Proceedings Paper:
Graham, S., Shaban, M., Qaiser, T. et al. (2018) Classification of lung cancer histology images using patch-level summary statistics. In: Tomaszewski, J.E. and Gurcan, M.N., (eds.) SPIE Medical Imaging 2018: Digital Pathology. SPIE Medical Imaging 2018: Digital Pathology, 10-15 Feb 2018, Houston, Texas, USA. Society of Photo-optical Instrumentation Engineers (SPIE). Article no: UNSP 1058119. ISBN: 9781510616516. ISSN: 1605-7422. EISSN: 2410-9045.

https://doi.org/10.1117/12.2293855

Classification of lung cancer histology images using patch-level summary statistics

Simon Graham, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali Khurram, et al.

Simon Graham, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali Khurram, Nasir Rajpoot, "Classification of lung cancer histology images using patch-level summary statistics," Proc. SPIE 10581, Medical Imaging 2018: Digital Pathology, 1058119 (6 March 2018); doi: 10.1117/12.2293855

SPIE.

Event: SPIE Medical Imaging, 2018, Houston, Texas, United States

Classification of Lung Cancer Histology Images using Patch-Level Summary Statistics

Simon Graham[a,b], Muhammad Shaban[a], Talha Qaiser[a], Navid Alemi Koohbanani[a], Syed Ali Khurram[c], and Nasir Rajpoot[a]

[a]Department of Computer Science, University of Warwick, United Kingdom
[b]Mathematics for Real-World Systems Centre for Doctoral Training, University of Warwick, United Kingdom
[c]School of Clinical Dentistry, University of Sheffield. United Kingdom

ABSTRACT

There are two main types of lung cancer: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), which are grouped accordingly due to similarity in behaviour and response to treatment. The main types of NSCLC are lung adenocarcinoma (LUAD), which accounts for about 40% of all lung cancers and lung squamous cell carcinoma (LUSC), which accounts for about 25-30% of all lung cancers. Due to their differences, automated classification of these two main subtypes of NSCLC is a critical step in developing a computer aided diagnostic system. We present an automated method for NSCLC classification, that consists of a two-part approach. Firstly, we implement a deep learning framework to classify input patches as LUAD, LUSC or non-diagnostic (ND). Next, we extract a collection of statistical and morphological measurements from the labeled whole-slide image (WSI) and use a random forest regression model to classify each WSI as lung adenocarcinoma or lung squamous cell carcinoma. This task is part of the Computational Precision Medicine challenge at the MICCAI 2017 conference, where we achieved the greatest classification accuracy with a score of 0.81.

Keywords: Non-small cell lung cancer, histology image classification, computational pathology, deep learning

1. INTRODUCTION

Lung cancer is the leading cause of cancer death worldwide, where almost a fifth of all cancer deaths are lung cancers.[1] Non-small cell lung cancer (NSCLC) accounts for over 80% of cases, where the two major histological types required to be diagnosed by pathologists are lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). Lung squamous cell carcinoma accounts for about 25-30% of all lung cancers and originates in the tissue that lines the air passages within the lung. In well differentiated LUSC, typical features include keratinization, often in pearl formation, and inter-cellular bridging. Lung adenocarcinoma originates in the mucous secreting glandular cells within the lung and accounts for about 40% of all lung cancers. Adenocarcinomas exhibit a high level of heterogeneity, where major subtypes include: acinar, papillary, micro-papillary, lepidic and solid tumor growth patterns. Over 80% of LUAD cases today are diagnosed as a mixed sub-type, consisting of two or more histological sub-types. Some typical examples from both lung squamous cell carcinoma and lung adenocarcinoma can be seen in Figure 1. Distinguishing between LUAD and LUSC is of clinical significance because classification can subsequently be done at scale to assist the pathologists and oncologists in making decisions about personalized treatment plans for NSCLC patients. For example, chemotherapy regimens are dependent on the type of NSCLC that is diagnosed. Despite the importance of distinguishing between NSCLC histological types, the task is non-trivial for poorly differentiated cases where typical morphological features are infrequent. Furthermore, manual inspection and analysis of WSIs to detect these types of lung cancer is a labor-intensive, subjective and time-consuming task particularly when the workload is high.

There is a growing trend towards a digitized pathology work-flow, where digital images are acquired from glass histology slides using a scanning device. The advent of digital pathology has led to a rise in computational

Corresponding author. E-mail: s.graham.1@warwick.ac.uk

Medical Imaging 2018: Digital Pathology, edited by John E. Tomaszewski, Metin N. Gurcan,
Proc. of SPIE Vol. 10581, 1058119 · © 2018 SPIE · CCC code:
1605-7422/18/$18 · doi: 10.1117/12.2293855

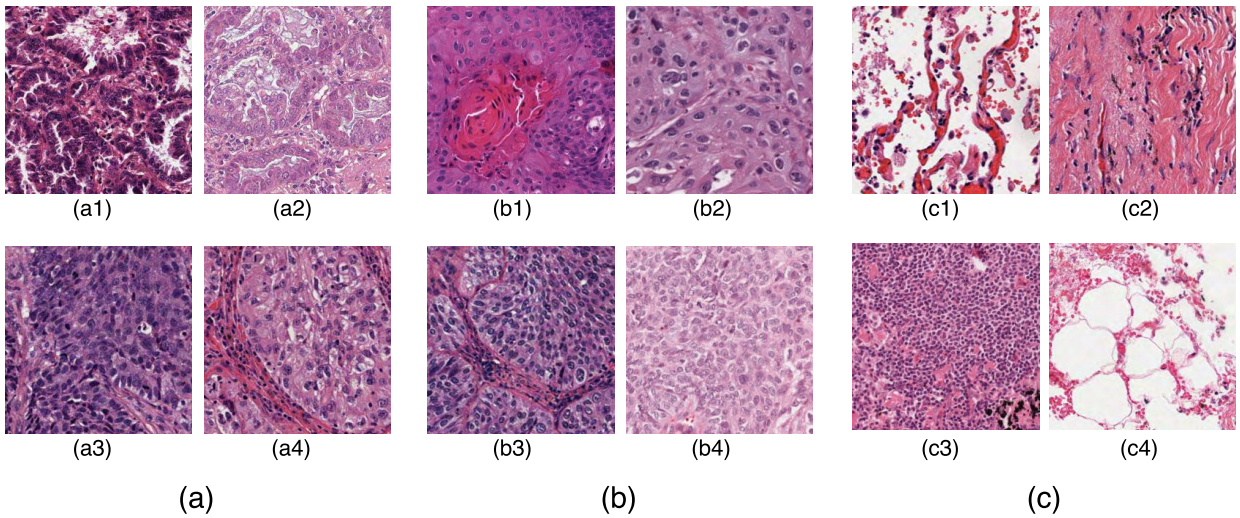

Figure 1. Examples of typical regions from each class. All patches are displayed at 256×256 at 10× resolution. Note, these patches are not used as input to the network, but are for display only. (a): LUAD patches; (b) LUSC patches; (c) ND patches. (a1) and (a2) show growth patterns typical in LUAD. (a3) and (a4) lack obvious non-solid growth patterns and are therefore harder to classify. (b1) displays a keratin pearl typical of LUSC. (b2-b4) display no keratinization. (c1) shows an alveolar region; (c2) shows a stromal region; (c3) shows a lymphocytic region; (c4) shows a fat region.

pathology, where algorithms are implemented to assist pathologists in diagnostic decision making. Whole slide images (WSIs) obtained from scanning the original glass slides can be leveraged to develop algorithms for classification tasks, where a single label is assigned to each slide. In particular, algorithms can be implemented for the purpose of classifying non-small cell lung cancer slides into either lung adenocarcinoma or lung squamous cell carcinoma. Similar to traditional methods, the automation of this task remains a challenge because typical histological features are not as obvious in poorly differentiated tumors and there is a high level of intra-class heterogeneity. Figure 1 highlights the difficulty in distinguishing between lung adenocarcinoma and lung squamous cell carcinoma when diagnostic features are uncommon. For example, (a1) and (a2) are more obviously lung adenocarcinoma than (a3) and (a4).

1.1 Related Work

In recent years, there have been a number of published methods for automated NSCLC classification. Yu *et al.*[2] extracted a range of quantitative image features from tissue regions and used an array of classical machine techniques to classify each WSI. Although hand crafted approaches perform well, there is a growing trend towards deep learning approaches, where networks are capable of learning a strong feature representation. As a result of this strong feature representation, recent deep networks[3–6] have achieved remarkable accuracy in large-scale image recognition tasks.[7] Most WSI classification methods use a patch-based approach due to the computational difficulty in processing multi-gigapixel images. Coudray *et al.*[8] classified NSCLC WSIs using deep learning on a patch-by-patch basis, but also predicted the ten most commonly mutated genes. For lung cancer classification, the authors used an inception v3 network architecture to classify input patches into LUAD, LUSC and normal. The authors assumed that all patches within each WSI had the same label and therefore did not differentiate between diagnostic and non-diagnostic regions. This method may result in a large amount false positives in non-diagnostic regions and training may take a long time to converge. Hou *et al.*[9] trained a patch-level classifier to classify glioma and NSCLC WSIs into different cancer types. This was done by aggregating discriminative patch-level predictions from a deep network using either a multi-class logistic regression model or support vector machine. The selection of discriminative patches was done in a weakly supervised manner, where an expectation-maximization approach was used to iteratively select patches. These patches were then fed into a conventional two-class CNN to classify input patches as LUAD or LUSC. The authors of this method counter the problem of differentiating diagnostic and non-diagnostic regions by only considering discriminative patches. Although successful, this technique would likely fail if presented with a small unrepresentative dataset.

1.2 Contributions

In this work, we propose a two-part automated method to address the challenge of classifying NSCLC histology images. Firstly, we classify all input patches from an unseen WSI as either LUAD, LUSC or non-diagnostic (ND) and obtain the corresponding probability maps for each class. Next, we extract a collection of statistical and morphological features from the LUAD and LUSC probability maps for input into a random forest regression model to classify each WSI. To our knowledge, this is the first 3-class network that aims to classify each WSI into diagnostic and non-diagnostic areas. This task has been organized as part of the Computational Precision Medicine challenge at the MICCAI 2017 conference, where we achieved the greatest accuracy.

1.3 Overview

The rest of the article is organized as follows: In section two we will give an overview of our framework for NSCLC classification. Section three will highlight the results of our experiments and finally, conclusions will be made in section four.

2. MATERIALS AND METHODS

In this section, we present the method for non-small cell lung cancer classification. The section is broken down into three main parts: (i) Explanation of the dataset; (ii) deep learning framework for patch based classification; (iii) Random forest regression model for classifying a whole slide image as LUAD or LUSC. An overview of the classification framework can be viewed in Figure 2.

2.1 The Dataset

As part of the Computational Precision Medicine (CPM) challenge[11] at the MICCAI 2017 conference, we used a total of 64 Hematoxylin and Eosin (H&E) NSCLC WSIs that were split into 32 training and 32 test images. Ground truth was supplied for the training images that gave the cancer type of each WSI, whereas this ground truth was held back by the challenge organizers for the test images. We had an even breakdown of NSCLC images in both the training and the test set, giving a total of 32 LUAD slides and 32 LUSC slides. We divided our dataset so that we had 24 WSIs for training and 8 for validation, with 4 validation images taken from LUAD and LUSC respectively. We extracted a 3 class dataset comprising of patches of size 256×256 at 20× magnification, from non-exhaustive labeled regions, confirmed by an expert pathologist (AK). This 3 class dataset consisted of LUAD, LUSC and non-diagnostic areas (ND). LUAD diagnostic regions within the slide consisted of: tumor; growth pattern structures and tumor stroma. LUSC diagnostic regions consisted of: tumor; keratin pearls and tumor stroma. Non-diagnostic regions included: fat; lymphocytes; blood vessels; alveoli; red blood cells; normal stroma; cartilage and necrosis. We considered necrosis to be non-diagnostic because, despite LUSC generally having more necrotic areas than LUAD, it is not indicative of lung squamous cell carcinoma on a patch-by-patch basis. Overall, our network is optimized on 65,788 training image patches.

Despite all slides being stained with H&E, there was a high level of stain variability from image to image. As a result, we applied Reinhard[10] stain normalization to all images to limit the adverse effect of the variation in the data. During training we performed random crop, flip and rotation data augmentation to make the network invariant to these transformations. After performing a random crop to all input patches, we were left with a patch size of 224×224.

2.2 Deep Neural Network for Patch-Based Classification

2.2.1 Convolutional Neural Network

An increase in the amount of labeled data coupled with a surge in computing power has allowed deep convolutional neural networks to achieve state-of-the-art performance in computer vision tasks. The hierarchical architecture of such networks allow them to have a strong representational power, where the complexity of learned features increases with the depth of the network. The proposed network f is a composition of a sequence of L functions of layers $(f_1, ..., f_L)$ that maps an input vector $\mathbf{x}$ to an output vector $\mathbf{y}$, i.e,

$$\boldsymbol{y} = f(\boldsymbol{x}; \boldsymbol{w}_1, ...\boldsymbol{w}_L) = f_L(\cdot; \boldsymbol{w}_L) \circ f_{L-1}(\cdot; \boldsymbol{w}_{L-1}) \circ ...f_2(\cdot; \boldsymbol{w}_2) \circ f_1(\cdot; \boldsymbol{w}_1) \tag{1}$$

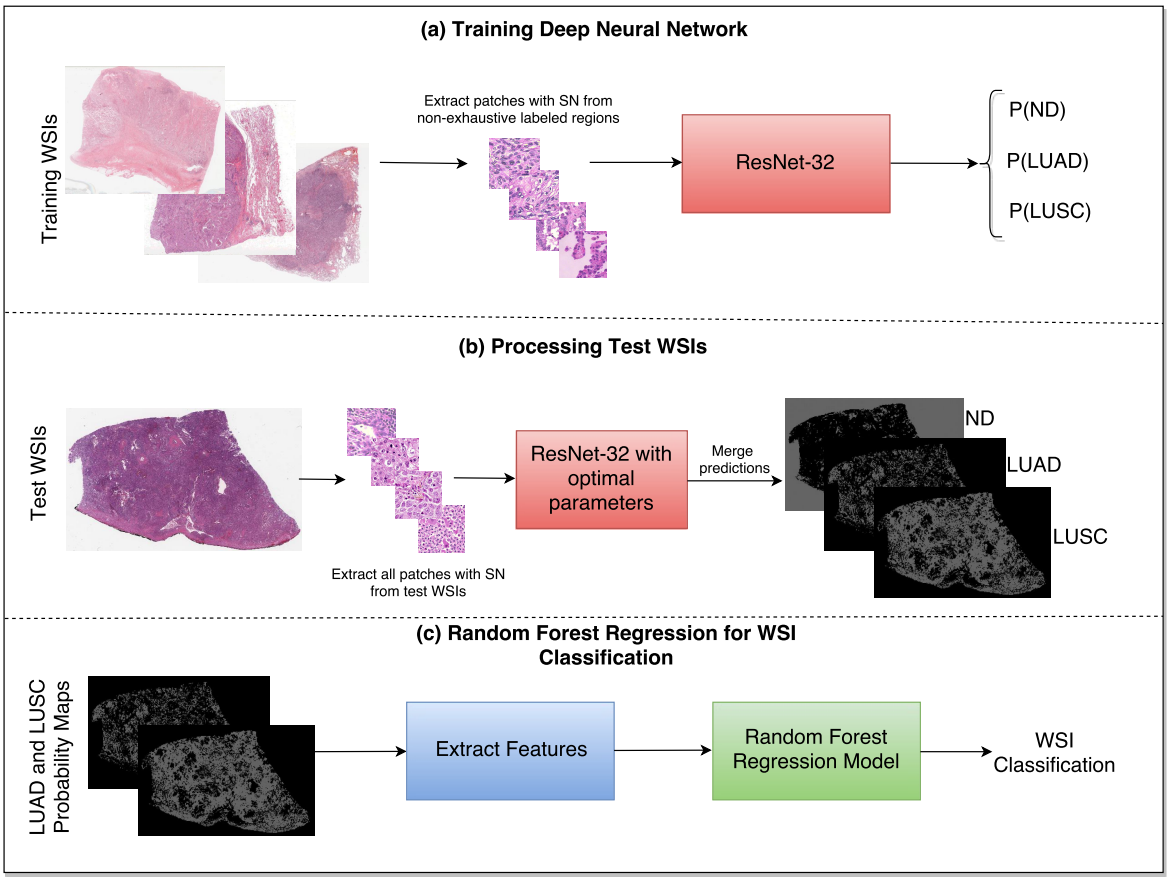

Figure 2. Overview of the NSCLC classification framework. (a) Work-flow for training the neural network to classify input patches as either non-diagnostic (ND), lung adenocarcinoma (LUAD) or lung squamous cell carcinoma (LUSC). (b) Work-flow for processing WSIs within the test set to obtain probability maps for each class. (c) Work-flow for the random forest regression model. Features are extracted from LUAD and LUSC probability maps and then fed as input into the random forest model. SN stands for stain normalization via method of Reinhard.[10]

where $\mathbf{w}_l$ is the weight and bias vector for the l^{th} layer f_l. In practice, f_l most commonly performs one of the following operations: a) convolution with a set of filters; b) spatial pooling; and c) non-linear activation.

Given a set of training data $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N}$, we can estimate the vectors $\mathbf{w}_1, .., \mathbf{w}_L$ by solving:

$$\underset{\boldsymbol{w}_1 \ldots \boldsymbol{w}_L}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \ell(f(\boldsymbol{x}^{(i)}; \boldsymbol{w}_1, \ldots \boldsymbol{w}_L), \boldsymbol{y}^{(i)}), \tag{2}$$

where l is the defined loss function. We perform numerical optimization of (2) conventionally via the back-propagation algorithm and stochastic gradient descent methods.

In addition to the above operations, residual networks (ResNets)[3] have recently been proposed that enable networks to be trained deeper and as a result, benefit from a greater accuracy. Current-state-of-the-art networks[3–6] indicate that network depth is of crucial importance, yet within conventional CNNs, accuracy gets saturated and then degrades rapidly as the depth becomes significantly large. The intuition behind a residual network is that it is easier to optimize the residual mapping than to optimize the original unreferenced mapping. Residual blocks are the core components ResNets and consist of a feed-forward skip connection, that performs identity mapping, without adding any extra parameters. These connections propagate the gradient throughout the model, which in turn enables the network to be trained deeper, often achieving greater accuracy.

2.2.2 Proposed network architecture

Inspired by the success of ResNet[3] in image-recognition tasks,[7] we implemented a deep neural network with residual blocks at its core to classify NSCLC input patches. This network architecture is a variant of ResNet50, as described by He *et al.*,[3] but we use a 3×3 kernel as opposed to a 7×7 kernel during the first convolution and reduce the number of parameters throughout the network. Using a 3×3 kernel is important in this domain because a smaller receptive field is needed to locate small features that are common in histology images. Reducing the amount of parameters allows the network to be more generalized and reduces the possibility of over-fitting. In order to reduce the amount of parameters, we modified ResNet50[3] by reducing the amount of residual blocks throughout the network so that we had 32 layers as opposed to 50. Due to the high variability between images, and therefore between the training and validation set, consideration for preventing over-fitting is crucial. Figure 3 gives an overview of the network architecture.

Once training was complete, we selected the optimal epoch corresponding to the greatest average validation accuracy and processed patches from each test WSI. This resulted in three probability maps; one for each class.

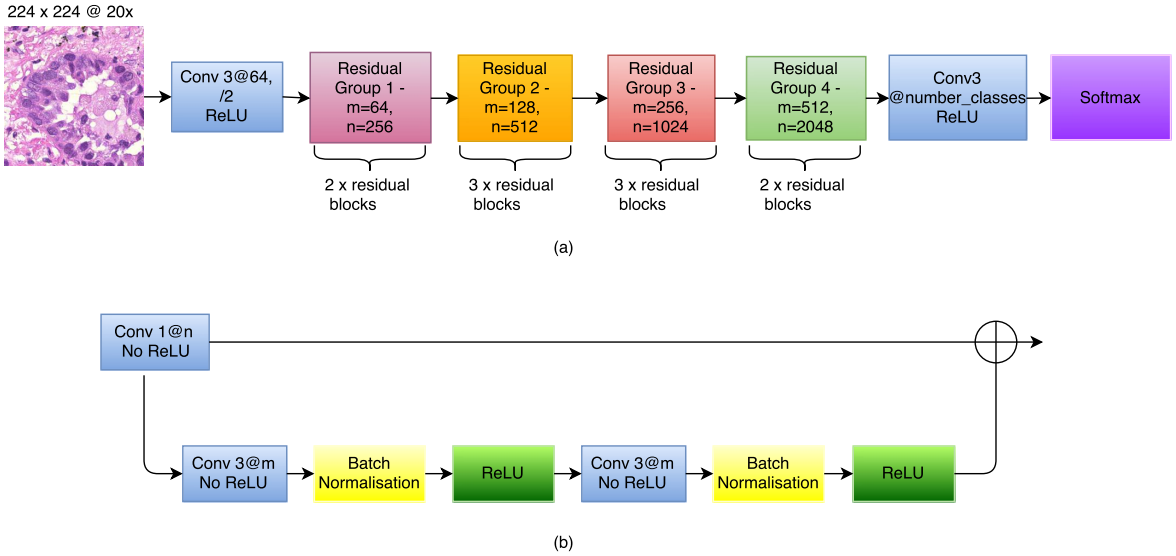

Figure 3. The proposed deep convolutional neural network. (a) Network architecture, (b) residual block. Within the residual block, ⊕ refers to a summation operator.

2.3 Extraction of Statistical and Morphological Features

For classifying each WSI as either lung adenocarcinoma or lung squamous cell carcinoma, we extracted features from both the LUAD and LUSC probability maps. We explored two post processing techniques: max voting and a random forest regression model. Max voting simply assigns the class of the WSI to be class with the largest amount of positive patches in its corresponding probability map. Therefore, max voting only requires the positive patch count for both the LUAD and LUSC probability maps in order to make a classification. For the random forest regression model, we extracted 50 statistical and morphological features from both the LUAD and LUSC training probability maps and then selected the top 25 features based on class separability. We gained the training probability maps by processing each training WSI with a late epoch. This ensured that the network had over-fit to the training data and gave a good segmentation of LUAD and LUSC diagnostic regions. In other words, using this method allowed us to transition from a non-exhaustive to an exhaustive labeled probability map. Once the model was trained with these features, they were then input as features into the random forest regression model. Statistical features that were extracted included: mean, median and variance of the probability maps. We also calculated the ratio between the LUAD and LUSC probability maps. Morphological features that were extracted included the size of the top five connected components at different thresholds.

2.4 Random Forest Regression Model

An ensemble method is a collection of classifiers that are combined together to give improved results. An example of such an ensemble method is a random forest, where multiple decision trees are combined to yield a greater classification accuracy. Decision trees continuously split the input data, according to a certain parameter until a criterion is met. Specifically, a random forest regression model fits a number of decision trees on various sub-samples of the data and then calculates the mean output of all decision trees. We optimized our random forest model by selecting an ensemble of 10 bagged trees, randomly selecting one third of variables for each decision split and setting the minimum leaf size as 5. We finally selected a threshold value to convert the output of the random forest regression model into a binary value, indicating whether the WSI was LUAD or LUSC.

3. RESULTS AND DISCUSSION

Table 1. Patch-Level accuracy. LUAD refers to lung adenocarcinoma, LUSC refers to lung squamous cell carcinoma, ND refers to area a non-diagnostic area of interest.

Network	Resolution	LUAD	LUSC	ND	Average
VGG	20x	0.634	0.663	0.826	0.708
GoogLeNet	20x	0.623	0.733	0.924	0.760
ResNet50	20x	0.601	0.597	0.889	0.695
ResNet32	20x	0.702	0.849	0.742	0.764

Table 1. summarizes the the experiments we carried out for classification of input patches into LUAD, LUSC and ND. We chose to train with the specified networks because of their state-of-the-art performance in recent image recognition tasks.[7] During training, we could see that our networks were over-fitting. This was because of two reasons: (i) The networks architectures that were used have been optimized for large-scale computer vision tasks with millions of images and thousands of classes; (ii) there is a large variability between the training set and the validation set. With such a small and visually diverse dataset, (ii) is unavoidable, and therefore we modify the network architecture to counter the problem of over-fitting. Modification of ResNet50 to give ResNet32 helped alleviate the problem of over-fitting and gave the best patch-level performance. Despite only achieving 0.4% greater accuracy than GoogLeNet, ResNet32 resulted in a significantly greater average LUAD and LUSC patch-level accuracy. The average LUAD and LUSC patch-level accuracy for GoogLeNet was 0.678, whereas the average accuracy for ResNet32 was 0.776. As a result, we chose to use ResNet32 for processing images in the test set. Figure 4 shows four test WSIs with their overlaid probability maps. Green regions show regions classified as LUSC, blue/purple regions show regions classified as LUAD and yellow/orange regions show regions classified as ND.

Table 2. Overall WSI classification accuracy. ResNet32-MV refers to classifying input patches using ResNet32, then using max voting as a post processing classification technique. ResNet32-RF refers to classifying input patches using ResNet32, then using a random forest regression model as a post processing classification technique.

Method	Accuracy
ResNet32-MV	0.78
ResNet32-RF	0.81

Table 2 shows the overall accuracy for NSCLC WSI classification, as processed by the challenge organizers. We observe that using the random forest regression model with statistical and morphological features from the labeled WSI increases the classification accuracy. Max voting is sufficient when either LUAD or LUSC is a dominant class within the labeled WSI, but when there is no obvious dominant class, the random forest regression model increases performance. This is because the features used as input to the random forest model are tailored to the task of NSCLC classification and can therefore better differentiate between each cancer type.

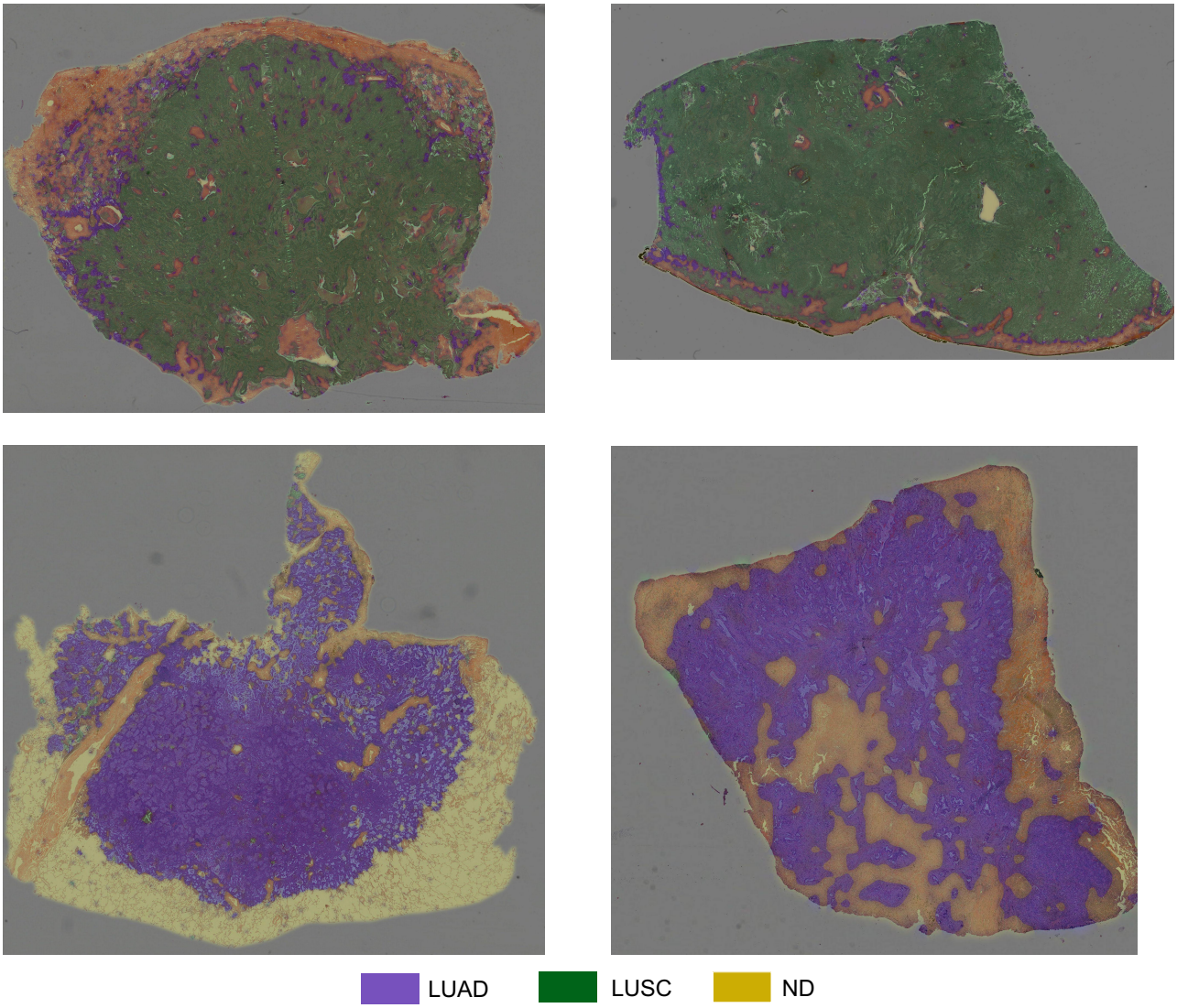

Figure 4. Test WSIs with overlaid probability maps. Blue/purple indicates a region classified as diagnostic LUAD, green indicates a region classified as diagnostic LUSC and yellow/orange indicates a region classified as ND.

4. CONCLUSIONS AND FUTURE WORK

This paper presented an automated method for non-small cell lung cancer classification into lung adenocarcinoma and lung squamous cell carcinoma. In the first step of our classification framework, we implement a deep neural network to classify input patches as lung adenocarcinoma, lung squamous cell carcinoma or non-diagnostic regions. In the second step, after processing each image, we extract a collection of statistical and morphological features from the LUAD and LUSC probability maps. These features are then used as input into a random forest regression model to classify each WSI as lung adenocarcinoma or lung squamous cell carcinoma. Overall we achieve the greatest accuracy with a score of 0.81 as part of the Computational Precision Medicine challenge at MICCAI 2017. Especially given the limitation of the dataset, classifying NSCLC WSIs into diagnostic and non-diagnostic regions seems to be of crucial importance.

The consideration of contextual information can provide additional assistance in classification tasks within computational pathology.[12,13] For example, growth patterns in LUAD cases and how the tumor grows with the stroma is of significant importance when classifying NSCLC cases. These patterns are often very hard to visualize

in a 224×224 patch at 20× resolution. In future work, developing our proposed network to accurately include more contextual information may improve patch-level accuracy and therefore overall classification accuracy.

4.1 Acknowledgments

We are grateful to the organizers of the Computational Precision Medicine WSI classification challenge, held in conjunction with MICCAI 2017.

REFERENCES

[1] "Lung cancer statistics, cancer research uk." http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer. Accessed: 2017-12-15.

[2] Yu, K.-H., Zhang, C., Berry, G. J., Altman, R. B., R, C., Rubin, D. L., and Snyder, M., "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nature Communications* **7**, 12474 (2016).

[3] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385* (2015).

[4] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., "Going deeper with convolutions," in [*Computer Vision and Pattern Recognition (CVPR)*], (2015).

[5] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *CoRR* **abs/1409.1556** (2014).

[6] Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q., "Densely connected convolutional networks," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], (2017).

[7] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "ImageNet: A Large-Scale Hierarchical Image Database," in [*CVPR09*], (2009).

[8] Coudray, N., Moreira, A. L., Sakellaropoulos, T., Fenyo, D., Razavian, N., and Tsirigos, A., "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," *bioRxiv* (2017).

[9] Hou, L., Samaras, D., Kurc, T. M., Gao, Y., Davis, J. E., and Saltz, J. H., "Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification," *ArXiv e-prints* (Apr. 2015).

[10] Reinhard, E., Ashikhmin, M., Gooch, B., and Shirley, P., "Color transfer between images," *IEEE Comput. Graph. Appl.* **21**, 34–41 (Sept. 2001).

[11] "Computational Precision Medicine challenge website." http://miccai.cloudapp.net/competitions/56. Accessed: 2017-12-15.

[12] Bejnordi, B. E., Zuidhof, G. C. A., Balkenhol, M., Hermsen, M., Bult, P., van Ginneken, B., Karssemeijer, N., Litjens, G. J. S., and van der Laak, J., "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images," *CoRR* **abs/1705.03678** (2017).

[13] Agarwalla, A., Shaban, M., and Rajpoot, N. M., "Representation-aggregation networks for segmentation of multi-gigapixel histology images," *CoRR* **abs/1707.08814** (2017).