EBioMedicine 27 (2018) 317–328

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images

Pegah Khosravi[a,b,1], Ehsan Kazemi[c,1], Marcin Imielinski[d,e,f,g], Olivier Elemento[a,b,d,g,*], Iman Hajirasouliha[a,b,d,g,*]

[a] Institute for Computational Biomedicine, Weill Cornell Medical College, NY, USA
[b] Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA
[c] Yale Institute for Network Science, Yale University, New Haven, CT, USA
[d] Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medical College, NY, USA
[e] Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, NY, USA
[f] The New York Genome Center, NY, USA
[g] The Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA

ARTICLE INFO

Article history:
Received 27 October 2017
Received in revised form 21 December 2017
Accepted 21 December 2017
Available online 28 December 2017

Keywords:
Biomarkers
Classification
Convolutional Neural Network
Deep learning
Digital pathology imaging
Tumor heterogeneity

ABSTRACT

Pathological evaluation of tumor tissue is pivotal for diagnosis in cancer patients and automated image analysis approaches have great potential to increase precision of diagnosis and help reduce human error.

In this study, we utilize several computational methods based on convolutional neural networks (CNN) and build a stand-alone pipeline to effectively classify different histopathology images across different types of cancer.

In particular, we demonstrate the utility of our pipeline to discriminate between two subtypes of lung cancer, four biomarkers of bladder cancer, and five biomarkers of breast cancer. In addition, we apply our pipeline to discriminate among four immunohistochemistry (IHC) staining scores of bladder and breast cancers.

Our classification pipeline includes a basic CNN architecture, Google's Inceptions with three training strategies, and an ensemble of two state-of-the-art algorithms, Inception and ResNet. Training strategies include training the last layer of Google's Inceptions, training the network from scratch, and fine-tunning the parameters for our data using two pre-trained version of Google's Inception architectures, Inception-V1 and Inception-V3.

We demonstrate the power of deep learning approaches for identifying cancer subtypes, and the robustness of Google's Inceptions even in presence of extensive tumor heterogeneity. On average, our pipeline achieved accuracies of 100%, 92%, 95%, and 69% for discrimination of various cancer tissues, subtypes, biomarkers, and scores, respectively. Our pipeline and related documentation is freely available at https://github.com/ih-_lab/CNN_Smoothie.

1. Introduction

Evaluation of microscopic histopathology slides by experienced pathologists is currently the standard procedure for establishing a diagnosis and identifying the subtypes of different cancers. Visual-only assessment of well-established histopathology patterns is typically slow, and is shown to be inaccurate and irreproducible in certain diagnosis cases of tumor subtypes and stages (Mosquera-Lopez et al., 2015).

Several recent studies attempted to employ machine learning approaches for determining subtypes of malignancies (Esteva et al., 2017; Yu et al., 2016). These computational approaches can be complementary with other clinical evaluation methods to improve pathologists' knowledge of the disease and improve treatments (Felipe De Sousa et al., 2013; Beck et al., 2011). For example, previous studies have shown more accurate diagnosis results are derived by integrating information extracted from computational pathology with patients' clinical data for various cancer types such as prostate cancer (Bhargava et al., 2011; Doyle et al., 2012), lung cancer (Hamilton et al., 2015), breast cancer (Wang et al., 2013; Dong et al., 2014), colorectal cancer (Korbar et al., 2017), and ovarian cancer (Janowczyk et al., 2012). In particular, computerized image

[*] Corresponding authors.
[1] Both authors contributed equally to this work.

https://doi.org/10.1016/j.ebiom.2017.12.026
2352-3964/

processing technology has been shown to improve performance, correctness, and robustness in histopathology assessments (Lemaître et al., 2015).

While new advanced approaches have improved image recognition (e.g., normal versus cancerous), the image interpretation of heterogeneous populations still suffers from lack of robust computerization approaches (Razzak et al., 2017; Carneiro et al., 2017; Gurcan, 2016; Jiang et al., 2016). Current available automatic methods focus on classification of just one type of cancer versus the corresponding normal condition. Although these studies achieved reasonable accuracy in detecting normal or cancerous conditions in specific kind of cancers, leveraging methods such as training Convolutional Neural Networks (CNNs)(LeCun et al., 1998), they have certain limitations which we address in this work:

1. Developing *ensemble* deep learning methods to employ state-of-the-art algorithms for improving training approaches in diagnosis and detection of various cancer subtypes (e.g., adenocarcinoma versus cell squamous lung cancer).
2. Improving the speed of deep learning, and investigating the trade-offs between performance (i.e., the size of the training set) and efficiency (i.e., the training speed).
3. Making decisions on selecting proper neural networks for different types of datasets.

One of the main challenges of computational pathology is that tumor tissue images often vary in color and scale batch effects across different research laboratories and medical facilities due to differences in tissue preparation methods and imaging implements (Kothari et al., 2013). Furthermore, erroneous evaluation of histopathology images and decision-making using tissue slides containing millions of cells can be time-consuming and subjective (Yu et al., 2016; Kothari et al., 2013). In this regard, utilization of the deep learning approaches with sufficient number of images to untangle color information can improve the computational approaches within a reasonable amount of time.

In addition, cancer is known to be a heterogeneous disease. i.e., a high degree of genetic and phenotypic diversity exists “within tumors” (intra-tumor) and “among tumors” (inter-tumor) (Polyak, 2011). Tumor heterogeneity leads to an important effect of disease progression and resistant responses to targeted therapies (Hardiman et al., 2016). We also aim to evaluate deep learning approaches for discrimination of digital pathology images from intra- and inter-tumor heterogeneous samples.

Deep learning approaches are emerging as leading machine-learning tools in medical imaging where they have been proven to produce precious results on various tasks such as segmentation, classification, and prediction (Greenspan et al., 2016). In this paper, we present an innovative deep learning based pipeline, CNN_Smoothie, to discriminate various cancer tissues, subtypes, and their relative staining markers and scores. We utilize the pathological images which stained by immunohistochemical markers of tumor differentiation to train CNNs for analyzing and identifying specific clinical patterns in different staining markers and scores of breast and bladder cancers. In addition, we applied deep learning methods on immunohistochemistry (IHC) and hematoxylin & esoin (H&E) stained images of squamous cell carcinoma and lung adenocarcinoma to investigate the performance of various classifiers.

This is a comprehensive study of applying a wide range of CNN architectures (all integrated in a single pipeline) on histopathology images from multiple different datasets. We evaluate performance of different architectures to detect and diagnosis of tumor images. Our results clearly demonstrate the power of deep learning approaches for distinguishing different cancer tissues, subtypes, IHC markers and their expression scores. Source codes and documentation of our pipeline containing training, evaluation and prediction methods are publicly available at https://github.com/ih-_lab/CNN_Smoothie.

2. Materials and Methods

2.1. Histopathology Images Resource

Our datasets come from a combination of open-access histopathology images, The Stanford Tissue Microarray Database (TMAD) and The Cancer Genome Atlas (TCGA). A total of 12,139 whole-slide stained histopathology images were obtained from TMAD (Marinelli et al., 2007). TMA database enables researchers have access to bright field and fluorescence images of tissue microarrays. This archive provide thousand human tissues which are probed by antibodies simultaneously for detection of protein abundance (immunohistochemistry; IHC), or by labeled nucleic acids (in situ hybridization; ISH) to detect transcript abundance. The extracted data included samples from three cancer tissues: (1) lung, (2) breast, comprising five biomarker types (EGFR, CK17, CK5/6, ER, and HER2), and (3) bladder with four biomarker types (CK14, GATA3, S0084, and S100P). Characteristics of all three cohorts and the comprised classes are summarized in Table 1. From the extracted TMA datasets, one dataset is stained by H&E method (BladderBreastLung) and one dataset is stained by both H&E and IHC methods (TMAD-InterHeterogeneity). The remaining datasets (BladderBiomarkers, BreastBiomarkers, BladderScores, and BreastScores) are stained by IHC markers including different polyclonal antiserums such as CK14, GATA3, S0084, S100P, EGFR, CK17, CK5/6, ER, and HER2 for their related proteins which play critical roles in tumor progression.

The markers are widely used in clinical immunohistochemistry as biomarkers for detection of various neoplasm types (Higgins et al., 2007; Vandenberghe et al., 2017). Several studies have acquired the expressions of biomarkers in biopsy samples of various cancer types to improve the distinction of specific pathological subtyping and understanding of molecular pathways of different cancers. For example, we can refer to the attempts made to discriminate morphologic subtyping of non-small call lung carcinoma (NSCLC), lung adenocarcinoma (LUAD) versus lung squamous cancer (LUSC) (Scagliotti et al., 2008; Khayyata et al., 2009; Conde et al., 2010; Fatima et al., 2011). Antiserums staining tissue are sub-classified according to the staining grade. Each tissue sample in this cohort was scored by a trained pathologist using a discrete scoring system (0, 1, 2, 3). A score zero represents no significant protein expression (negative) because there is no staining color, whereas a score three indicates high expression. Positive results were scored based on both the extent and the intensity of staining. For score three, intense staining was required in more than 50% of the cells. Other scores including one and two staining comprise in fewer than 50% of the total cells (Higgins et al., 2007).

We also obtained the TCGA (Network et al., 2012, 2014) images by extracting them from the Cancer Digital Slide Archive (CDSA) (Gutman et al., 2013) that is accessible to the public and, at the time of writing this, hosts 31,999 whole-slide images from 32 cancer tissues. For the purpose of this study, we analyze 1520 H&E stained whole-slide histopathology images as well as 1629 H&E stained high resolution image patches (40× magnification) of two TCGA lung cancer subtypes (i.e., LUAD versus LUSC).

2.2. Classification and Diagnostic Framework

This study presents a framework (see Fig. 1) to discriminate different cancer types, subtypes, immunohistochemistry markers, and marker staining scores of histopathology images (Table 1). For the first step of our study, the stained whole-slide images with 1504 × 1440 and 2092 × 975 pixels were obtained from TMA and TCGA databases, respectively. Note that we did not use any pre-processing

Table 1
Eight datasets are selected to assess the performance of the pipeline across different conditions.

Number	Datasets	The database representation	Labels of inputs and outputs	Dataset size	Class size
1	BladderBreastLung	H&E-stained images for bladder, breast and lung cancers	Discrimination of different tissues of cancer (bladder, breast, and lung)	3 classes and 1918 images	Bladder: 543, breast: 962, lung: 413
2	BladderBiomarkers	IHC-stained images of cancer biomarkers comprising GATA3, CK14, S100P, and S0084 in bladder cancer	Discrimination of different types of biomarkers (GATA3, CK14, S100P, and S0084)	4 classes and 2139 images	GATA3: 542, CK14: 514, S100P: 544, S0084: 539
3	BreastBiomarkers	IHC-stained images of cancer biomarkers including ER, CK17, CK5/6, EGFR, and HER2 in breast cancer	Discrimination of different types of biomarkers (ER, CK17, CK5/6, EGFR, and HER2)	5 classes and 2542 images	ER: 637, CK17: 639, CK5/6: 635, EGFR: 307, HER2: 324
4	TMAD-InterHeterogeneity	H&E- and IHC-stained whole-slides of adenocarcinoma and squamous cell lung cancers	Discrimination of different subtypes of cancer (adenocarcinoma vs. squamous cell lung tumors) for TMAD images	2 classes and 860 images (H&E: 572, IHC: 288)	Adenocarcinoma: 637, squamous cell: 223
5	TCGA-IntraHeterogeneity	H&E-stained high-resolution image patches of adenocarcinoma and squamous cell lung tissues	Discrimination of different subtypes of cancer (adenocarcinoma vs. squamous cell lung tumors) within high-resolution image patches of TCGA images	2 classes and 1629 images	Adenocarcinoma: 845, squamous cell: 784
6	TCGA-InterHeterogeneity	H&E-stained whole-slides images of adenocarcinoma and squamous cell lung tissues	Discrimination of different subtypes of cancer (adenocarcinoma vs. squamous cell lung tumors) within whole-slide images of TCGA images	2 classes and 1520 images	Adenocarcinoma: 761, squamous cell: 759
7	BladderScores	IHC-stained images with various staining scores comprising Score 0, Score 1, Score 2, and Score 3 in bladder cancer	Discrimination of different staining scores (Score 0, Score 1, Score 2, and Score 3) of biomarkers	4 classes and 2137 images	Score 0: 680, Score 1: 235, Score 2: 284, Score 3: 938
8	BreastScores	IHC-stained images with various staining scores including Score 0, Score 1, Score 2, and Score 3 in breast cancer	Discrimination of different staining scores (Score 0, Score 1, Score 2, and Score 3) of biomarkers	4 classes and 2543 images	Score 0: 1817, Score 1: 263, Score 2: 184, Score 3: 279

methods such as color deconvolution to separate the images from staining (van der Laak et al., 2000) or any watershed algorithms to identify cells (Vincent and Soille, 1991) manually. The whole images directly used as the input to the pipeline.

The images are then divided in different classes based on the classification aims and the CNN algorithms are applied on these classes. For each class, images divided in three groups including training, validation, and test groups. For this purpose, 70% of all images are allocated to the training group and 30% of the remaining images devoted to validation and test sets. Although the ratio is not always stringent due to image limitation, we set the train:validation:test ratio to 70:15:15 while the training, validation, and test sets are not identical and contain different images. The allocation ration was selected 70/30 since various references (Akram et al., 2015; Lam et al., 2014) in tumor classification concept have separated their datasets into training and testing set with the composition of 70/30 which yields the best results.

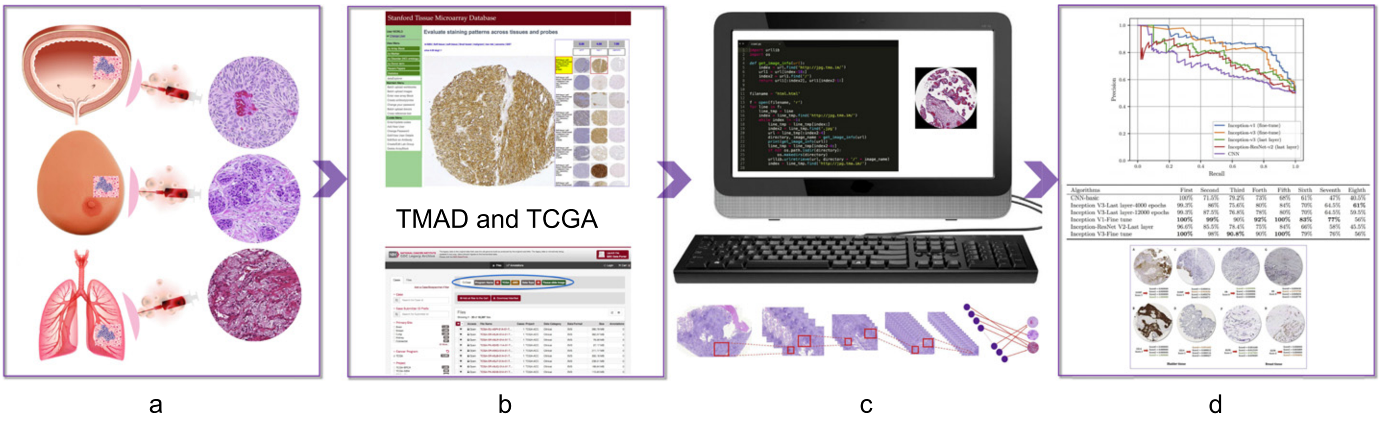
2.3. Convolutional Neural Networks (CNNs)

In this study, we use various architectures of CNN algorithms (i.e., deep neural network methods). The most well-known traditional neural networks is called the multi-layered perceptrons (MLP) that have many layers of transformations. A neural network which contains multiple hidden layers, in between the input and output, is considered a “deep neural network”. A survey on deep neural network approaches and their application in medical image analysis is described in (Litjens et al., 2017).

Convolutional neural networks have become the technique of choice for using deep learning approaches in medical images analysis since the first time in 1995 by (Lo et al., 1995). Before deep neural networks (DNN) gained popularity, they were considered hard to train large networks efficiently for a long time. Their popularity indebted to good performance of training DNNs layer by layer in an unsupervised manner (pre-training), followed by supervised fine-tuning of the stacked network. In this project, we are going to utilize DNNs for histopathology image analysis. They are the most successful type of models for image analysis because they comprise multiple layers which transform their input with convolution filters (Bengio et al., 2007; Hinton et al., 2006; Hinton and Salakhutdinov, 2006).

A convolutional neural network is a type of deep, feed-forward artificial neural networks which obtain simple features as input and then return them into more complex features as output (Simard et al., 2003). The CNNs use the spatial structure of images to share weights across units and benefit of some parameters to be learned a rotation, translation, and scale invariance. So, each image patch around each image can be extracted and directly used as input to CNNs model. One of the very first successful application of deep CNNs was shaped for hand-written digit recognition in LeNet (LeCun et al., 1998). Then, various novel techniques were developed for training deep networks through efficient ways. The contribution of Krizhevsky and his colleagues (Krizhevsky et al., 2012) to the ImageNet challenge made a watershed advance in core computing systems. They proposed a new architecture of CNN, AlexNet, that won the mentioned competition in December 2012. Currently, the CNNs with deeper architecture and hierarchical feature representation learning have made dramatic changes in object recognition related problems (Russakovsky et al., 2014; Krizhevsky et al., 2012; Szegedy et al., 2015; Simonyan and Zisserman, 2014; Chen et al., 2015).

Simonyan and Zisserman (2014) explored much deeper networks containing 19-layer model which called OxfordNet and won the ImageNet challenge of 2014. Then, Szegedy et al. (2015) introduced a 22-layer network named GoogLeNet which later referred to as Inception and made use of so-called inception blocks (Lin et al., 2013). This Inceptions family architectures allow a similar function



Algorithms	First	Second	Third	Forth	Fifth	Sixth	Seventh	Eighth
CNN-basic	100%	71.5%	79.2%	73%	68%	61%	47%	40.5%
Inception V3-Last layer-4000 epochs	99.3%	86%	75.6%	80%	84%	70%	64.5%	**61%**
Inception V3-Last layer-12000 epochs	99.3%	87.5%	76.8%	78%	80%	70%	64.5%	59.5%
Inception V1-Fine tune	**100%**	**99%**	90%	**92%**	**100%**	**83%**	**77%**	56%
Inception-ResNet V2-Last layer	96.6%	85.5%	78.4%	75%	84%	66%	58%	45.5%
Inception V3-Fine tune	**100%**	98%	**90.8%**	90%	**100%**	79%	76%	56%

Fig. 1. This flowchart demonstrates the pipeline, which includes extracting data, training and evaluation of CNN algorithms, and prediction of various classes. a: tumor image preparation of biopsy samples, b: extracting biopsy-derived tissue slides from TMA and TCGA databases, c: analysis of images using CNN_smoothie, and d: evaluation of the algorithms performance and annotation of the output results.

to be represented with less parameters. Also, the ResNet architecture (He et al., 2016) won the ImageNet challenge in 2015 and consisted of so-called ResNet-blocks. However, the majority of recent landmark studies in the field of medical imaging use a version of GoogLeNet called Inception-V3 (Esteva et al., 2017; Gulshan et al., 2016; Liu et al., 2017). Recently Esteva et al. (2017) utilized a deep CNN as a pixel-wise classifier which is computationally demanding in cancer research to detect melanoma malignant with high performance.

The advantage of Google's Inception architectures is their good performance even under strict constraints on memory and complexity of computational problems. For example, GoogLeNet (Szegedy et al., 2015) used 5 million parameters, which represented a significant reduction in parameters with respect to AlexNet (Krizhevsky et al., 2012) and VGGNet (Simonyan and Zisserman, 2014). This is the reason of using Inception networks in big data analysis where huge amount of data needed to be processed at reasonable time and computational cost (Movshovitz-Attias et al., 2015; Schroff et al., 2015). Various version of Inceptions are the attempt of Google team to scale up deep networks. For example, in 2014 (Szegedy et al., 2015) proposed Inception-V1 and then in 2015 (Ioffe and Szegedy, 2015) revealed batch normalization. Then, the authors proposed Inception-V2; they presented a derivative form of Inception-v2 which refers to the version in which the fully connected layer of the auxiliary classifier is also-normalized. Then, they call the new model as Inception-v3 which comprising Inception-V2 plus batch-normalization (BN) auxiliary (Szegedy et al., 2016). The Google team also tried various versions of the residual version of Inception such as Inception-ResNet-V1 which is high computational cost version of Inception-v3. Another version is Inception-ResNet-V2 that its computational cost matches with the newly introduced Inception-V4 network (Szegedy et al., 2017).

2.4. Transfer Learning

Image classification was one of the first areas in which deep learning made a principal contribution to medical image analysis. In medical image classification multiple images are considered as inputs with a single diagnostic result as output (e.g., cancerous or normal). A dataset comprising diagnostic image samples have typically bigger sizes with smaller numbers compared to those in computer vision. The popularity of transfer learning for such applications is therefore not surprising that essentially refers a method with two popular and have been widely applied strategies on medical data. Transfer learning refers to pre-train a network architecture on a very large dataset and use the trained model for new classification tasks for a dataset with limited size.

The first strategy includes using a pre-trained network as a feature extractor. A major benefit of this method is not requiring a deep network to be trained and the extracted features smoothly applied to the existing image analysis pipelines (Litjens et al., 2017). The second strategy is fine-tuning a pre-trained network (Litjens et al., 2017). Empirical investigation about different strategies have revealed conflicting results. For example, Antony et al. (2016) showed that fine-tuning clearly outperformed feature extraction, achieving 57.6% accuracy in multi-class grade assessment of knee osteoarthritis versus 53.4%. While, Kim et al. (2016) showed that using pre-trained network as a feature extractor slightly outperformed fine-tuning in cytopathology image classification (70.5% versus 69.1%). Besides, two recent published papers presented fine-tuned method by pre-trained version of Google's Inception-V3 architecture on medical data and achieved a high performance close to human experts (Esteva et al., 2017; Gulshan et al., 2016). In addition, CNNs developers also train their own network architectures from scratch instead of using pre-trained networks as the third strategy. For instance, Menegola et al. (2016) compared few experiments using training from scratch to fine-tuning of pre-trained networks, and indicated that fine-tuning worked better for a small data set (i.e., 1000 images of skin lesions).

Given the prevalence of CNNs in medical image analysis, we focused on the most common architectures and strategies with a preference for far deeper models that have lower memory footprint during inference. In this study, we compare various strategies and architectures for application of CNN algorithm to assess their performance on classification of histopathology images. These are included basic architecture of CNN, pre-trained network (training the last layer) of Google's Inceptions versions 1 and 3, fine-tunning the parameters for all layers of our network derived from the data using two pre-trained version of Google's Inception architectures, and the ensemble of two the state of the art algorithms (i.e., Inception and ResNet).

2.5. Implementation Details

In order to deploy the central architecture, we used a Tensorflow (version: 1.4.0) (Abadi et al., 2016) framework. This open source software solution was originally created by the Google Brain team for machine learning applications on textual data sets. The framework supports running the training operation of the network on graphics processing units (GPUs) or traditional computer microprocessors (CPUs). This platform also supports several machine learning algorithms with the same optimizer. The Python programming language version 2.7 (including libraries such as numpy, cv2, matplotlib, and random) was used for all aspects of this project. Also, TF-Slim which is a library for defining, training, and evaluating models in TensorFlow was used in this study. This library enables defining complex networks quickly and concisely while keeping a model's architecture transparent and its hyperparameters explicit.

A fixed image (with the JPG format) size of 20 × 20 pixels was selected for CNN-basic architecture to ensure that all images have the same size and large cells were entirely captured. CNN with the basic architecture consist of a two layer CNN network with max-pooling blocks; at the end we have two fully connected layers. The image sizes for Inception-V1, Inception-V3, and Inception-ResNet were automatically selected as 224 × 224 , 229 × 229, and 229 × 229 pixels by the algorithms, respectively.

All design and training of our method was performed on a desktop computer running the Mac operating system. This computer was powered by an Intel i5 processor at 3.2 GHz, 16 GB 1867 MHz DDR3 of RAM, and a solid state hard drive which allowed ruling out bottlenecks in these components. Although we were able to run all experiments without a GPU (≈7 Gigabyte data), high levels of system memory and a fast storage medium make this application faster since it depends on loading a significant number of medical images for training and validation.

The experimental section is split into two parts: While the aim of the first part of experiment is to reach reliable classification accuracy on the digital pathological images, the goal of the latter is to apply various architectures of CNNs to better understand the choice for the parameters.

2.6. *Metrics for Performance Evaluation of Algorithms*

To assess the performance of different algorithms and to select the most appropriate architectures for a given task and classification aim, we carried out several experiments on the reference datasets. precision-recall curves (PRCs) are typically generated to evaluate the performance of a machine learning algorithm on a given dataset. In this study, precisions and recalls are presented by average for multi-class datasets. Furthermore, to quantify and comparing the performance of various architectures of CNN algorithm on a sample dataset, commonly used accuracy measures, receiver operating characteristic (ROC), were estimated. The ROC curve depicts by plotting the true positive rate (TPR) versus the false positive rate (FPR) at various threshold settings. We defined a hard threshold range (e.g., from 0 to 1 across a dataset with two classes) for confidence of our predictions. Then, we observe a trade-off between two operating characteristics, TPR and FPR, by varying this threshold. Probability scores can be used to trade-off precision with recall. In this way, each input image is predicted by the algorithm using a probability score which could be smaller, equal, or greater probability score across the threshold range. Meanwhile, allocation of an image with less confidence to a specific class decreases the precision and increase the recall and vice versa.

Therefore, accuracy is measured by the area under the ROC curve (AUC) (Hanley and McNeil, 1982; Zawistowski et al., 2017).

To evaluate the algorithms performance on all datasets, we also used of two defined measures of accuracy retrieval curve (ARC): true number (TNu) and false number (FNu) (Khosravi et al., 2015). Therefore, to measure the algorithm performances for theses datasets, the accuracy, defined as TNu/(TNu + FNu) which is the fraction of correctly identified images among all images identified by algorithms, while retrieval is the total number of images identified by algorithms.

We also address other measures such as Cohen's kappa (Cohen, 1960) which is a popular way of measuring the accuracy of presence and absence predictions because of its simplicity and its tolerance to zero values in the confusion matrix (Allouche et al., 2006). The kappa statistic ranges from 1 to +1, where +1 indicates perfect agreement and values of zero or less indicate a performance no better than random (Cohen, 1960).

The other measure is the Jaccard coefficient measures similarity as the intersection divided by the union of the objects. The Jaccard coefficient ranges between 0 and 1; it is 1 when two objects are identical and 0 when the objects are completely different (Chen et al., 2016).

The Log-loss or cross entropy which is defined as $-\sum_j t(j|x)\log_2 \frac{p(j|x)}{t(j|x)}$ where $p(j|x)$ is the probability estimated by the method for example x and class j, and $t(j|x)$ is the true probability of class j for x (Bottou and GO, 1998; Drish, 2001). It is used to obtain a solution for a wide variety of loss functions and mathematically convenient because it can be computed for each example separately (Prashanth et al., 2017; Fogel and Feder, 2017; Zadrozny and Elkan, 2001).

3. Results

For the purpose of evaluating our pipeline, we obtained 9649 IHC stained whole-slide images as well as 2490 H&E stained histopathology images of lung, breast, and bladder cancers from TMAD. We also obtained 1520 H&E stained whole-slide histopathology images and 1629 H&E stained high resolution image patches of squamous cell carcinoma and lung adenocarcinoma from TCGA project. In summery, we used eighth different datasets comprising 26 classes (See Table 1). As demonstrated in Table 2, we utilized six state-of-the-art CNN architectures. The first three datasets cover the tasks that are primarily designed for setting up the pipeline (CNN_Smoothie) across different conditions (i.e., discrimination of different cancer tissues and markers). The other datasets are designed to assess the pipeline application and refer to challenging problems in clinical context. In addition of investigating different algorithms, we studied the effect of *step number* and *training strategies* on the accuracy and compared the performance of various architectures of CNN algorithm for classification and detection of tumor images.

3.1. *Evaluation of Various CNN Architectures in Pathological Tumor Images*

In this section, we present details of our evaluations on various CNN architectures. There are two basic subjects in analysis of digital histopathology images including classification and segmentation (Xu et al., 2017). We restricted the evaluations to image-based classification. Also, the basic architecture of CNN was utilized as well as Inception-V1 and Inception-V3 architectures (with fine-tuning the parameters for the last layer as well as all the layers). In addition, we evaluated the ensemble of Inception and ResNet (Inception-ResNet-V2) on all datasets.

Our results show that CNN_Smoothie is able to detect different cancer tissues, subtypes, and their related markers with highly reliable accuracy which depends on the dataset content, dataset size, and the selected algorithm (Table 2). For example, the pipeline can detect various cancer tissues by about 100% accuracy (Tumor tissue discrimination dataset in Table 2). While, the results of cancer subtype detection are varied from 61% to 100% based on the selected database, algorithm architecture, and the presence of heterogeneity in a tumor image (Tumor subtype discrimination datasets in Table 2). In addition, separating various bladder immunohistochemical markers results in 71.5% to 99% accuracy for CNN-basic and Inception-V1 fine-tune, respectively (bladder biomarker discrimination dataset in Table 2). Application of the mentioned algorithms on breast immunohistochemical markers lead to 79.2% and 90% accuracy, respectively (breast biomarker discrimination dataset in Table 2).

Table 2
The results of six state-of-the-art architectures of deep learning algorithms on various datasets using ARC. The numbers are measured based on TNu and FNu and represent accuracy percentages. The bold fonts indicate the best classification accuracies on datasets.

Algorithms	Tumor tissue discrimination	Bladder biomarker discrimination	Breast biomarker discrimination	Lung tumor subtype discrimination (TMAD images)
CNN-basic	100%	71.5%	79.2%	73%
Inception V3-Last layer-4000 steps	99.3%	86%	75.6%	80%
Inception V3-Last layer-12000 steps	99.3%	87.5%	76.8%	78%
Inception V1-Fine tune	**100%**	**99%**	90%	**92%**
Inception-ResNet V2-Last layer	96.6%	85.5%	78.4%	75%
Inception V3-Fine tune	**100%**	98%	**90.8%**	90%

Algorithms	Lung tumor subtype discrimination (TCGA intra-images)	Lung tumor subtype discrimination (TCGA inter-images)	Score discrimination in bladder	Score discrimination in breast
CNN-basic	68%	61%	47%	40.5%
Inception V3-Last layer-4000 steps	84%	70%	64.5%	**61%**
Inception V3-Last layer-12000 steps	80%	70%	64.5%	59.5%
Inception V1-Fine tune	**100%**	**83%**	**77%**	56%
Inception-ResNet V2-Last layer	84%	66%	58%	45.5%
Inception V3-Fine tune	**100%**	79%	76%	56%

Closer look at the Inception-V1 result of bladder cancer (99%) shows S0084 and S100P were misclassified with GATA3 and S0084, respectively, in two cases out of 200 cases. Moreover, the Inception-V1 result (90%) for discrimination of breast biomarkers revealed that all 10% contradictions have happened between CK17 and CK5/6 due to high similarity between them. This result is in concordant to previous studies such as (Tang et al., 2008) that compared different IHC markers in breast cancer and showed CK17 and CK5/6 have similar expression patterns.

We configured three datasets (BladderBreastLung, BladderBiomarkers, BreastBiomarkers) to set up the pipeline for all the proposed experiments (pathologists typically know cancer tissues or type of markers in advance). As expected, the algorithms were successfully able to discriminate various tissues of cancer with 100% accuracy (Table 2). Furthermore, we designed the experiments to investigate whether keeping the background color might have the potential to introduce certain inherent biases in datasets and affect the result for discrimination of various markers. The slides across BladderBiomarkers and BreastBiomarkers datasets are stained with different IHC staining colors. However, the results show that Inception architectures (V1 and V3) provide accuracies more than 90% in case of the colored version of the dataset (Table 2). When designing the experiments, we were concerned that the convolutional neural networks might only learn with biases associated to the colors, but the results showed the algorithm's adaptability in the presence of color information, and their ability to learn higher level of structural patterns typical to particular markers and tumors. This result is in concordance with a previous study that compared three dataset types based on different configurations (i.e. segmented, gray and colored) (Mohanty et al., 2016). Mohanty et al. (2016) showed that the performance of the model using segmented images is consistently superior than gray-scaled images, but slightly lower than colored version of the images.

The low concordance of the classification results (by algorithms) for BladderScores and BreastScores datasets (Table 2) to the labels that were determined by pathologists, could be related to the high heterogeneity within tumor cell populations of each slide. Moreover, because we did not have enough images to separate each classes individually, we blended all markers with the same score together (e.g. class score 0 contains GATA3-score 0, CK14-score 0, S100P-score 0, and S0084-score 0). Thus, discrimination of various images in these classes became more challenging. The algorithms are then trained for each score disregard to the markers.

Our findings are in agreement with previous studies which showed significant variability between pathologists in score discretization (Vogel et al., 2011; Roche et al., 2002; Perez et al., 2006; Gavrielides et al., 2011; Bueno-de Mesquita et al., 2009; Bloom and Harrington, 2004; Kaufman et al., 2014) and confirmed that 4% of negative and 18% of positive cases are misclassified even for one type of marker. Consequently, S0084 marker had the minimum cases of misclassification in bladder cancer. Furthermore, the minimum misclassification is related to the score 3 and EGFR marker which is a well known basal marker for breast cancer therapy (Lakhani et al., 2005). Results are comparable with those ones which classified by expert pathologists despite the difficulty of the task (Vandenberghe et al., 2017).

Although medical images are mostly interpreted by clinicians, the accuracy of their interpretation is reduced due to subjectivity, large variations across interpreters, and exhaustion (Greenspan et al., 2016; Webster and Dunstan, 2014). We reviewed BreastScores and BladderScores datasets and the content images that are labeled as negative and positive scores. We perceived the low concordance in our result also could be indeed due to significant human errors in labeling, particularly among positive scores (i.e. scores 1, 2, or 3) (Fig. 2).

In this regard, we categorized the image datasets into two negative an positive classes for the breast cancer and applied CNN-basic and Inception-V3 (last layer training) on them. The result showed significant increasing of the algorithms performance. The CNN algorithm with basic architecture could discriminate the positive (scores 1, 2, and 3) and negative (score 0) images with 94% accuracy. Besides, applying the Inception-V3 which its last layer was trained indicated 96% accuracy for the same dataset.

3.2. Discrimination of Tumor Subtypes Across Heterogeneous Images

Tumor tissues are highly heterogeneous (Marino et al., 2015) that lead in great limitation for the correct diagnosis. Tumor heterogeneity is the result of genetic disorders which potentially reflects on a variability of morphological features (Nassar et al., 2010).

We randomly selected 1629 H&E stained high resolution image patches (i.e. a few patches of each tumor slide) from TCGA (Network et al., 2012, 2014) comprising lung adenocarcinoma and squamous cell carcinoma. Then, we trained all CNN architectures for the selected images to discriminate two subtypes. Consequently, we assessed the performance of the trained algorithms for a separated

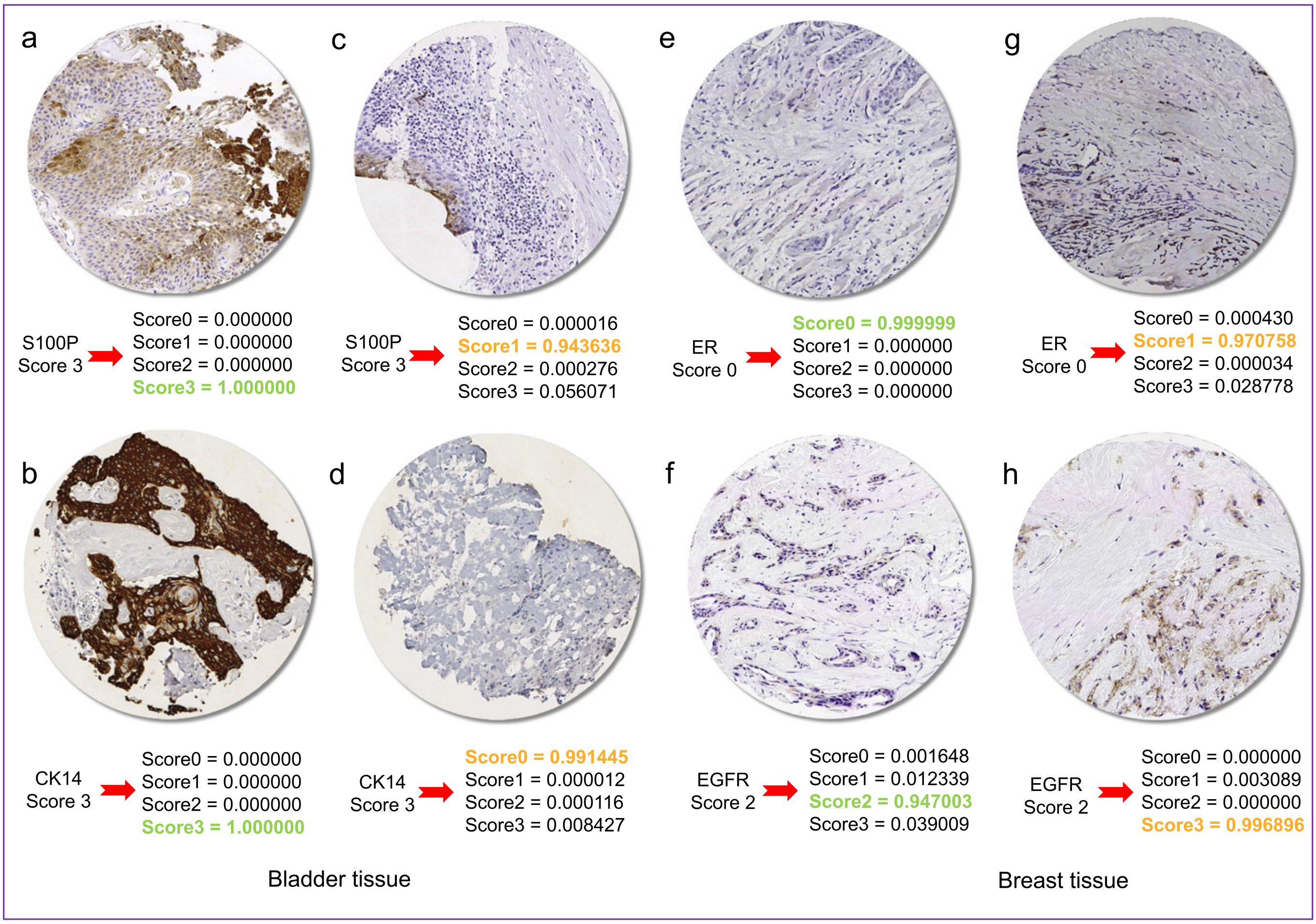


Fig. 2. Low accuracy may be associated with human errors in labeling of IHC scores. For example, figures a, b, c, and d labeled to score 3 by pathologists, while the algorithm (Inception-V1) has classified them to score 3, 3, 1, and 0, respectively. In particular, figures e and g are both labeled to score 0 by pathologists; however, the algorithm correctly has classified them into score 0 and 1, respectively. Finally, figures f and h are labeled to score 2 by pathologists while the algorithm has classified them into score 2 and 3, respectively. Closer manual inspection of images indicate the algorithm results are indeed more reliable. Highlighted probability scores in *green* and *orange* indicate concordance and discordance between algorithm classification and pathologist labeling, respectively.

test set. The test set includes 50 different high resolution image patches of the tumor slides (i.e. we considered it as the Intra-tumor test set) (Fig. 3). The result showed that while CNN-basic cannot dedicate various cell populations to each subtype, the complex architectures such as Inception-V1 and -V3 can successfully distinguish adenocarcinoma and squamous cell carcinoma across heterogeneous tissue of the tumor slides with no error (TCGA-IntraHeterogeneous dataset in Table 2).

In addition, we assess the performance of algorithms on inter-tumor heterogeneity of lung cancer. We selected 1520 whole H&E stained histopathology images from TCGA as well as 860 H&E and IHC stained images from TMA database for both lung cancer subtypes (adenocarcinoma and squamous cell carcinoma). Then, we randomly selected and extracted 100 images of each TCGA and TMA datasets separately and trained all algorithms' architectures for the remaining images. Since the test set images were selected from different patients (tumor slides) that the algorithm never trained for their whole slides or patches, we considered it as Inter-tumor test set. In this way, the algorithms should cope with wide range of cell population variance (intra each individual image and inter different images).

The result indicated 92% and 83% accuracy using networks which their all layers are fine-tuned based on Inception-V1 parameters for TMAD and TCGA test sets, respectively (Table 2). The low accuracy of Inter-tumor test set in compare to the Intra-tumor test set can be associated to the high heterogeneity that present across lung cancer for various patients. The mentioned heterogeneity may link to various growth patterns (lepidic, acinar, papillary, and solid) (Marino et al., 2015), grades, and stages in a mixed LUAD and LUSC (or cancer and normal) of the obtained images from various lung cancer patients (Fig. 3).

Based on the overall results, it could be useful to use suitable architectures of CNN algorithms based on the goal of projects. For example, we can use simpler and complex architectures of CNN for discrimination of tumor subtypes through intra- and inter-heterogeneity, respectively.

3.3. *Selecting Optimal Step Number and Training Approach of CNN Algorithm*

In order to find the optimal step number for CNN architectures over different datasets, we stop the training process when the validation accuracy converges to its maximum. We consider that stopping point as the optimal step number for the tested architecture and dataset (e.g. see Figs. 4 and 5). The final classification for images in

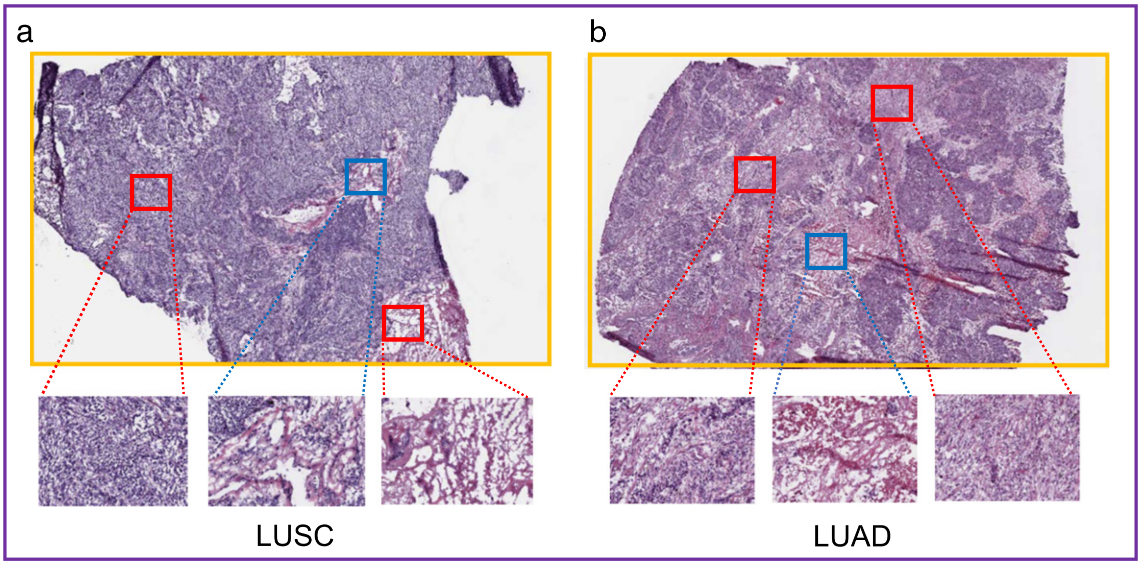


Fig. 3. Intra- and inter-tumor heterogeneity. The figure shows the squamous cell lung cancer in the left (A) and adenocarcinoma cell lung cancer in the right (B). The top images (A and B) represent whole-slide images and the down images represent the extracted high-resolution patches from TCGA datasets. The *red* cubes show the patches that algorithms are trained for them and the *blue* cubes indicate the patches comprising test set.

the test set is performed by re-training the proposed architecture over both training and validation sets with the optimal step number.

As Table 2 demonstrates, the inceptions-based architecture networks (V1 and V3) that are fine-tuned for all layers, are consistently superior. We also compare various architectures of CNN algorithm using ROC (Fig. 6) and PRC (Fig. 7) for a dataset using various thresholds. In this experiment, we consider outputs of an algorithm if prediction's confidence of a sample pass the determined threshold. We observe a trade-off between precision and recall (for PRC) and TPR and FPR (for ROC) by varying this threshold. These figures reveal that algorithms are able to classify more images through larger recall and smaller threshold.

We also compare accuracy of different strategies for training Inception-V1. In this regard, we train the model on the marker dataset of breast cancer across training the last layer, fine-tuning of the parameters for all layers, and the training of our own network from scratch (Fig. 5). As the figure shows, the best performance is obtained using a pre-trained network and fine-tuning the parameters for all layers of the network, which is in concordance with the results of previous studies (Esteva et al., 2017; Gulshan et al., 2016).

3.4. Robustness and Limitations of CNN_Smoothie

To demonstrate the robustness of the CNN_Smoothie method, we apply it to eight different datasets of histopathological images with different spectrum of apparent colors to show the uniformity of its performance. The image set spans multiple tumor types, along

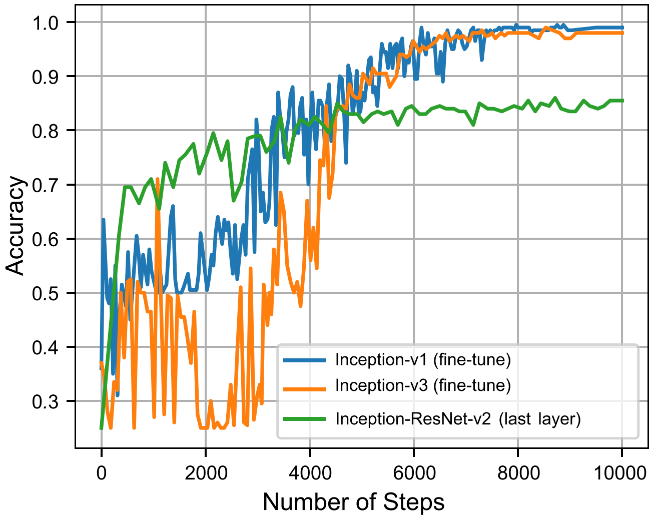


Fig. 4. The graph shows the optimal number of steps for Inception-ResNet (last layer training), Inception-V1 (fine-tuning all layers), and Inception-V3 (fine-tuning all layers) to get the highest accuracy in BladderBiomarkers.

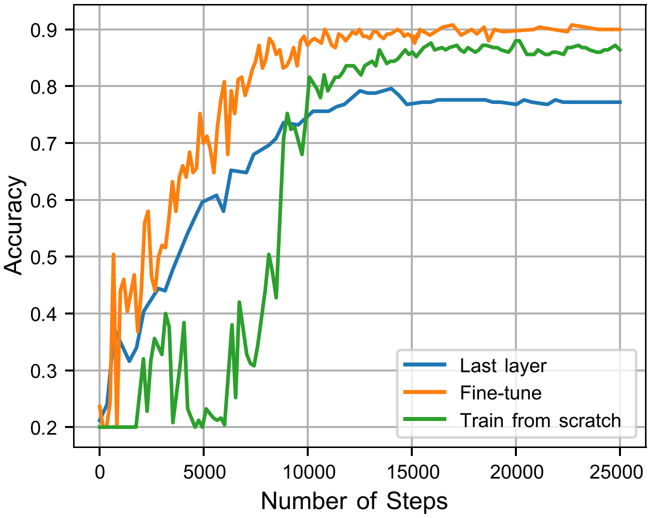


Fig. 5. Inception-V1 via three different training strategies (last layer training, fine-tuning the parameters for all layers, and training from the scratch) in BreastBiomarkers dataset.

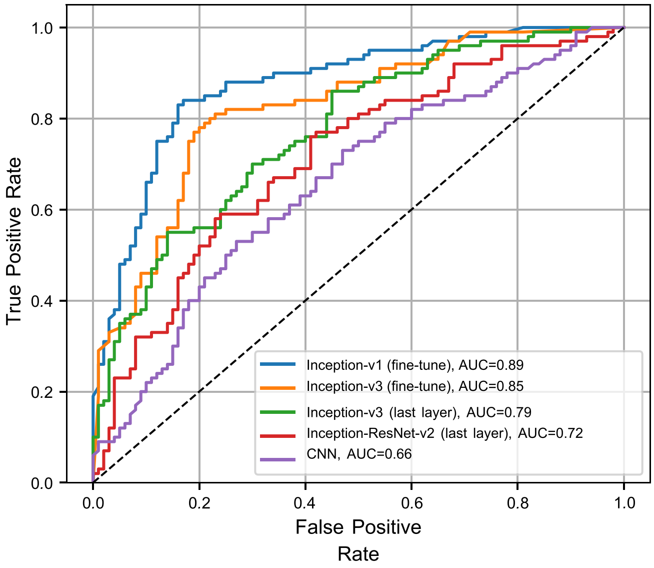


Fig. 6. Receiver operating characteristic (ROC) curve for the TCGA-InterHeterogeneity dataset.

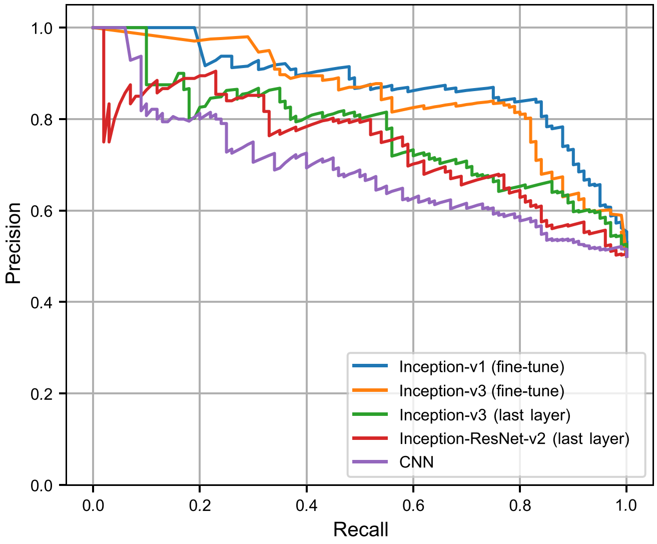


Fig. 7. Precision versus recall for the TCGA-InterHeterogeneity dataset. The 4000-step version is used for Inception-V3 (training the last layer).

with several different image colors. The results show that although the colors space for different images have different distributions, our CNN_Smoothie method can successfully identify and register tumor variations and discriminate them consistently and robustly (Fig. 8).

In addition, we evaluate the performance of algorithms using various statistical measurements on TMAD-InterHeterogeneity and TCGA-InterHeterogeneity datasets to assess the robustness of results (Table 3). These measures include AUC, average of Precision and Recall, Cohen's kappa, Jaccard Coefficient, and Log-loss. The Youden index (Youden, 1950) also referred to the ROC which is an indicator for the performance of a classifier and measured as specificity + sensitivity 1 (Table 3).

4. Discussion

The era of computational pathology is rapidly evolving and there are enormous opportunities for computational approaches to provide additional prognostic and diagnostic information that cannot be provided by pathologists alone (Bouzin et al., 2016; Louis et al., 2014; Roth and Almeida, 2015; Sarnecki et al., 2016). The CNN_Smoothie pipeline presented here provides a novel framework that can be easily implemented for a wide rang of applications, including immunohistochemistry grading and detecting tumor biomarkers. Recently several papers have been published that utilize various methods such as classical machine learning approaches including support vector machine (SVM) and random forest (RF) (Yu et al., 2016), and deep learning methods such as CNN-basic (Vandenberghe et al., 2017) or Inception methods (Esteva et al., 2017). However, this is the first report that utilize various architectures of CNN algorithms and compare their performance on histopathological tumor images across various configurations.

The aim of this project is to evaluate the utility of convolutional neural networks to automatically identify cancer tissues, subtypes, related markers, and their staining scores. We indicate deep learning approaches can provide accurate status assessments in clinical conditions. Our results show the accuracy of convolutional neural networks primarily depends on the size, complexity, algorithm architecture, and noise of the dataset utilized. We also show that our study raise several important issues regarding tumor heterogeneity since different response of deep learning could be due to genetic heterogeneity. Further studies required in order to clarify the efficiency of the deep learning application in detection of heterogeneity through digital images.

In terms of computation cost, note that we optimized our pipeline so that it can be run on CPUs. However, GPUs are indeed preferable to scale up the method to Pan-Cancer Analysis and accelerate training speed for future work.

The discordance of our findings and pathology results are due to the low number of tumor images. In certain cases, we blended some images to increase the number of images in each class. In particular, the images associated with biomarkers were blended for each score in BreastScores and BladderScores datasets. Then, the algorithms were trained for different scores disregard of the biomarkers associated with bladder and breast cancers. In addition, the number of images in some classes are not balanced which lead to compliance biases. Finally, we did not train all the algorithms from scratch because GPU is necessary for some datasets and architectures due to their higher complexity. We leave this for future work.

Our method yields cutting edge sensitivity on the challenging task of detecting various tumor classes in histopathology slides, reducing the false rate. Note that, our CNN_Smoothie pipeline requires no prior knowledge of an image color space or any parameterizations from the users. It provides pathologists or medical technicians a straightforward platform to use without requiring sophisticated computational knowledge.

Research in Context

Computational pathology approaches are complementary to other clinical evaluation methods in order to improve pathologists' knowledge of disease and to improve treatment strategies. In this paper, we develop an open source pipeline to detect various cancer tissues and subtypes with the aim of increasing accuracy of diagnosis with focus on applying deep learning algorithms. The pipeline does not require any prior knowledge of the image color space or any parameterization input from the user, which allows pathologists or medical technicians to apply this approach without extensive knowledge of optimization or mathematical tools.

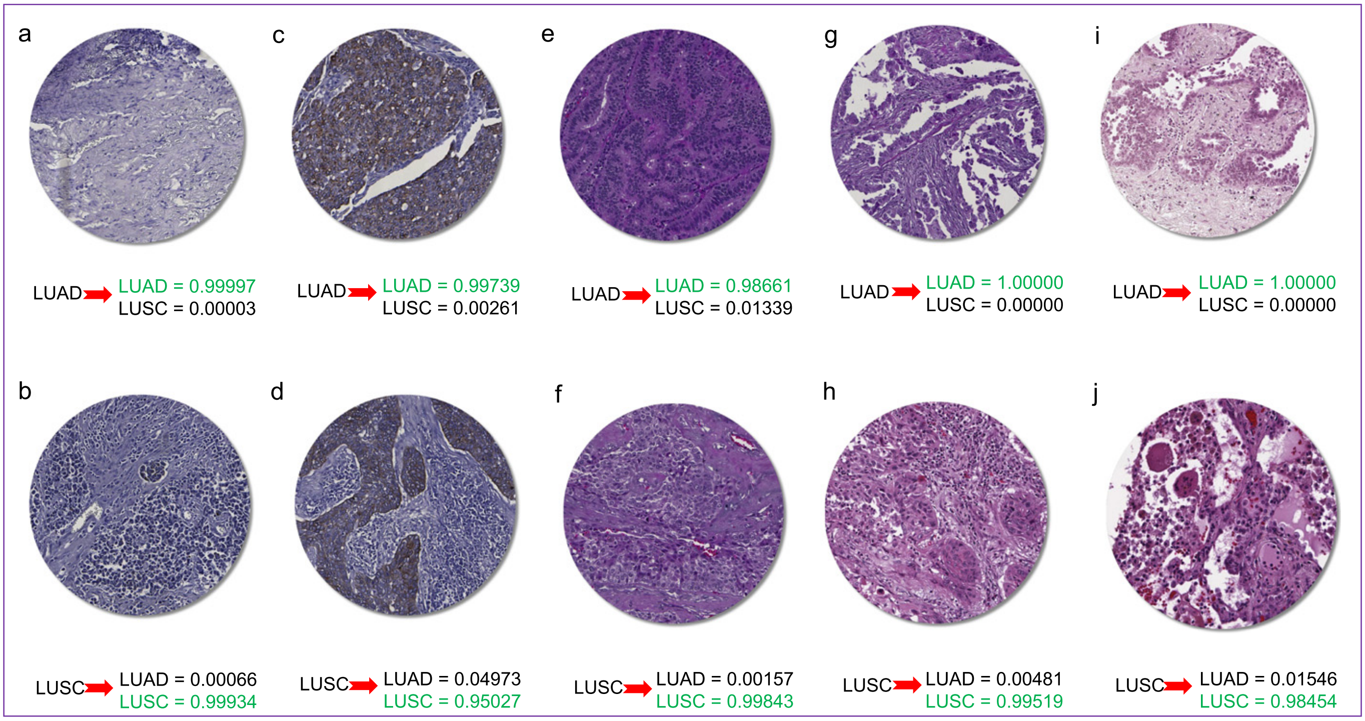


Fig. 8. CNN_Smoothie successfully identifies tumor subtypes (LUAD vs. LUSC) and discriminates them consistently and robustly across different spectrum of colors. Highlighted probability scores in *green* indicate the output of classification using Inception-V1.

Table 3
The result on TMAD-InterHeterogeneity and TCGA-InterHeterogeneity datasets using various statistics measures. The number in parentheses correspond to the Youden Index. The bold fonts indicate higher classification accuracies for the measures.

Algorithms	AUC		Precision		Recall		Cohen's kappa		Jaccard coefficient		Log-loss	
	TMA	TCGA	TMA	TCGA	TMA	TCGA	TMA	TCGA	TMA	TCGA	TMA	TCGA
CNN-basic	0.64 (0.27)	0.61 (0.22)	0.71	0.62	0.73	0.61	0.30	0.22	0.73	0.61	1.34	1.4
Inception-V3 Last-layer 4000-steps	0.79 (0.59)	0.70 (0.40)	0.81	0.71	0.80	0.70	0.56	0.4	0.80	0.70	0.45	**0.57**
Inception-V3 Last-layer 12000-steps	0.76 (0.52)	0.70 (0.40)	0.79	0.70	0.78	0.70	0.50	0.4	0.78	0.70	0.55	0.64
Inception-V1 Fine-tune	**0.89 (0.80)**	**0.83 (0.66)**	**0.92**	**0.84**	**0.92**	**0.83**	**0.81**	**0.66**	**0.92**	**0.83**	0.39	0.66
Inception-ResNet-V2 Last-layer	0.68 (0.35)	0.66 (0.32)	0.74	0.68	0.75	0.66	0.38	0.32	0.75	0.66	0.48	0.63
Inception-V3 Fine-tune	0.87 (0.75)	0.79 (0.58)	0.90	0.83	0.90	0.79	0.76	0.58	0.90	0.79	**0.36**	1.16

Acknowledgments

Ehsan Kazemi is supported by Swiss National Science Foundation under grant number 168574. This work was supported by start up funds (Weill Cornell Medicine) to Iman Hajirasouliha. We would also like to thank Mehdi Totonchi, Mehdi Habibzadeh-Motlagh and Ali Sharifi for valuable related discussions to this project.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Author Contribution

PK, EK, OE and IH conceived of the study and designed the algorithms and experiments. PK and EK developed the algorithms, wrote the codes, built the pipeline and performed analysis. OE and IH supervised all aspects of the project implementation. MI contributed to the manuscript and provided additional pathological insights. PK, EK and IH wrote the manuscript. All authors read, edited and approved the final manuscript.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. 2016. Tensorflow: A system for large-scale machine learning. OSDI. pp. 265–283.

Akram, S., Javed, M.Y., Qamar, U., Khanum, A., Hassan, A., 2015. Artificial neural network based classification of lungs nodule using hybrid features from computerized tomographic images. Appl. Math. Inf. Sci. 9, 183.

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (tss). J. Appl. Ecol. 43, 1223–1232.

Antony, J., McGuinness, K., O'Connor, N.E., Moran, K., 2016. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE., pp. 1195–1200.

Beck, A.H., Sangoi, A.R., Leung, S., Marinelli, R.J., Nielsen, T.O., Van De Vijver, M.J., West, R.B., Van De Rijn, M., Koller, D., 2011. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci. Transl. Med. 3, 108ra113–108ra113.

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., 2007. Greedy layer-wise training of deep networks. Advances in neural information processing systems. pp. 153–160.

Bhargava, R., Sinha, S., Kwak, J.T., 2011. Multimodal Microscopy for Automated Histologic Analysis of Prostate Cancer. US Patent App. 13/090,384.

Bloom, K., Harrington, D., 2004. Enhanced accuracy and reliability of her-2/neu immunohistochemical scoring using digital microscopy. Am. J. Clin. Pathol. 121, 620–630.

Bottou, Y.L.L., GO, M., 1998. K.: Efficient backprop. Neural Networks: Tricks of the Trade, Springer.

Bouzin, C., Saini, M.L., Khaing, K.-K., Ambroise, J., Marbaix, E., Grégoire, V., Bol, V., 2016. Digital pathology: elementary, rapid and reliable automated image analysis. Histopathology 68, 888–896.

Bueno-de Mesquita, J.M., Nuyten, D., Wesseling, J., van Tinteren, H., Linn, S., van De Vijver, M., 2009. The impact of inter-observer variation in pathological assessment of node-negative breast cancer on clinical risk assessment and patient selection for adjuvant systemic treatment. Ann. Oncol. 21, 40–47.

Carneiro, G., Zheng, Y., Xing, F., Yang, L., 2017. Review of deep learning methods in mammography, cardiovascular, and microscopy image analysis. Deep Learning and Convolutional Neural Networks for Medical Image Computing. Springer., pp. 11–32.

Chen, X., Xu, Y., Wong, D.W.K., Wong, T.Y., Liu, J., 2015. Glaucoma detection based on deep convolutional neural network. Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. IEEE., pp. 715–718.

Chen, Y.-J., Fan, C.-Y., Chang, K.-H., 2016. Manufacturing intelligence for reducing false alarm of defect classification by integrating similarity matching approach in cmos image sensor manufacturing. Comput. Ind. Eng. 99, 465–473.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46.

Conde, E., Angulo, B., Redondo, P., Toldos, O., García-García, E., Suárez-Gauthier, A., Rubio-Viqueira, B., Marrón, C., García-Luján, R., Sánchez-Céspedes, M., et al. 2010. The use of p63 immunohistochemistry for the identification of squamous cell carcinoma of the lung. PLoS One 5, e12209.

Dong, F., Irshad, H., Oh, E.-Y., Lerwill, M.F., Brachtel, E.F., Jones, N.C., Knoblauch, N.W., Montaser-Kouhsari, L., Johnson, N.B., Rao, L.K., et al. 2014. Computational pathology to discriminate benign from malignant intraductal proliferations of the breast. PloS One 9, e114885.

Doyle, S., Feldman, M., Tomaszewski, J., Madabhushi, A., 2012. A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. IEEE Trans. Biomed. Eng. 59, 1205–1218.

Drish, J., 2001. Obtaining calibrated probability estimates from support vector machines. Technique Report, Department of Computer Science and Engineering.University of California, San Diego, CA.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118.

Fatima, N., Cohen, C., Lawson, D., Siddiqui, M.T., 2011. Ttf-1 and napsin a double stain. Cancer Cytopathol. 119, 127–133.

Felipe De Sousa, E.M., Wang, X., Jansen, M., Fessler, E., Trinh, A., De Rooij, L.P., De Jong, J.H., De Boer, O.J., Van Leersum, R., Bijlsma, M.F., et al. 2013. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. Nat. Med. 19, 614–618.

Fogel, Y., Feder, M., 2017. On the problem of on-line learning with log-loss. Information Theory (ISIT), 2017 IEEE International Symposium on. IEEE., pp. 2995–2999.

Gavrielides, M.A., Gallas, B.D., Lenz, P., Badano, A., Hewitt, S.M., 2011. Observer variability in the interpretation of her2/neu immunohistochemical expression with unaided and computer-aided digital microscopy. Arch. Pathol. Lab. Med. 135, 233–242.

Greenspan, H., van Ginneken, B., Summers, R.M., 2016. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans. Med. Imaging 35, 1153–1159.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama 316, 2402–2410.

Gurcan, M.N., 2016. Histopathological image analysis: Path to acceptance through evaluation. Microsc. Microanal. 22, 1004–1005.

Gutman, D.A., Cobb, J., Somanna, D., Park, Y., Wang, F., Kurc, T., Saltz, J.H., Brat, D.J., Cooper, L.A., Kong, J., 2013. Cancer digital slide archive: an informatics resource to support integrated in silico analysis of tcga pathology data. J. Am. Med. Inform. Assoc. 20, 1091–1098.

Hamilton, P.W., Wang, Y., Boyd, C., James, J.A., Loughrey, M.B., Hougton, J.P., Boyle, D.P., Kelly, P., Maxwell, P., McCleary, D., et al. 2015. Automated tumor analysis for molecular profiling in lung cancer. Oncotarget 6, 27938.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology 143, 29–36.

Hardiman, K.M., Ulintz, P.J., Kuick, R., Hovelson, D.H., Gates, C.M., Bhasi, A., Grant, A.R., Liu, J., Cani, A.K., Greenson, J., et al. 2016. Intra-tumor genetic heterogeneity in rectal cancer. Lab. Investig. J. Tech. Methods Pathol. 96, 4.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778.

Higgins, J.P., Kaygusuz, G., Wang, L., Montgomery, K., Mason, V., Zhu, S.X., Marinelli, R.J., Presti, J.C., Jr, van de Rijn, M., Brooks, J.D., 2007. Placental s100 (s100p) and gata3: markers for transitional epithelium and urothelial carcinoma discovered by complementary dna microarray. Am. J. Surg. Pathol. 31, 673–680.

Hinton, G.E., Osindero, S., Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18, 1527–1554.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504–507.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning. pp. 448–456.

Janowczyk, A., Chandran, S., Singh, R., Sasaroli, D., Coukos, G., Feldman, M.D., Madabhushi, A., 2012. High-throughput biomarker segmentation on ovarian cancer tissue microarrays via hierarchical normalized cuts. IEEE Trans. Biomed. Eng. 59, 1240–1252.

Jiang, M., Zhang, S., Huang, J., Yang, L., Metaxas, D.N., 2016. Scalable histopathological image analysis via supervised hashing with multiple features. Med. Image Anal. 34, 3–12.

Kaufman, P.A., Bloom, K.J., Burris, H., Gralow, J.R., Mayer, M., Pegram, M., Rugo, H.S., Swain, S.M., Yardley, D.A., Chau, M., et al. 2014. Assessing the discordance rate between local and central her2 testing in women with locally determined her2-negative breast cancer. Cancer 120, 2657–2664.

Khayyata, S., Yun, S., Pasha, T., Jian, B., McGrath, C., Yu, G., Gupta, P., Baloch, Z., 2009. Value of p63 and ck5/6 in distinguishing squamous cell carcinoma from adenocarcinoma in lung fine-needle aspiration specimens. Diagn. Cytopathol. 37, 178–183.

Khosravi, P., Gazestani, V.H., Pirhaji, L., Law, B., Sadeghi, M., Goliaei, B., Bader, G.D., 2015. Inferring interaction type in gene regulatory networks using co-expression data. Algorithms Mol. Biol. 10, 23.

Kim, E., Corte-Real, M., Baloch, Z., 2016. A deep semantic mobile application for thyroid cytopathology. Proc. SPIE. pp. 97890A.

Korbar, B., Olofson, A.M., Miraflor, A.P., Nicka, K.M., Suriawinata, M.A., Torresani, L., Suriawinata, A.A., Hassanpour, S., 2017. Deep-learning for Classification of Colorectal Polyps on Whole-slide Images. arXiv preprint arXiv:1703.01550.

Kothari, S., Phan, J.H., Stokes, T.H., Wang, M.D., 2013. Pathology imaging informatics for quantitative analysis of whole-slide images. J. Am. Med. Inform. Assoc. 20, 1099–1108.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. pp. 1097–1105.

Lakhani, S.R., Reis-Filho, J.S., Fulford, L., Penault-Llorca, F., van der Vijver, M., Parry, S., Bishop, T., Benitez, J., Rivas, C., Bignon, Y.-J., et al. 2005. Prediction of brca1 status in patients with breast cancer using estrogen receptor and basal phenotype. Clin. Cancer Res. 11, 5175–5180.

Lam, K.-M., He, X.-J., Choi, K.-S., 2014. Using artificial neural network to predict mortality of radical cystectomy for bladder cancer. Smart Computing (SMARTCOMP), 2014 International Conference on. IEEE., pp. 201–207.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.

Lemaître, G., Martí, R., Freixenet, J., Vilanova, J.C., Walker, P.M., Meriaudeau, F., 2015. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. Comput. Biol. Med. 60, 8–31.

Lin, M., Chen, Q., Yan, S., 2013. Network in Network. arXiv preprint arXiv:1312.4400.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I., 2017. A Survey on Deep Learning in Medical Image Analysis. arXiv preprint arXiv:1702.05747.

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., Corrado, G.S., et al. 2017. Detecting Cancer Metastases on Gigapixel Pathology Images. arXiv preprint arXiv:1703.02442.

Lo, S.-C., Lou, S.-L., Lin, J.-S., Freedman, M.T., Chien, M.V., Mun, S.K., 1995. Artificial convolution neural network techniques and applications for lung nodule detection. IEEE Trans. Med. Imaging 14, 711–718.

Louis, D.N., Gerber, G.K., Baron, J.M., Bry, L., Dighe, A.S., Getz, G., Higgins, J.M., Kuo, F.C., Lane, W.J., Michaelson, J.S., et al. 2014. Computational pathology: an emerging definition. Arch. Pathol. Lab. Med. 138, 1133–1138.

Marinelli, R.J., Montgomery, K., Liu, C.L., Shah, N.H., Prapong, W., Nitzberg, M., Zachariah, Z.K., Sherlock, G.J., Natkunam, Y., West, R.B., et al. 2007. The Stanford tissue microarray database. Nucleic Acids Res. 36, D871–D877.

Marino, F.Z., Liguori, G., Aquino, G., La Mantia, E., Bosari, S., Ferrero, S., Rosso, L., Gaudioso, G., De Rosa, N., Scrima, M., et al. 2015. Intratumor heterogeneity of alk-rearrangements and homogeneity of egfr-mutations in mixed lung adenocarcinoma. PloS One 10, e0139264.

Menegola, A., Fornaciali, M., Pires, R., Avila, S., Valle, E., 2016. Towards Automated Melanoma Screening: Exploring Transfer Learning Schemes. arXiv preprint arXiv:1609.01228.

Mohanty, S.P., Hughes, D.P., Salathé, M., 2016. Using deep learning for image-based plant disease detection. Front. Plant Sci. 7,

Mosquera-Lopez, C., Agaian, S., Velez-Hoyos, A., Thompson, I., 2015. Computer-aided prostate cancer diagnosis from digitized histopathology: a review on texture-based systems. IEEE Rev. Biomed. Eng. 8, 98–113.

Movshovitz-Attias, Y., Yu, Q., Stumpe, M.C., Shet, V., Arnoud, S., Yatziv, L., 2015. Ontological supervision for fine grained classification of street view storefronts. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1693–1702.

Nassar, A., Radhakrishnan, A., Cabrero, I.A., Cotsonis, G.A., Cohen, C., 2010. Intratumoral heterogeneity of immunohistochemical marker expression in breast carcinoma: a tissue microarray-based study. Appl. Immunohistochem. Mol. Morphol. 18, 433–441.

Network, C.G.A.R., et al. 2012. Comprehensive genomic characterization of squamous cell lung cancers. Nature 489, 519.

Network, C.G.A.R., et al. 2014. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543.

Perez, E.A., Suman, V.J., Davidson, N.E., Martino, S., Kaufman, P.A., Lingle, W.L., Flynn, P.J., Ingle, J.N., Visscher, D., Jenkins, R.B., 2006. Her2 testing by local, central, and reference laboratories in specimens from the north central cancer treatment group n9831 intergroup adjuvant trial. J. Clin. Oncol. 24, 3032–3038.

Polyak, K., 2011. Heterogeneity in breast cancer. J. Clin. Invest. 121, 3786.

Prashanth, R., Deepak, K., Meher, A.K., 2017. High accuracy predictive modelling for customer churn prediction in telecom industry. International Conference on Machine Learning and Data Mining in Pattern Recognition. pp. 391–402.

Razzak, M.I., Naz, S., Zaib, A., 2017. Deep Learning for Medical Image Processing: Overview, Challenges and Future. arXiv preprint arXiv:1704.06825.

Roche, P.C., Suman, V.J., Jenkins, R.B., Davidson, N.E., Martino, S., Kaufman, P.A., Addo, F.K., Murphy, B., Ingle, J.N., Perez, E.A., 2002. Concordance between local and central laboratory her2 testing in the breast intergroup trial n9831. J. Natl. Cancer Inst. 94, 855–857.

Roth, K.A., Almeida, J.S., 2015. Coming Into Focus: Computational Pathology as the New Big Data Microscope.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. 2014. Imagenet Large Scale Visual Recognition Challenge. arXiv preprint arXiv:1409.0575.

Sarnecki, J.S., Burns, K.H., Wood, L.D., Waters, K.M., Hruban, R.H., Wirtz, D., Wu, P.-H., 2016. A robust nonlinear tissue-component discrimination method for computational pathology. Lab. Investig. 96, 450–458.

Scagliotti, G.V., Parikh, P., Von Pawel, J., Biesma, B., Vansteenkiste, J., Manegold, C., Serwatowski, P., Gatzemeier, U., Digumarti, R., Zukin, M., et al. 2008. Phase iii study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. J. Clin. Oncol. 26, 3543–3551.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 815–823.

Simard, P.Y., Steinkraus, D., Platt, J.C., et al. 2003. Best practices for convolutional neural networks applied to visual document analysis. ICDAR. pp. 958–962.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. AAAI. pp. 4278–4284.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826.

Tang, P., Wang, J., Bourne, P., 2008. Molecular classifications of breast carcinoma with similar terminology and different definitions: are they the same? Hum. Pathol. 39, 506–513.

van der Laak, J.A., Pahlplatz, M.M., Hanselaar, A.G., de Wilde, P., 2000. Hue-saturation–density (hsd) model for stain recognition in digital images from transmitted light microscopy. Cytometry A 39, 275–284.

Vandenberghe, M.E., Scott, M.L., Scorer, P.W., Söderberg, M., Balcerzak, D., Barker, C., 2017. Relevance of deep learning to facilitate the diagnosis of her2 status in breast cancer. Sci. Rep. 7,

Vincent, L., Soille, P., 1991. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans. Pattern Anal. Mach. Intell. 13, 583–598.

Vogel, C., Bloom, K., Burris, H., Gralow, J., Mayer, M., Pegram, M., Rugo, H., Swain, S., Yardley, D., Chau, M., et al. 2011. P1-07-02: Discordance Between Central and Local Laboratory her2 Testing From a Large her2-negative Population in Virgo, A Metastatic Breast Cancer Registry.

Wang, L.-W., Qu, A.-P., Yuan, J.-P., Chen, C., Sun, S.-R., Hu, M.-B., Liu, J., Li, Y., 2013. Computer-based image studies on tumor nests mathematical features of breast cancer and their clinical prognostic value. PLoS One 8, e82314.

Webster, J., Dunstan, R., 2014. Whole-slide imaging and automated image analysis: considerations and opportunities in the practice of pathology. Vet. Pathol. 51, 211–223.

Xu, Y., Jia, Z., Wang, L.-B., Ai, Y., Zhang, F., Lai, M., Eric, I., Chang, C., 2017. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinf. 18, 281.

Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3, 32–35.

Yu, K.-H., Zhang, C., Berry, G.J., Altman, R.B., Ré, C., Rubin, D.L., Snyder, M., 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat. Commun. 7,

Zadrozny, B., Elkan, C., 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. ICML. pp. 609–616.

Zawistowski, M., Sussman, J.B., Hofer, T.P., Bentley, D., Hayward, R.A., Wiitala, W.L., 2017. Corrected roc analysis for misclassified binary outcomes. Stat. Med. 36, 2148–2160.