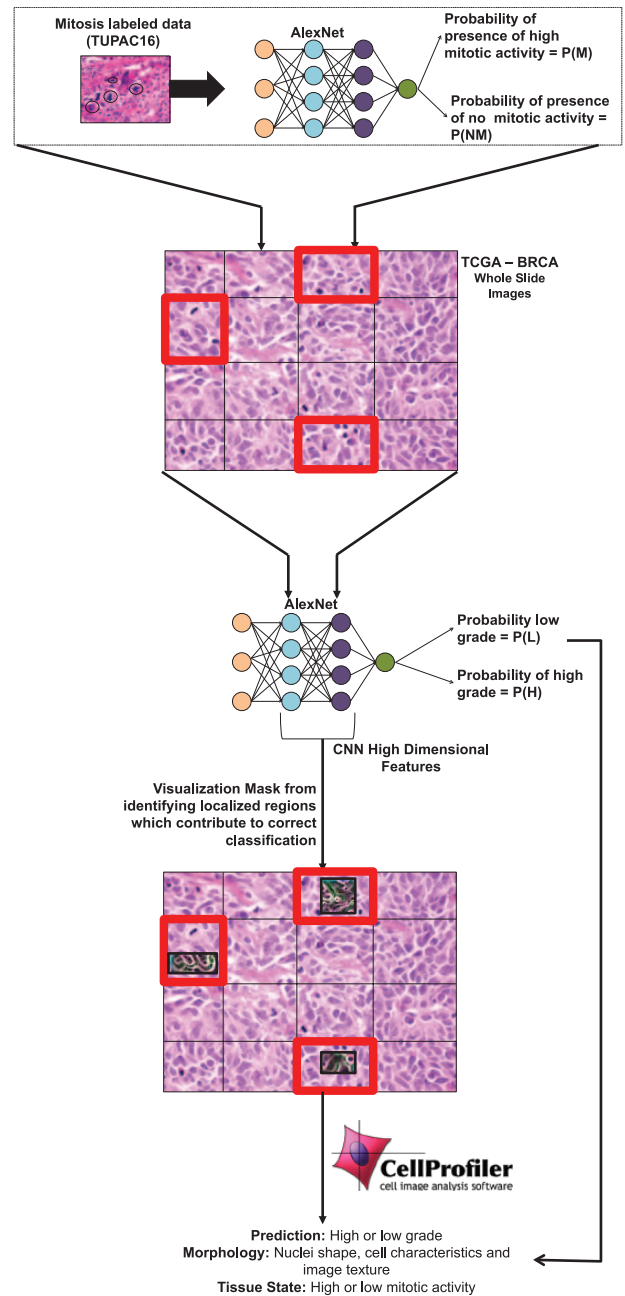

Figure 2. Workflow schematic for interpretable context-based CNN modeling of histology images. The main four steps of the workflow are as follows: (1) Build a model to select disease-relevant patches from the WSI based on tissue state; (2) perform intelligent sampling of WSI tiles for training of a deep learning model to predict patient attributes, based on whether they exhibit disease-relevant tissue state; (3) perform patient attribute prediction and extract CNN relevant regions of interest (CNN-ROIs), which were most informative to the deep learning model; and (4) assess these CNN-ROIs in terms of qualitative, observable features that associate model learning to interpretable features.

all images, followed by appropriate changes to the image histogram to capture most of the intensity dynamic range.

Neural network modeling. To build a model which identifies areas of high mitotic activity, we chose to utilize the traditional AlexNet architecture (Supplementary Table 1) with a few key

changes to adapt it to the nature of our problem (Supplementary Table 1). AlexNet is a popular and fundamental CNN architecture that has achieved widespread success in both traditional imaging challenges[9] and histology specific modeling.[3] Additionally, the resulting network is not extremely deep, which suits data that has a few underlying distinctive features in the presence of largely similar images. AlexNet uses rectified linear unit (ReLU) activation function to add non-linearity to the network and speed up training, and dropout instead of regularization to combat overfitting. Additionally, overlap pooling is employed to reduce the size of network. Our version of the model, the codebase, the necessary utilities and the complete trained models are available on request.

Selecting tiles presenting high mitotic activity. Utilizing the modified AlexNet architecture, and the mitosis annotated data from TUPAC2016, we built a model for identifying mitotic activity probability in tiles of histology images. This model was trained with tiles from TUPAC2016 whole slide images that were labeled according to the existence of mitotic cells within them. Upon successful training, the model was then employed to rank tiles from single whole slide images according to the probability of mitotic activity within that tile. The top 1% tiles that present high probability for mitotic activity are chosen for each whole slide image to represent disease-relevant areas for that patient.

Prediction of patient attributes based on regions of high mitotic activity

Data and pre-processing. The TCGA breast cancer study[18] is a well-characterized and thoroughly comprehensive experimental study of breast invasive carcinoma.[19,20] It consists of +1000 whole slide images from tumor sites and associated clinical information detailing AJCC stage, tumor subtypes and relevant mutational status is available. This work utilizes the same high-quality 163 whole slide images (105 patients) from the TCGA-BRCA compendium that were analyzed in our previous work on trans-omics features[5] as these images were histopathologically documented by pathologists and thus had extensive clinical information available. Each image is digitized at 40x and contains upward of 10 billion pixels.[21] The TCGA-BRCA compendium images were tiled and pre-processed employing the same methodology as detailed in the section above.

Neural network modeling. The mitotic activity prediction model described in the section above was employed on the 224px x 224px tiles from each TCGA-BRCA WSI and the top 1% of highly mitotic tiles are used to represent each whole slide image. Two separate models, using the AlexNet architecture as described above, were built. One using all the generated tiles (baseline) and the second using top 1% tiles showcasing mitotic activity, both training a predictor for patient staging and node status. The ensuing performance comparison ensured that the
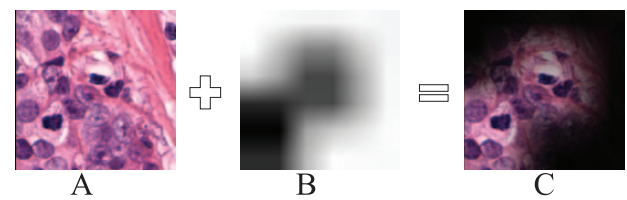

Figure 3. (A) Whole slide image tile containing mitotic cells. (B) Binary mask obtained using class activation mapping to highlight the discriminative localized region utilized by the CNN and (C) Composite tile, highlighting only the regions deemed "important" by the CNN, zeroing in on mitotic cells.

intelligent sampling of training data in accordance with disease-relevant tissue state was indeed producing superior results.

Visualizing CNN relevant ROIs

While there are multiple methods for visualizing a trained CNN's feature weights and network filters with respect to an input image, we choose to use a method, namely—CAM,[10] which identifies, across the entire trained CNN, localized regions that contribute most to the classification task. This technique utilizes a global average pooling layer at the penultimate step of the CNN in order to identify discriminative localized regions for each class. Global average pooling enables a generalized view across all network layers of the optical cues in an image that drive the model to a certain classification. Figure 3 presents an example of the visualization mask we obtain using CAM and the mitotic activity prediction model, for a histology image tile containing mitosis. We observe that the region highlighted using the CAM visualization mask contained mitotic cells and other (non mitotic) cells were ignored. Visualization masks such as these were generated for tiles from TCGA-BRCA histology images on which the disease stage prediction model was employed. Those visualization masks hone in on regions informative to stage prediction.

Image and morphological feature extraction

Once CAM enabled the visualization of CNN relevant ROIs, the concluding step involves extracting these regions and extracting qualitative features from them. We focus on three different types of features when assessing these CNN-ROIs to find a cohesive and interpretable signature that can be associated with the labels that a CNN model is aiming to predict. Namely, these three features are (a) color/hue, (b) cell size and shape, and (c) image texture. The procedure for extracting these features from CNN-ROIs is outlined below.

Finding dominant colors in CNN-ROI. The protocol for assessing histology images is highly dependent on the visible colors in the image (different colors of the staining mark for different structures within the tissue). It stands to reason that dominant colors visible in the CNN-ROIs evidence the predominance of

a certain structure relevant to the CNN modeling. The dominant colors are extracted from an image by utilizing unsupervised k-means clustering across the RGB vectors of all the pixels of an image.[22–24] With k = 4(accounting for distinct visible colors observed across a sampling of whole slide images), we extract clusters consisting of the RGB vectors for each pixel. The color corresponding to the largest cluster's centroid is then deemed the dominant color in the image. The method, corresponding sources, and our codebase is available on request. By identification of cells, tissues, and gaps in TCGA-BRCA whole slide images and subsequent extraction of RGB vectors from a sampling of these areas, we closely approximated the main colors visible and their corresponding RGB vectors. By euclidean distance-based proximity to these RGB vectors, we classify the dominant color to either be "purple" (cells), "pink" (muscle), or "white" (gaps or artifacts).

Assessing cell size and shape in CNN-ROI. Cells and their attributes are known to be relevant for pathologists to study and grade histology samples. We perform cell-specific segmentation in the tissue and analyze size and shape characteristics of the cells present. This is achieved by a combination of cell segmentation (performed by Ilastik) and object detection (performed by CellProfiler). Ilastik is a tool for image classification and segmentation. We train an Ilastik model on a subset of tiles from available whole slide images, where we manually demarcate cell regions. Ilastik then uses features based on color, intensity, and brightness and a random forest classifier to label pixels of the image if they are predicted as belonging to a cell. Using the results from Ilastik prediction, we extract a binary mask of the image, which identifies cell regions. This mask can be now used with CellProfiler and the "IdentifyPrimaryObjects" and "MeasureObjectSizeShape" module to extract size and shape features from the identified objects. The identification of objects is performed using a 2-class Otsu thresholding on the binary mask. These features include area, compactness, eccentricity, form factor, and Zernike features. For each CNN-ROI image, we extract the mean, median, and standard deviation of all of these features across the objects identified in the image for downstream analysis. This aggregation tuple is useful as due to the aberrant shape of tumor cells, the dynamic ranges of these features are highly relevant and distinctive. A full list of the relevant features and their descriptions is available in Supplementary Table 2.

Assessing CNN-ROI texture. Finally, to complete the qualitative signature for CNN-ROI, we extract the texture features of these images. This is also performed using a CellProfiler pipeline with the help of the module "MeasureTexture" and "MeasureGranularity." We perform texture extraction after separating the native CNN-ROI image to Red, Green and Blue channels, using the "ColortoGray" module in CellProfiler. Features include well-characterized texture features such as Haralick features. Similar to size and shape features, we aggregate them for each histology image using mean, median and standard deviation due to the varying dynamic ranges of each feature. A full list of the relevant features and their descriptions is available in Supplementary Table 3.

Results

This work and the resulting exploration can be divided to four distinct findings (a) successful predictions of enhanced mitotic activity in whole slide image tiles, (b) prediction of patient attributes using a model built with selected tiles that display evidence of mitotic activity, (c) isolating the tile regions discriminative for each class using class activation mappings, and (d) performing morphological assessment of selected regions to extract interpretable signatures for each class. The sections below highlight the main results for each aforementioned section.

Prediction of regions containing high mitotic activity

The AlexNet architecture was utilized to build a model to predict probability of a histology image tile containing mitotic activity. This model utilized the training data from the annotated whole slide images from the TUPAC16 challenge. As mentioned previously, the model trains on tiles of size 224px x 224px, and the labels are generated based on the existence of mitotic activity on the slide. This model seemed to successfully isolate mitotic activity within tiles as evidenced by all performance measures (~82% Accuracy, precision, recall and F-score) when employed on the testing subset of the dataset (20% of all tiles). A subset of tiles was also presented to a pathologist and the correct identification of mitosis was verified.

Patient attribute prediction

Patient stage and node status prediction models were trained with all generated tiles in TCGA-BRCA histology set as well as when trained with selected 1% tiles presenting mitotic activity for each whole slide image. For the prediction of patient stages, between a model trained from all tiles and a model trained from selected tiles presenting high mitotic activity, accuracy increased from 42.67% to 44.73%, and precision, recall, and F-Score increased from 25.09% to 26.99%, 24.26% to 27.13%, and 24.45% to 26.69%, respectively. Displaying a similar trend, a model predicting node status in these patients, utilizing all tiles, compared to a model trained on select tiles showed an increase in accuracy, precision, recall, and F-Score (28.64% to 38.17%, 24.09% to 28.86%, 23.02% to 28.19%, and 22.29% to 28.1%, respectively). As we observe from these results, the approach of intelligently sampling training data based on relevant tissue state (eg, mitotic activity) is justified as it shows marked improvement in the performance of a prediction model. We can hypothesize that this reduces noise from artifacts as well as ignores non-tumor areas and hones in on regions that pathologists would ideally focus
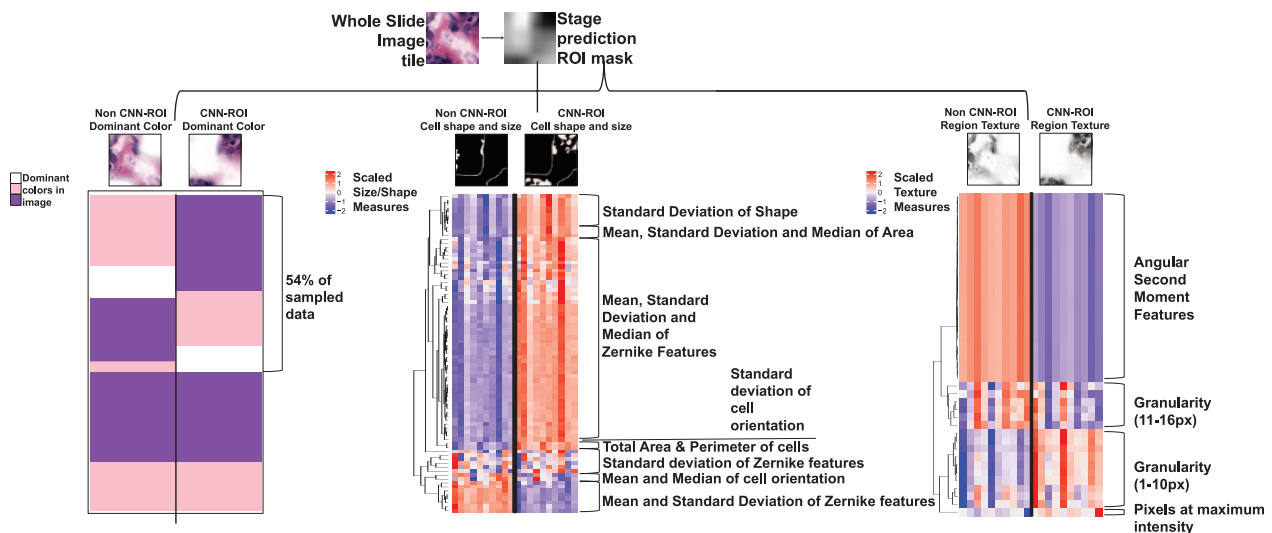
Figure 4. Results for the qualitative feature extraction (dominant color, cell size and shape, and image texture) of a random sampling of WSI tiles (~10 000 tiles across 10 patients, spanning multiple stages). Each heatmap compares CNN-ROI image and its inverse (Non CNN ROI) across all tiles. About 54% of tiles show dominant color changes from between CNN-ROI and Non CNN ROI images. Zernike features, orientation and area are distinct for CNN ROI cells. Uniformity and Granularity are image texture features that characterize CNN-ROI Images.

on when assigning attributes to a whole slide image, and consequently a patient.

Isolation and visualizing discriminative localized pixel regions

As described above, using the CAM technique, visualization masks were generated for each tile that highlighted the ROI to the CNN. The tiles selected by the mitotic activity predictor are used to train and test a CNN model to predict staging and for each prediction, an associated ROI mask is generated. The technique was tested by employing the CAM technique to the mitosis-predicting model as well, and on observation, we confirmed that the ROI masks were highlighting mitotic cells and ignoring typical circular non-mitotic cells. ROI masks are used to generate CNN-ROI images and qualitative features are extracted from these regions to assign an interpretable signature to the same.

Interpretable signatures of CNN-ROIs

In the penultimate step of this analysis, once we obtain the regions that were important to the CNN when defining the final prediction for patient stage, we can assess these regions interpretably to isolate signatures that represent the model and its encompassing labels. While multiple observable facets are available for exploration of the CNN-ROIs, we focus primarily on three aspects of these patches. Namely, (1) dominant colors, (2) cell size and shape, and (3) texture features. Utilizing the tools, as described in the methods section above, we assess these tuple attributes for both CNN-ROI and non-ROI patches for each tile. Comparing the two facilitates the identification of the unique signatures as distinguished by the CNN. Figure 4 presents an overall comparison of all three qualitative features, between the CNN-ROI and non-ROI patches from a sampling of 10,000 tiles, spanning 10 patients and multiple stages.

Dominant colors comparison between CNN-ROI and Non-ROI. In over 54% of the samples tiles, the dominant colors between CNN-ROI and Non ROI patches were distinct. A majority of these listed "purple" as the dominant color in the CNN-ROI patch and "pink" or "white" as the dominant color in the corresponding non-ROI patch. This provides evidence for the fact that hyperchromaticity of cells is a factor of distinction during the decision-making of the CNN patient staging model.

Cell size and shape. Specific features relating to shape, area, and orientation were observed to be distinct (twofold change) between CNN-ROI and non-ROI patches for each tile across the sampled set. These features included—standard deviation of compactness (cells in CNN-ROI images have highly varying numbers of close and well-defined enclosed structures), mean, median and standard deviation of cell area (cells in CNN-ROI images are bigger and vary more in terms of area), standard deviation of minimum and maximum feret diameter (cells shape varies more in CNN-ROI images) and multiple moments of Zernike shape features. This provides evidence for the hypothesis that pleomorphic, aberrant, atypical, and large cells characterize patient staging according to the predictive CNN model.

Image texture. Lastly, the texture measures (multiple features of granularity and texture angular second momentum (ASM)[25] describing structure and uniformity of texture respectively, were consistently distinct (twofold change) across CNN ROI versus non-ROI patches. Consistently, all different moments of ASM, which describes uniformity, are drastically lower in CNN-ROIs versus the non-ROIs, which present high uniformity. This is consistent with our previous findings, as pleomorphic cells contained within CNN-ROI patches are not well ordered, which would result in this texture feature presenting lower values. Granularity on the other hand, showcases the opposite trend

(higher in CNN-ROI vs. non- ROI), as it describes the size distribution of objects across a certain pixel scale, which provides further evidence that there is a higher concentration of cells within the CNN-ROI patches than the non-ROI patches. These findings are consistent with previous results and further establish the qualitative feature signature built for this model.

Discussion

The goal of this manuscript was to understand and interpret the unique signatures between subtypes of whole slide images, as understood and interpreted by a CNN model. We wished to quantify, whether controlling the input and assessing the output manifests in better performance and understanding of deep histology models. To this end, we identified meaningful regions in whole slide images by automatically classifying tissue state (high mitotic activity), which is a crucial facet of histologically assessing breast cancer. Following which, we performed experiments which predicted staging with the selected neural network model, using all tiles and only highly mitotic tiles. The performance enhancement in prediction confirmed our correct selection of whole slide image patches. Lastly, we used these selected tiles and ROIs as identified by the deep learning model to explore and understand the exact features of regions the CNN deemed interesting, and on which it based its predictions. We believe this work will enable the community to better understand the high dimensional neural network models that have slowly become the standard in automatic histology modeling. Additionally, it has the potential to identify new histopathological features that are markers of disease as understood by data driven deep modeling.

Author Contributions

AS, RM, PM Conceived and designed the analysis; RM, AP and KH contributed to the interpretation of the results; AS, CK performed the analysis and AS, RM and PM wrote the paper.

REFERENCES

1. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Nasir R, Yener B. Histopathological image analysis: a review. *IEEE Rev Biomed Eng*. 2009;2: 147–171.
2. Riordan DP, Varma S, West RB, Brown PO. Automated analysis and classification of histological tissue features by multi-dimensional microscopic molecular profiling. *PLoS ONE*. 2015;10:e0128975.
3. Xu Y, Jia Z, Wang LB, et al. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinformatics*. 2017;18:281.
4. Araujo T, Aresta G, Castro E, et al. Classification of breast cancer histology images using convolutional neural networks. *PLoS ONE*. 2017;12:e0177544.
5. Mallick P, Huang K, Machiraju R. Building trans-omics evidence: using imaging and "omics" to characterize cancer profiles. *Pac Symp Biocomput*. 2018;23:377–387.
6. Nagarkar DB, Mercan E, Weaver DL, et al. Region of interest identification and diagnostic agreement in breast pathology. *Mod Pathol*. 2016;29:1004–1011.
7. Wang H, Cruz-Roa A, Basavanhally A, et al. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *J Med Imaging*. 2014;1:034003.
8. Al-Janabi S, van Slooten HJ, Visser M, van der Ploeg T, van Diest PJ, Jiwa M. Evaluation of mitotic activity index in breast cancer using whole slide digital images. *PLoS ONE*. 2013;8:e82576.
9. Krizhevsky A, Sutskever I, Hinton GE. Alexnet. *Adv Neural Inf Process Syst*. 2012. doi: 10.1016/j.protcy.2014.09.007.
10. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. arXiv:1512.04150; 2016.
11. Sommer C, Straehle C, Kothe U, Hamprecht FA. Ilastik: interactive learning and segmentation toolkit. Paper presented at: IEEE International Symposium on Biomedical Imaging: From Nano to Macro; March 30-April 2, 2011; Chicago, IL. New York: IEEE.
12. Carpenter AE, Jones TR, Lamprecht MR, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol*. 2006;7:R100.
13. Tomczak K, Czerwiska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*. 2015;19:A68–A77.
14. He L, Long LR, Antani S, Thoma GR. Histology image analysis for carcinoma detection and grading. *Comput Methods Programs Biomed*. 2012;107:538–556.
15. Yu KH, Zhang C, Berry GJ, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun*. 2016;7:12474.
16. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. Paper presented at: IEEE International Symposium on Biomedical Imaging: From Nano to Macro; June 28-July 1, 2009; Boston, MA. New York: IEEE.
17. Vink JP, Van Leeuwen M, Van Deurzen CH, De Haan G. Efficient nucleus detector in histopathology images. *J Microsc*. 2013;249:124–135.
18. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors. *Nature*. 2012;490:61–70.
19. Yao F, Zhang C, Du W, Liu C, Xu Y. Identification of gene-expression signatures and protein markers for breast cancer grading and staging. *PLoS ONE*. 2015;10:e0138213.
20. Kim D, Li R, Dudek SM, Ritchie MD. Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer. *J Biomed Inform*. 2015;56:220–228.
21. Gutman DA, Cobb J, Somanna D, et al. Cancer digital slide archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J Am Med Inform Assoc*. 2013;20:1091–1098.
22. Spannbauer A. iPhone App Icons. GitHub Repos. 2018. https://github.com/AdamSpannbauer/iphone_app_icon
23. Adorio EP. Python, finding nearest matching color name for RGB values. 2012. http://my-other-life-as-a-programmer.blogspot.com/2012/02/python-finding-nearest-matching-color.html.
24. Jdiscar. RGB Value to Nearest Color Name (Python). *GitHub Repos*. 2014. https://gist.github.com/jdiscar/9144764.
25. Broad Institute. Welcome to CellProfiler's documentation! 2017. http://cellprofiler-manual.s3.amazonaws.com/CellProfiler-3.0.0/index.html.