Patch-based Head and Neck Cancer Subtype Classification

Wanyi Qian, Guoli Yin, Frances Liu,
Advisor: Olivier Gevaert, Mu Zhou, Kevin Brennan
Stanford University
wqian2@stanford.edu, guoliy@stanford.edu, francesl@stanford.edu
ogevaert@stanford.edu, muzhou1@stanford.edu, kbren@stanford.edu

Abstract

Head and neck cancer(HANC) is one of the most common cancer in the US. The subtype of HANC is mainly categorized based on the location of the tumor. Previous researches[1][4][6][7] confirmed the existance of genotype variation associating with molecular subtype. However, to our best understanding, there was no previous research exploring molecular subtype variation in terms of pathology image. To explore the molecular subtype variation in terms of pathology image, the project implemented a patch-based deep learning approach and VGG16, InceptionV3, and HNSVNet inherited from Le Hou[2] were trained and tested. The result of our project indicated the possibility of molecular subtype variation in terms of pathology image, and prediction power of CNN on HANC subtype classification.

1. Introduction

Head and neck cancer (HANC), ranks seventh most common form of cancer in the US, is used to describe cancers developed around throat, larynx, nose, sinuses, and mouth[1]. According to the location of the tumor, typically HANC is composed of laryngeal cancer, hypopharyngeal cancer, nasal cavity and paranasal sinus cancer, nasopharyngeal cancer, salivary gland cancer, oral and oropharyngeal cancer. Moreover, according to the deepness of the cancer located in the tissue layer, HANC can be divided into carcinoma in situ and squamous cell carcinoma, and the latter indicates deeper location in the tissue layer and most HANC are squamous cell carcinomas.

According to Vonn[1], head and neck squamous cell carcinoma (HNSCC) is a heterogeneous disease and there are no valid molecular characterization currently for cancer subtype analysis besides the effect of human papilloma virus (HPV). Classification of the subtype of cancer is of importance, since different subtypes correspond to different treatments. By identifying molecular subtype of cancer correctly and efficiently, targeted therapies can be arranged on time for patients.

Previously there were large amount of researches focusing on classifying molecular subtype of lung cancer, and there are few researches done on molecular subtype of head and neck cancer. Moreover, there were several researches focusing on analyzing gene expression difference analysis for possible subtype differentiation[1][4][6][7], but to the best of our knowledge, there is no previous research focusing on analyzing pathology images of HNSCC for molecular subtype differentiation in HNSCC. Due to the heterogeneity of the cancer that multiple gene sequences can result in the same disease, analyzing pathology images could be a more direct way for cancer subtype classification.

Currently, convolutional neural networks(CNNs) shows outstanding performance for image classification. However, one of the difficulty for utilizing CNNs for image analysis is for high resolution images due to high computational cost. For example, typical Whole Slide Tissue Images (WSI) are about gigapixel level and their size can be in gigabytes level. Since classifying cancer subtypes for potential appropriate treatment and progression is of importance, previously there were several researches aims for applying machine learning and CNN for WSI classification[2][3]. However, by applying CNN directly on WSI, the discriminative information can be diluted due image downsampling, which might also lead to data inefficiency[2]. Therefore, previously, Le Hou[2] proposed a patch-based CNN model for lung cancer subtype (binary) classification. Inspired by their architecture, we implemented HNSCNet in same architecture with half of the hyper-parameters for computational efficiency consideration and a VGG16 on HNSCC subtype classification. Among five subtypes of HNSCC, we are focusing on differentiating HPV-positive and HPV-negative subtype. The challenges in this project relies in unlabeled patches for whole image classification, no previous research in this cancer, and extremely large dataset in terms of the size.

In our project, we aim to explore the possibility of expression of molecular subtypes of HNSC on cell level.

After generating patches from each image, we proposed the label of each patches according to their associate images. InceptionV3, VGG16, and HNSCNet inheriting the architecture in Le Hou[2] were trained the make prediction on each individual patch. Majority vote was used to aggregate the predictions of each patch.

2. Related Work

Previously, in Camelyon16, ISBI challenge on cancer metastasis detection in lymph node competition, the team from Harvard Medical School and MIT compared different current neural networks' architecture including GoogLeNet, VGG16, FaceNet, and AlexNet for breast cancer metastasis detection in lymph node WSI[5]. According to the special structure of WSI, appropriate resolution should be chosen to preserve the information of WSI. By utilizing foreground segmentation, the region with lymph node is extracted from the white background. Moreover, in order to use all discriminative information in the images, instead of directly predict on WSI, they cropped the original image with 40X magnification into 256*256 patches. The similar preprocessing method is also mentioned in Le Hou[2]. In their project, GoogLeNet and VGG16 showed extraordinary performance in terms of accuracy 98.4% and 97.9% separately, therefore, VGG16 and InceptionV3 were decided to be used for HNSC HPV-positive and HPV-negative subtypes classification.

In Le Hou[2], the author proposed a model with expectation-maximization(EM) and CNN for discriminative patches generation and lung cancer subtype classification. After assuming all the patches generated from the image discriminative initially, their model made prediction on the probability of discrimination for each patch and eliminate patches with low probability according to certain threshold[2]. By iteratively eliminating indiscriminative patches, the first step provided discriminative patches for patch-level prediction and aggregation for image-level prediction. Their EM step achieved improvements of 6% accuracy compared to model without EM step. Given consideration of no previous research utilizing WSI in HNSC, the CNN model without EM step was tested first for simplicity concern. Since the CNN model showed reasonable prediction power in terms of 71% accuracy, the HNSCNet directly used the architecture mentioned in this paper and was tested on HNSC WSI.

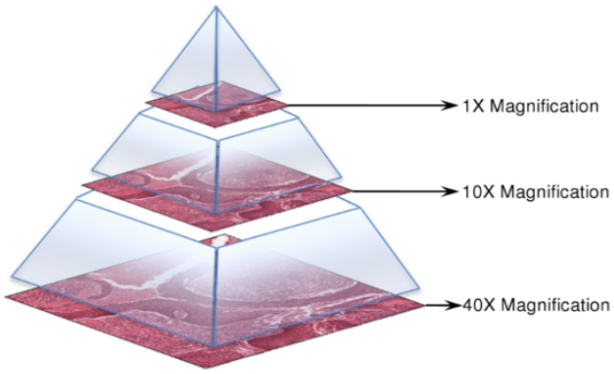


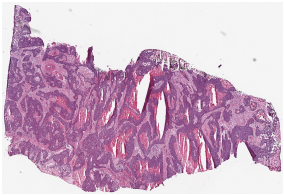
Figure 1: Examples of patch segmentation[8]

3. Experiments

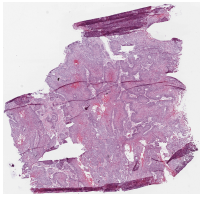
3.1. Dataset

The whole dataset is obtained according to the TCGA case ID in TCGA-HNSC program from national cancer institute GDC legacy archive[8] and contains WSI of 582 patients with different HNSC subtypes. For patient with multiple WSI, only one of the WSI is selected. Among the whole dataset, 79 patients are with HPV-positive subtypes and the rest are with HPV-negative subtypes. Within the HPV-negative samples, 4 major subtypes are identified based on cellular differences or location of tumor, but we choose to focus on the HPV binary subtypes due to its direct linkage to treatment response.

Given consideration of computational cost and training time, the final dataset contained 79 HPV and 86 HPV-negative patients, whose size is about 38GB. Unlike traditional images saved in png or jpeg format, a typical WSI file is of SVS format, and contains images with more than one resolutions as shown in Figure 1. We had attempted to convert the SVS images into lossless TIFF format, but the patch-generation pipeline dictates that the level information must be retained. Moreover, the TIFF file is much larger than the jpeg file of the same patch. Therefore, for computational space efficiency, the individual patches are saved in jpeg format for storage and computational efficiency in the following training and testing progress. In order to preserve the characteristic of cells, 40X magnification image is selected to generate patches for further data preprocessing.



(a) HPV-Positive WSI



(b) HPV-Negative WSI

Figure 2: HPV-positive and HPV-negative WSI

3.2. Data Preprocessing

For each WSI, two types of the patch size (224*224, and 500*500) were used to fit default VGG16 and our HNSC-Net. According to different size of the image, more than 70,000 patches maybe generated for one single slide. Some of these patches contain white to gray background with no tissue, or small proportions of tissue with respect to the size of the patch. We aim to save computing power and curate the learning process by only selecting patches from the area of interest. To visualize and extract patches from the image, OpenSlide[9], a python package, is used to extract 40X image from SVS file and generate patches. We use the data level with the highest magnifer to help better investigate cellular differences in cancer subtypes with the 224*224 patches, while maintaining some larger scale pattern with the 500*500 patches. The generated patches are in RGB channel. We index into the loaded Openslide instance, and access the patches (tiles according to Openslide) by grid. The extracted tile then gets converted into a numpy array of size 224*224*3 or 500*500*3. Subsampling the patches reveals that the background is normally set to be above 200 in value. We use this as a threshold to find the percentage of background versus tissue area, and set the filter at 30% to ensure that only patches with predominantly tissue area are fed into the neural networks. We agree that sometimes these background pixels have predictive power in the context of larger-scale patterns. However, it could also be an artifact of stretching or misplacement when pathology sections are prepared in the laboratories.

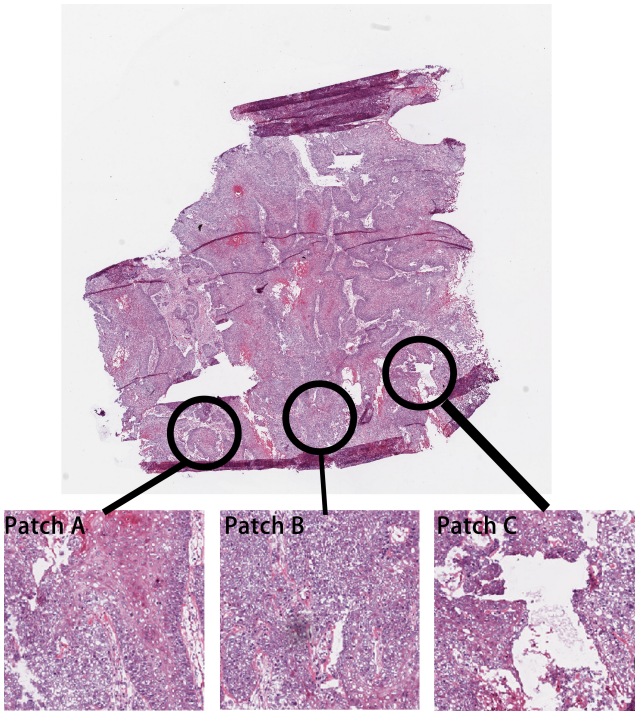


Figure 3: Examples of patch segmentation

We had also tried using edge detection or Otsu's method from OpenCV to delineate the tissue region. However, the binary mask is computationally expensive to generate due to WSIs' large file size. A high-resolution slide may take up to 300MB of space. It is much easier to apply numpy calculations to individual tiles after they had been produced by the optimized partitioning method in OpenSlide.

Besides foreground extraction, during data preprocessing, we also found several "broken" images in terms of gray gradient. In order to filter out those images, the variation of the patches is calculated and the patches with low variances (smaller than 2000) were filtered out. Eventually, with each slide, approximately 13,000 224*224 patches and 1,400 500*500 patches were generated. Moreover, to increase the training efficiency, only 10% and 20% of the generated 224*224 and 500*500 patches were used for further data training and testing, respectively. We sample the data randomly from preprocessed valid patches. The ratio of the training, validation, and testing dataset is 6:2:2. For each patch, it is assumed that it has the same label as its corresponding image. The model is trained on patches level and make prediction on each patch.

Moreover, to make each features' distribution even, the input is normalized before enter into the models. For data augmentation, since the number of patches for training are about 50,000, which is large for training process, we only introduced width and height shift, horizontal and vertical

flip to further increase the size of the training set.

3.3. Method

Loss Function: Cross-entropy loss

$$L_i = -log(\frac{e^{f_{yi}}}{\sum_j e^{f_j}}) \tag{1}$$

$$Loss_{softmax} = \sum_{i=0}^{N-1} L_i \tag{2}$$

The cross-entropy loss is a generalized version of logistic function. It is typically used for classification problem. By providing the "probabilities" corresponding to each label for a sample, the cross-entropy loss aims to maximize the probability of the sample belonging to a certain label.

3.3.1 VGG16

VGG16 is a convolutional neural network architecture named after the Visual Geometry Group from Oxford. It won the ILSVR (ImageNet) competition in 2014. Since it is still recognized as an excellent model, a pre-trained VGG16 model was implemented and trained on our dataset and two fully connected layers were added on the top of the model. Relu as a non-linear operation is also added on top of linear mapping. Then we used the softmax method to predict the probability of each class for each patch. For the VGG16 model, the first 15 layers (5 blocks) were frozen to train the model, which saves much time to get a baseline.

The test samples are approximately 12,000 patches generated from 25 examples and the accuracy is around 68% for 12,000 patches. In the meantime, a majority vote based on corresponding patches for each WSI image sample was implemented to give the final prediction. The CNN-Vote method can predict each WSI input image from its generated patches and the test accuracy for CNN-Vote is 70%.

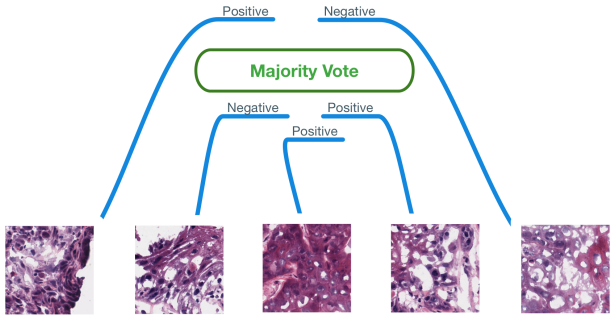


Figure 4: Examples of patch segmentation

Layer	Filter size, Filter
Conv	3×3, 64
Conv	3×3, 64
Max-pool	-
Conv	3×3, 128
Conv	3×3, 128
Max-pool	-
Conv	3×3, 256
Conv	3×3, 256
Conv	3×3, 256
Max-pool	-
Conv	3×3, 512
Conv	3×3, 512
Conv	3×3, 512
Max-pool	-
Conv	3×3, 512
Conv	3×3, 512
Conv	3×3, 512
Max-pool	-
FC + ReLu	1024
FC + ReLu	2
Softmax	-

Table 1: Architecture of VGG16

3.3.2 InceptionV3

The InceptionV3 was released by Google and the most important improvement is the implementation of factorization. It factorizes a 7×7 convolutional layer into two one dimensional convolutional layers. The advantage of this implementation is that it can speed up the computation, make the architecture much deeper and increase the non-linearity. We also implemented a pre-trained InceptionV3 model and added two fully connected layers on the top of the model which is similar to the top part of our fine-tune VGG model. And we used the softmax method to predict the probability of each class for each patch. For the InceptionV3 model, we freeze the first 172 layers to train the model, in order to save time. The test accuracy is around 72%.

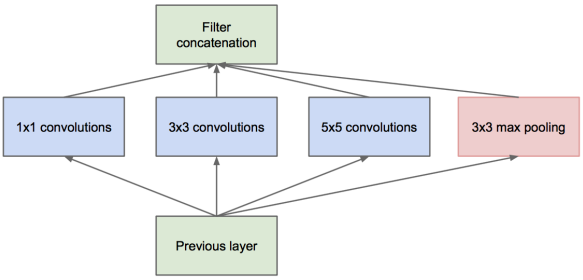


Figure 5: Architecture of naive Inception[10]

3.3.3 HNSCNet

A custom CNN model was implemented due to the limitations of pretrained VGG16 or InceptionV3 based on ImageNet. HNSCNet is a deep convolutional neural network inherited from Le Hou[2] and is built for HPV-positive and HPV-negative binary classification in HNSC. Its original size of 500×500 is larger than the input size for vgg model and it contains less layers but more filters for the top layers. The number of parameters is around 8 million.

HNSCNet was trained for 30 epochs, using around 45000 training and 15000 validation samples. The data is dynamically loaded and the batch-size is 32. In order to speed up and memory limitation, we use train steps and validation steps for each epoch. Validation loss decreases from 1.8 to 0.54, and the validation accuracy achieves 75%.

Due to the lack of ground truth annotation of the WSI and previous confirmation of difference in pathology image corresponding to HNSC subtypes, the convergence of HNSCNet is unpredictable at the beginning. However, the reasonable validation accuracy indicated that there exits difference in WSI associating with different HNSC subtypes.

Layer	Filter size, Filter
Conv	10×10, 80
ReLu + Batch Normalization	-
Max-pool	6×6
Conv	5×5, 120
ReLu + Batch Normalization	
Max-pool	3×3
Conv	3×3, 160
ReLu	-
Conv	3×3, 200
ReLu	
Max-pool	3×3
FC	320
ReLu + Dropout	-
FC	320
ReLu + Dropout	-
FC	2
Softmax	-

Table 2: Architecture of HNSCNet

	Predicted Positive	Predicted Negative
Actual Positive	4336	1037
Actual Negative	2749	3878

Table 3: Confusion matrix of VGG16

4. Results & Discussion

Take the test result of VGG16 model as an example. The confusion matrix showed that the total accuracy is about 68%. Moreover, false positive samples are more than false negative samples. With the majority vote, the confusion matrix highlights this phenomenon. The generated patches for each WSI image is around 1400. Based on large amount of patches for each WSI image, HPV test samples (positive) can be classified correctly into the correct class. This suggested that our models can distinguish HPV-positive cell type and may catch the visual features of those HPV-positive cells, while negative samples may obtain some similar tissue areas as positive samples, like blood tissue. The generated filters were checked for VGG model, as is shown in Figure 6&7. Those centralized bright pixels are related to cells extracted by the model. And this visual feature may suggest underlying difference between HPV cells (positive) and other cells (negative).

One of the challenge in this project is the lack of ground-truth labels for individual patches. Previous literature related to HNSC was almost focusing on genotype variation,

	VGG16	InceptionV3	HNSCNet
Test accuracy	0.69	0.72	0.75

Table 4: Test accuracy for patches level prediction comparison among three models

	Predicted Positive	Predicted Negative
Actual Positive	11	0
Actual Negative	8	7

Table 5: Confusion matrix of majority vote for result generated from VGG16

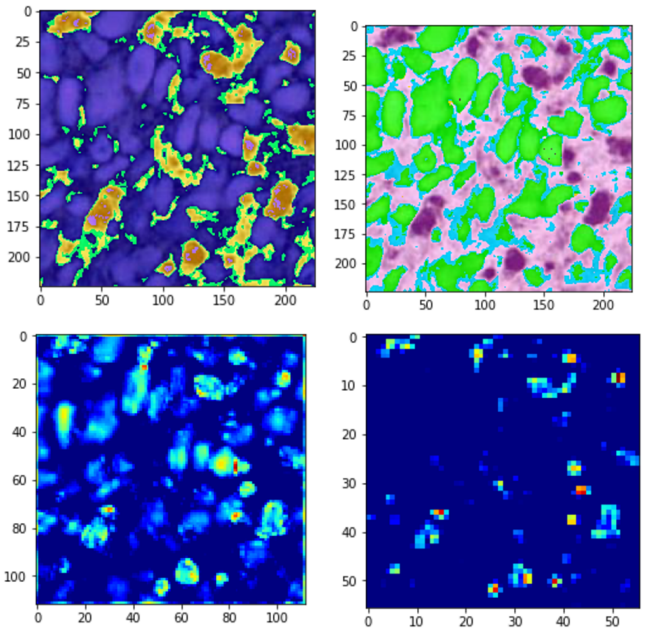


Figure 6: Filter visualizations of trained VGG model.
Top-left:raw(cmap-viridis).
Top-right:preprocessed(cmap-viridis).
Bottom-left:$block1\_pool$(cmap-jet).
Bottom-right:$block2\_pool$(cmap-jet)image(cmap:jet).
The clustered bright pixels mean some cell types are extracted by the model.

gene expression analysis, and other laboratory methods for testing HPV-positive in HNSC tumors, and no previous researches, to our best understanding, explicitly stated the capability of using pathology images in HANC subtype classification. Since there wasn't any publications detailing the annotation of HPV-positive region in HNSC pathology images, during the data preprocessing, the label of the image is distributed to all the patches generated. For the HPV-negative samples, it is safe to assume that all patches are HPV-negative. However, for the HPV-positive samples, there might be distinct regions in the WSI that show HPV-positive characteristics, while the rest are indistinguishable from HPV-negative cases. A major drawback of this assumption is that the training dataset is diluted by the noisy labeled HPV-positive patches.

For possible improvements of our model in the future, on feature selection side, since the shape of cell could vary from subtype to subtype, edge features could be potentially added to provide more information on cell shape. Moreover, on data preprocessing side, potential clustering method such as K-means(naive) or EM could be used to generate more discriminative patches from each image. For K-means, one of the challenge will reside in the definition of distance

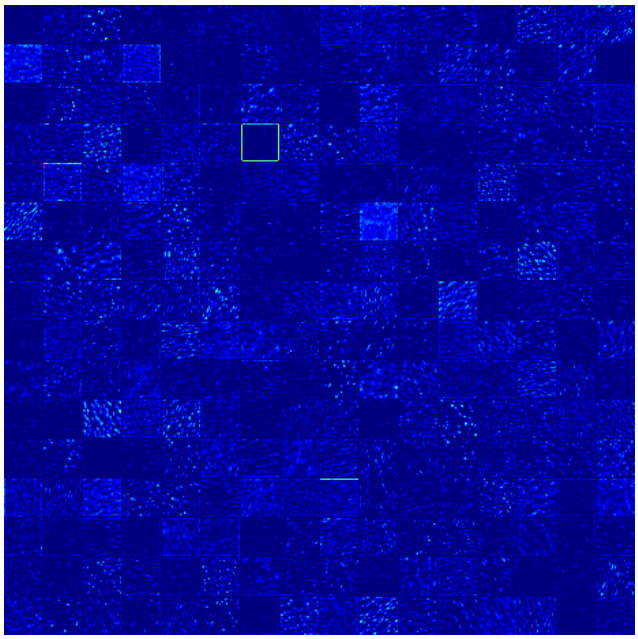


Figure 7: all filters from $block3\_Pool$ shown as a full image. The bright dots suggest cell types are extracted by the model.

among patches generated. For EM, one of the challenge would reside in the weight initialization, which will influence the prediction of discriminative patches. Since in the EM, the patches with low probability to be discriminative will be eliminated in the next round, and the patches for the next round will be generated from the region indicated to be discriminative, the initial weight could potentially have large influence on the convergence of the model.

In conclusion, our project confirmed the capability of using WSI for HNSC molecular subtype classification and the prediction power of CNN through patch-based CNN classification.

5. References

[1] Walter, Vonn, et al. "Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes." PloS one 8.2 (2013): e56823.

[2] Hou, Le, et al. "Patch-based convolutional neural network for whole slide tissue image classification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[3] Yu, Kun-Hsing, et al. "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features." Nature

Communications 7 (2016).

[4] De Cecco, Loris, et al. ”Head and neck cancer subtypes with biological and clinical relevance: Meta-analysis of gene-expression data.” Oncotarget 6.11 (2015): 9627-42.

[5] Wang, Dayong et al. ”DEEP LEARNING BASED CANCER METASTASES DETECTION”. Camelyon16 Presentation. 2016.

[6] Gevaert, Olivier, and Sylvia Plevritis. ”Identifying master regulators of cancer and their downstream targets by integrating genomic and epigenomic features.” Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. NIH Public Access, 2013.

[7] De Cecco, Loris, et al. ”Head and neck cancer subtypes with biological and clinical relevance: Meta-analysis of gene-expression data.” Oncotarget 6.11 (2015): 9627-42.

[8] National Cancer Institute GDC Data Portal, https://portal.gdc.cancer.gov/. Accessed 15 March. 2017.

[9] Goode, Adam, et al. ”OpenSlide: A vendor-neutral software foundation for digital pathology.” Journal of pathology informatics 4.1 (2013): 27.

[10] Szegedy, Christian, et al. ”Going deeper with convolutions.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.