CS231n Project: Prediction of Head and Neck Cancer Submolecular Types from Patholoy Images

Kuy Hun Koh Yoo[*]
Energy Resources Engineering
Stanford University
Stanford, CA 94305
kohykh@stanford.edu

Muhammad M. Almajid[*]
Energy Resources Engineering
Stanford University
Stanford, CA 94305
majimm0a@stanford.edu

Markus Zechner[*]
Energy Resources Engineering
Stanford University
Stanford, CA 94305
mzechner@stanford.edu

Abstract

Currently no individual treatment for head and neck cancer is available for its different subgroups. Recent research indicates possible five molecular subgroups of head and neck cancer based on DNA data which could lead to individual treatment plans in the future. Because DNA tests can be costly, this report aims to classify the molecular subgroups from histological images.

In order to handle the large dimensions of the histological section several preprocessing steps are introduced to tile the section. A pre-trained VGG16 and a self-designed convolutional neural network are used to predict subgroups on a tile level. First experiments showed that the number of 520 available histological sections has to be reduced to 50 to obtain feasible run times on the available computational resources. Very low validation accuracy ($\sim$ 20%) was obtained when the tiles were classified into five classes. This fact is a combination of a three reasons. First, there is currently no definite evidence that a histological section contains information about the molecular subgroups. Second, a subgroup classification is only provided on a histological section but not on a tile level. Because there is a possibility that the information about the subgroup is localized in the histological section, training the neural network is quite challenging. Third, there exists uncertainty in the number of subgroups. In order to address one of these challenges, we reduced the classification task to a two-class problem indicating HPV+ and HPV- patients. Validation accuracies of 85% were achieved. Generated heatmaps indicate that the entire histological section should be used in training as well as that the subgroup information might be present at several locations in the histological section. Future work has to focus on localization of subgroup information in the histological section and, therefore, select the correct tiles that classify a whole slide image.

1. Introduction

Across all cancer types, an early diagnosis can significantly increase the survival rate of patients. About $65,000$ people are diagnosed with head and neck cancer every year. That represents approximately 3% of the total number of cancer cases in the US [3]. Currently, no individual treatment procedure is present for head and neck cancer patients. Recent research indicates that head and neck cancer can be possibly divided into five subgroups based on their DNA methylation [2]. Because alternations of the DNA methylation have been recognized as an important component of cancer development, this data might be crucial for treatment selection [8]. While most patients undergo a biopsy and therefore have a histological image, DNA testing is only sporadically performed. It would be very practical if the molecular subtypes could be classified from the histological image since it would lessen the need to acquire the lengthy and expensive DNA tests. One of the biggest challenges for histology sections is whether or not the information of the molecular subgroups are present in the tissue image. An additional challenge is the fact that diagnosis accuracy depends on optical variations of colors and textures in the tissue image due to non-standardized laboratory and experi-

[*] All authors contributed equally to this work

mental protocols. Other factors are the biological heterogeneities such as cell type and state.

In this project, convolutional neural networks (CNNs) are used to build a classifier that categorizes whole slide images (WSIs) into the possible five molecular subgroups as proposed by Brennan et al.[2]. The goal is to achieve a high accuracy on classifying WSIs. Others have used CNNs to analyze and classify pathology images for other types of cancer with success [6]. This exercise is also an opportunity to validate the proposed molecular subgroups.

2. Background Related Work

There are several studies that have applied deep learning methods to pathology images. Unfortunately, an extensive literature search shows that there is a lack of studies with regards to head and neck cancer as opposed to other types of cancer such as lung or breast cancer.

Wang *et al.* [7] developed a deep learning framework to detect and localize metastasis using the Camelyon16 dataset [1]. The dataset contains WSIs with annotations indicating where the tumor is located. Wang *et al.* extracted random positive and negative patches to train their model and then produced a heatmap that shows where the tumor is likely to be. They achieved a patch classification accuracy of 98.4%. Similarly, Liu *et al.* [6] recently used an inception-inspired CNN to perform metastasis detection and localization for breast cancer using the same dataset as Wang *et al.* . They were able to detect 92.4% of the tumors at only 8 false positives per image.

Hou *et al.* [5] used a CNN-based method to perform subtype classification of glioma and non-small-cell lung carrcinoma (NSCLC) cases. Their approach consists of two levels. In the first level, an Expectation Maximization (EM) based method is combined with CNN to output the patch-level predictions iteratively. Particularly, this level distinguishes between discriminative and non-discriminative patches (i.e. the label of the patch is the same as the true label of the image or not). In the second level, histograms of patch-level predictions are fed into a multiclass logistic regression or Support Vector Machine (SVM) model that predicts the image-level labels. Their best accuracies were 77.1% and 79.8% for the glioma and NSCLC, respectively. It should be noted that this approach is problematic in that the EM algorithm depends on the initial guess. Ideally, the best approach would have a pathologist look at the training set WSIs and annotate where the cancer is likely to be present.

3. Dataset and Data Preparation

The pathology images were obtained from the Genomic Data Commons (GDC) data portal [4].The dataset consists of 520 cases that have been uploaded in the data portal for Head and Neck Squamous Cell Carcinoma under the TCGA program (TCGA-HNSC). The WSIs are stored in a muti-resolution pyramid structure as shown in Fig. 1. The labels of the WSIs refer to the five possible molecular subtypes of HNSC. These labels were received from the Gevaert lab in Medicine School of Stanford University [2]. Specifically, Brennan *et al.* performed unsupervised clustering of the 520 patients into the five subtypes based on their profiles of epigenetically deregulated genes. The five subtypes include one HPV+ subtype, two smoking-related subtypes, and two atypical subtypes. It should be noted that the HPV+ subtype is distinctively different than the other four subtypes based on the study of Brennan *et al.* . In this study the HPV+ subtype is referred to as label 4 (when 5 classes are predicted) or label 1\HPV+ (when 2 classes are predicted).

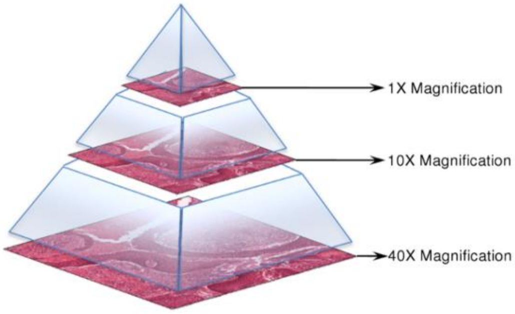


Figure 1: A schematic showing the multi-resolution pyramid structure the WSIs are generally stored in.

The WSIs were preprocessed to obtain the RGB images at a specified level. The highest level of resolution was chosen so that the WSI could be cropped into smaller images while maintaining important features of the image as shown in Fig. 2. One would expect that label information is most likely to be present in high resolution features such as deformation of cells. The WSIs were cropped into images of size $256 \times 256 \times 3$, Fig. 2A. Each cropped image was binarized using Otsu's method with a threshold of 0.8, Fig. 2B. The binarized image was then accepted as a valid image for our dataset if it had more than 70% white pixels ensuring that we have enough features in the image for our neural network to learn on. More than 2 million tiles were produced using this procedure. The only thing that is left to do is to assign labels to these individual tiles. We make the assumption that all tiles in a WSI have the same label as the WSI (i.e. if the WSI label is 1, all its tiles will have label 1). As we learned later, this assumption is problematic. The ”Experiment” section discusses this in more details. After assigning the labels to the individual patches, the tiles and their labels are then ready to be taken as input to a neural network for training, validation, and testing.

4. Approach

Our initial approach was to test different architectures and perform hyper-parameter tuning with cross-validation in order to achieve high performance on classification of the molecular subgroups.

Given the large number of tiles generated by the preprocessing step (2.2 million tiles with size $256 \times 256 \times 3$), the limited computational resources and time constraints, we had to reduce the number of tiles used to train our convolutional neural network. In particular, we initially chose to work with 2.5% of the area of each WSI and to test two CNN architectures in order to achieve run times of approximately one day. The first architecture we used is the VGG-16 architecture, shown in Fig.4, and the second one is an architecture developed by ourselves, shown in Fig. 3, which we refer to as "our model". The various models and parameters used for tuning are listed in Table 1. Since the features of a subgroup might only be seen in a very specific area of the histological section and therefore could be easily missed when using only 2.5% of the total image, we decided to reduce the number of patients to 50 but then use the entire histology section. We selected only 50 patients to keep run times below one day. Going ahead with these settings, we encountered limited validation accuracy (detailed explanation can be found in the "Experiment" section). In the next step, after consulting medical experts, we decided to reduce the classes to the most distinct categories (HPV+ and HPV-).

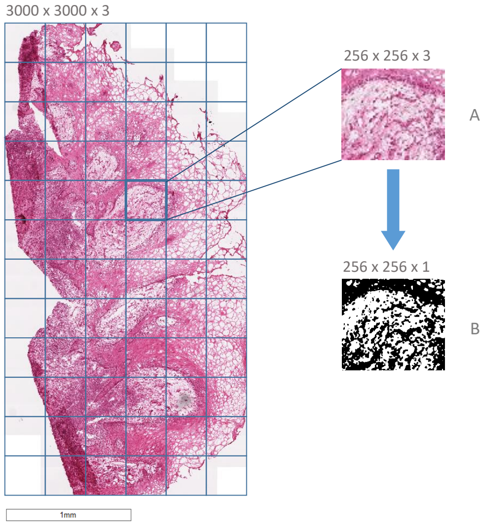


Figure 2: A super-resolved whole slide image shown on the left. A) cropped image, and B) binarized image.

5. Experiment

Our initial exercises with 2.5% of the slide (aerially) were quickly deemed incorrect after speaking to medical specialist and understanding that cancer may be localized in specific places. Taking such a small number of tiles in a random manner could completely ignore the places where the cancer was actually present. Furthermore, run times were not practical given the time constraints of this project for both the VGG-16 and our model.

Our subsequent approach was to consider a smaller number of patients but use the 50% or 100% of the WSI to address the localization issue. A summary of our exercises is shown in Table 1. We performed exercises considering both all five subtypes proposed by Brennan at al. and only two subtypes (HPV+ subtype vs. rest). The latter was suggested by the authors of the study since they noted that these were distinctively different. Results are summarized in Table 2. It is clear that best results are achieved when only two subtypes are considered. Figure 7 shows training and validation accuracies over all training epochs for our best case where we were able to achieve an 85% validation accuracy. Because the dataset became imbalanced when we combined all labels other than label 4 into one label, we compute the F1 score to ensure that our high accuracies are not just an artifact of using a highly imbalanced dataset. Therefore, Fig. 8 shows the F1 score over all epochs.

Figures 6 and Figure 9 show confusion matrices for our best models considering five or two subtypes. Both models are better at recognizing HPV+ as expected by the medical specialists, indicating a true distinction in this subtype.

In order to further investigate the reason for the initially low validation accuracy we decided to develop heatmaps of the best validation results. In a heatmap we superimpose the tiles back onto the original histological sections and color them by their predicted subgroup class.

One of the biggest assumption that we have made throughout our experimentation is that all the tiles that were extracted from a WSI will have the same label as the WSI's label. This, in reality, might not be correct because indications of subgroups in the WSI are most likely localized at one or several areas of the WSI. Most of the generated heatmaps propose that the indications are distributed over the WSI since no distinct accumulation of certain class-tiles was observed, see Fig. 10. This uniform distribution of true label tiles (label 1 in Fig. 10) makes it a difficult task to predict the true label of the WSI. A majority vote for the WSI cannot be applied since the indications might be localized. Another idea was to apply a support vector machine to identify accumulations of true label tiles in order to identify the true label for the WSI but because the true labels are uniformly distributed that will not work either. Hence, prediction is very low in the case of five classes because there is no distinction between the patches that discrimina-

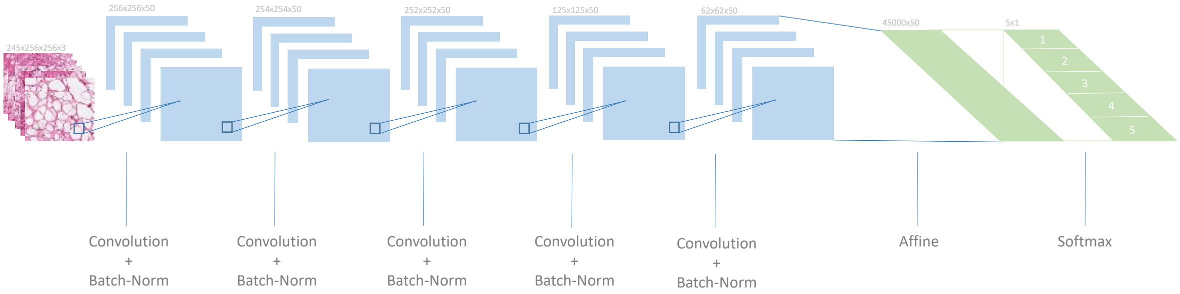


Figure 3: Convolutional neural network used (Our Model).

#	Model	Learning Rate	Area %	Num. of Patients	Num. of Classes
1	VGG-16	$10^{-3}, 10^{-2}$	2.5	520	5
2	Our model				
3	VGG-16	10^{-3}	$50, 100$	50	5
4	Our model				
5	Our model				2

Table 1: The models used for hyperparameter tuning and area selection.

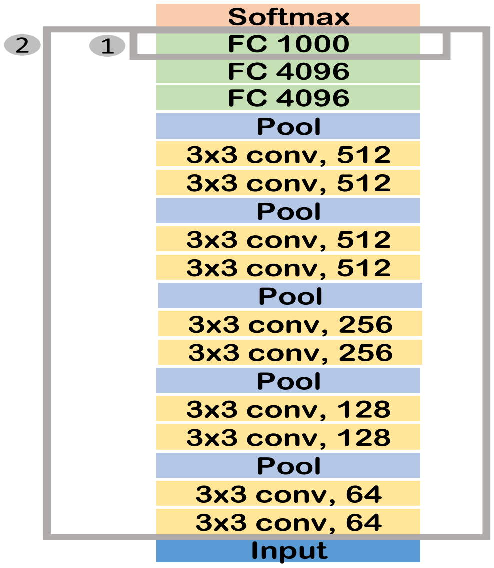


Figure 4: VGG-16 architecture. 1) the last affine layer that was trained when using pretrained model weights, and 2) the full model is trained on the dataset.

tively classify the WSI to have the correct label and those that do not really matter to this classification task. In other words, because the training dataset did not have annotations indicating where the cancer is present in each WSI and our algorithm did not explicitly determine those regions of each

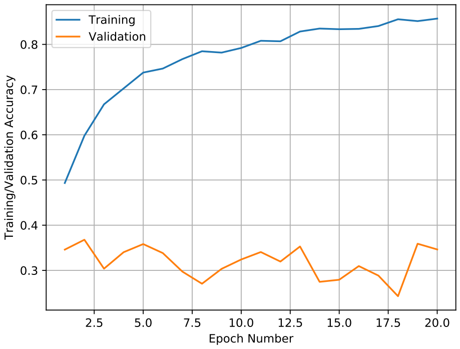


Figure 5: Training and validation accuracies when using 50 patients, 5 classes with 50% of the tiles.

WSI before predicting the labels of all the patches, the accuracy for predicting the five subtypes is expected to be very low (as we experienced in Fig. 5). We are proposing to use these heatmaps as a starting point for the pathologist to identify critical regions that contain information about the subgroups. The localization of the true label by a pathologist is essential to make a classification with convolutional neural networks possible. In a further experiment we investigated the effect of areal coverage on prediction. In one case we only randomly select 50% (Fig. 10 a and c) of the WSI's tiles. Alternatively, the second case uses all tiles obtained from a WSI (Fig. 10 b and d). The fact that the 100% case does not cover the entire WSI results from the chosen

Model	Area %	Num. of classes	Best training accuracy	Best val accuracy	Best F1 score
Our model	50	2	0.96	0.85	0.82
Our model	100		0.90	0.76	0.78
Our model	50	5	0.86	0.38	0.38
Our model	100		0.56	0.19	0.19
VGG-16	100		0.66	0.42	0.42

Table 2: Results of the models ran with the selected 50 patients

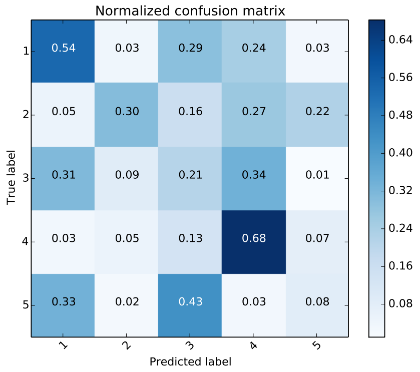


Figure 6: Confusion matrix for the dataset of 50 patients, 5 classes with 50% of the tiles.

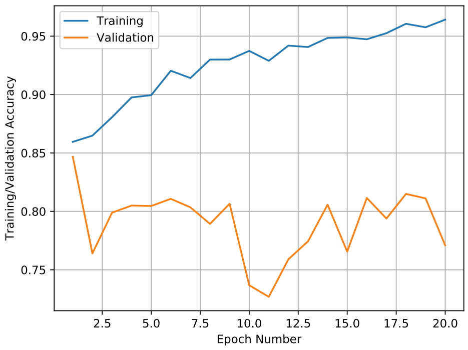


Figure 7: Training and validation accuracies when using 50 patients, 2 classes with 50% of the tiles.

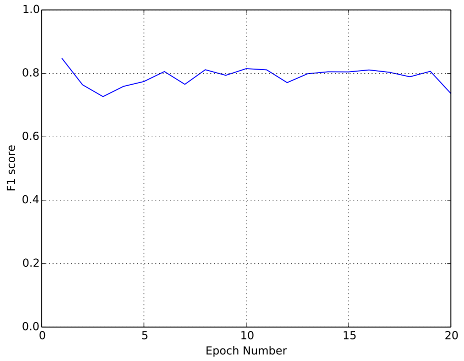


Figure 8: F1 score when using the imbalanced dataset of 50 patients, 2 classes with 50% of the tiles.

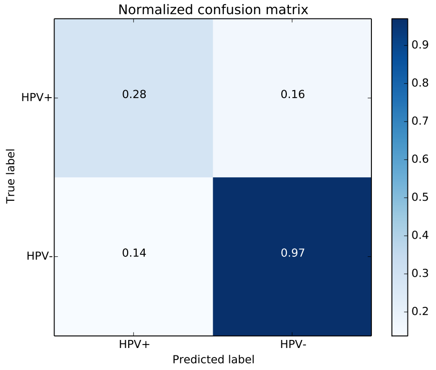
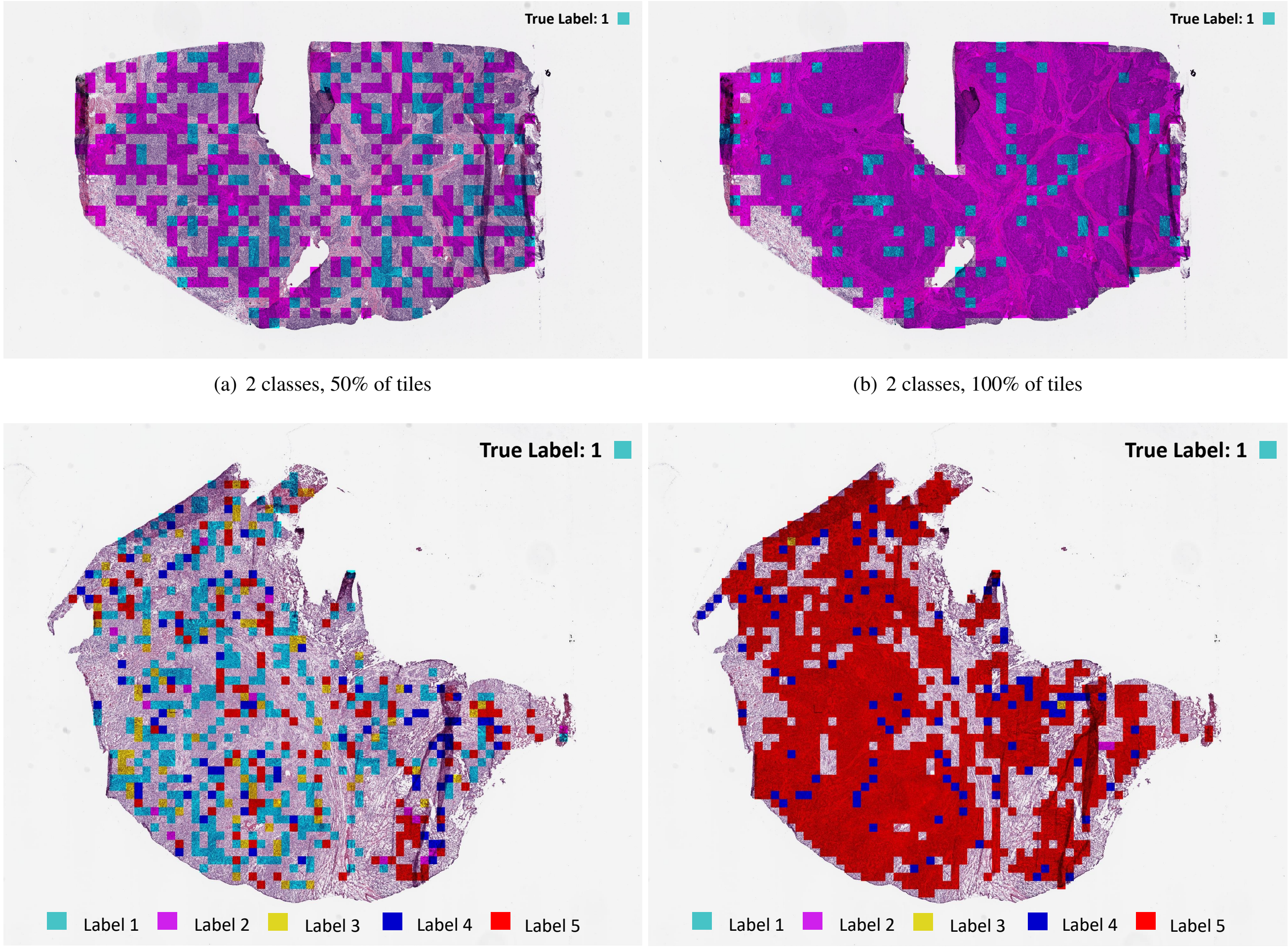


Figure 9: Confusion matrix for the dataset of 50 patients, 2 classes with 50% of the tiles.

threshold in the preprocessing step (see section Dataset and Data Preparation). We interestingly observed that when using the entire information of the WSI we obtain a distinct background label with another less frequent label. In some cases the change is not very drastic (Fig. 10 a and b) but in some cases a clear change can be observed (Fig. 10 c and d). Even though this is an interesting observation we would need further expert input in order to use that observation for future work. Another interesting observation we made is the fact that class 1 and 5 are mislabeled in most cases (see confusion matrix Fig. 6). This fact is an indication that classes 1 and 5 might be very similar in their appearance and, therefore, hard to distinguish. A discussion with Brennan ([2]) showed that the number of subgroups might also vary since it is a model choice in the k-means algorithm. This similarity of different classes could be a valuable input for an update of the molecular subgroups.



(a) 2 classes, 50% of tiles

(b) 2 classes, 100% of tiles

(c) 5 classes, 50% of tiles

(d) 5 classes, 100% of tiles

Figure 10: Heatmaps of predicted labels of the patches overlaid on top of the whole slide image. Note that each colored square is a $256 \times 256 \times 3$ image.

6. Conclusion and Future Ideas

1. High resolution WSIs are required to capture important cell features that may be indicative of a specific HNSC subtype. Unfortunately, this leads to significant down sampling, as the images are usually very large. Down sampling images can be problematic as the chosen tiles could lack information about the subtype.

2. It is suboptimal to train a model using labels at the WSI level. The cancer is not necessarily present in the whole tissue and is usually localized in small areas. This kind of discrimination could be performed by a medical specialist (i.e. pathologist) and the dataset could be significantly improved.

3. Given the localized nature of cancer, simple decision procedures such as majority voting or max-pooling are not effective. The problem must be approached in a way that labels are provided at a tile level.

4. The problem is better posed towards finding probability maps where cancer is most likely to be present serving as an aid and a second opinion to that of the pathologist. The heatmaps are a can be a valuable input for the pathologist to identify areas that contain subgroup information.

5. The similarity encountered in the confusion matrix (e.g. class 1 and 5) could be a valuable additional input for further investigations of the subgroups.

Acknowledgment

We are very thankful to Prof. Olivier Gevaert and his team for providing the labels of the data and introducing us to the new field. We would also like to thank our project mentor, Rishi Bedi, for the fruitful discussions about the project. Lastly, thanks are also due to the whole CS231n teaching staff for an awesome experience with convolutional neural networks.

References

[1] Camelyon2016. `https://camelyon16.grand-challenge.org/`. Accessed: 2017-06-04.

[2] K. Brennan, J. Koenig, A. Gentles, J. Sunwoo, and O. Gevaert. Identification of an atypical etiological head and neck squamous carcinoma subtype featuring the cpg island methylator phenotype. *EBioMedicine*, 17:223–236, 2017.

[3] Cancer.Net. Head and neck cancer: Statistics, 2017. [Online; accessed 16-May-2017].

[4] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.

[5] L. Hou, S. Dimitris, K. M. Tahsin, G. Yi, D. James E., and S. Joel H. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2424–2433, 2016.

[6] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*, 2017.

[7] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

[8] Wikipedia. Dna methylation — wikipedia, the free encyclopedia, 2017. [Online; accessed 15-May-2017].