Prediction of Gene Expression from Histopathology Images via Deep Learning in Gastric Cancer

Stanford CS271 Project Report (Autumn 2020)
Mentor: Yuming Jiang

Rui Yan
Stanford University
ruiyan@stanford.edu

Justin Xu Huang
Stanford University
jhuan104@stanford.edu

Victoria Valverde
Stanford University
vvalverd@stanford.edu

Abstract

Patients with gastric cancer are often treated with surgery followed by adjuvant chemotherapy. However, adjuvant chemotherapy's therapeutic benefits are inconsistent across patients. In studying the tumor, current methods usually ignore tissue context specifics, such as microsatellite instability (MSI), and gene expression levels. Both MSI and the expressions of CDX1, GZMB, SFRP4, and WARS genes are known to be linked to adjuvant chemotherapy response in patients with gastric cancer. However, patients do not normally receive gene expression panels for these four genes, restricting the utility of these genes as prognostic tools. In this paper, we successfully developed a deep learning model to predict gene expression of these 4 genes from readily-available, whole slide H&E images from patients with gastric cancer. Our prediction results are comparable to the state-of-the-art results. Further work is required to run the models with more sampled patches and confirm the potential for our results to predict adjuvant chemotherapy response and select patients that will benefit from treatment.

1 Introduction

Adjuvant chemotherapy after surgery is known to improve the survival of patients with gastric cancer. However, adjuvant chemotherapy's therapeutic benefits are inconsistent across patients, helping some while hurting others. Both tissue context and molecular profiling are crucial in predicting patient response to chemotherapy. Kather et al. 2019 showed microsatellite instability (MSI) can determine whether patients with gastric cancer will respond well to immunotherapy [1]. They applied deep learning to predict MSI status from whole slide images (WSI). Another recent study by Cheong at al. showed that expression levels of 4 genes (CDX1, GZMB, SFRP4, and WARS) can help predict whether patients will respond to adjuvant chemotherapy, as well as their overall prognoses [2]. However, not every patient is tested for expressions of these 4 genes, as this requires additional tests. We will build a deep learning model to directly predict gene expression from whole slide H&E images, which are readily available for all cancer patients. These gene expression values can be combined with Cheong et al.'s work to predict whether patients with gastric cancer will respond to adjuvant chemotherapy, from their WSIs of their tissues. This model may help identify patients who can benefit from adjuvant chemotherapy after surgery while preventing non-responders from receiving the treatment. We successfully trained and fine-tuned a deep learning model with a ResNet-50 backbone [3]. The model accepts tumor patches, extracted from whole slide images, as input, and outputs gene expression values for the four genes of interest. The predicted gene expressions achieved an average correlation of 0.312 across the the CDX1, GZMB, SFRP4, and WARS genes.

2 Related Work

Gastric cancer outcomes prediction. The primary motivation of this project was based on two papers. First, previous work by Kather et al. (2019) has been able to predict whether a patient with gastrointestinal cancer will benefit from immunotherapy directly from H&E histology images [1]. They applied deep learning methods directly to whole slide images (WSI) and predicted microsatellite instability (MSI), which is known to determine immunotherapy response in gastrointestinal cancer. The main assumption made by this group was that all areas of the tumor were equally correlated with MSI. We plan to build on this assumption for gene expression prediction by building a model that predicts response to chemotherapy by sampling from known tumor areas in the WSI rather than the whole slide. Second, Cheong et al. (2018) presented a parallel insight that the expression of 4 genes (CDX1,GZMB, SFRP4, and WARS) together were highly predictive of quality of response to adjuvant chemotherapy and prognosis in patients with stage II-III gastric cancer [2]. They developed two rule-based classifier algorithms to classify patients as likely responders or nonresponders, as well as low/intermediate/high risk patients. However, to use this algorithm to identify candidates for adjuvant chemotherapy, physicians must order additional tests for gene expressions, incurring additional costs. We plan to develop a model to predict the expression of these 4 genes from HE whole-slide images, which are readily available for all cancer patients. We hope to combine our predictions with Cheong et al.'s findings to create an efficient, cost-effective method to identify candidates for adjuvant chemotherapy.

Predicting molecular markers from histology images. With the above two insights informing our hypothesis, we looked for previous work predicting molecular signatures from whole slide images. There has been much previous work seeking to predict tumor mutations from whole slide images in various cancer types, such as lung cancer [4], melanoma [5] [6], and others [7] [8]. However, less work has been done in predicting gene expression from whole slide images. Schmauch et al. predicted gene expression of many genes from whole slide imaging tissues of 28 different cancer types [9]. We based our baseline model on their work. Their method predicts gene expression from an image as follows:

- Sample 8000 224$\times$224 patches from the image, filtering out patches with only white space.
- Run a 50-layer ResNet, pretrained on ImageNet, on each tile to produce a 2048-feature output.
- Clusters the 8000 images into 100 groups based on tile location
- Compute representative "supertiles" for each group by averaging the 2048 features within each group.
- Run a multi-layer perceptron (MLP) on these supertiles to predict gene expression per supertile.
- Sample a random integer K between 1 and the number of supertiles.
- Computes expression for any gene X for the full image by taking the average predicted expression of the K supertiles with the highest predicted expressions of X.

As we look to build our tailored model structure to make gene expression predictions, we have been searching for other previous work in the space to consider for choosing: (1) where to sample tumorous patches in the whole slide image, and (2) how to predict gene expression from the sample. More recent work by Kather et. al (2020) was able to predict immune related gene expression signatures, known to be involved in response to cancer treatment, from routine histology images [10]. Their image pre-processing approach, neural network training, model selection and hyperparamater optimization was different from Schmauch et al., and we plan to take inspiration from their work in the expression prediction. Other work from Fu et. al (2020) used deep transfer learning to identify molecular patterns in whole slide images, and was able to predict mutations, tumor composition and prognosis [11]. The spatially resolved tumor and non-tumor tissue distinction they achieve could be a good addition to our process of sampling tumor tiles, which we currently do through previously computed heatmaps showing presence of tumor probability.

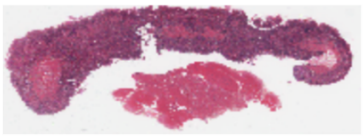
3 Data

For our analysis, we will develop a model to predict expression of the CDX1, GZMB, SFRP4, and WARS genes from whole slide, H&E-stained tissue images of gastric cancer tumor biopsies. Our

data comprises paired samples of whole slide biopsy images (∼450 patients, 82 GB, Fig.1) and gene expression panels (∼420 patients, 300 KB, Fig. 3). Thus, we have over 400 paired samples of data, all coming from The Cancer Genome Atlas Stomach Adenocarcinoma data collection (TCGA-STAD) (https://portal.gdc.cancer.gov/). From our mentor's lab, we also received heat maps identifying likely tumorous regions of each slide (Fig.2, 36 MB). For each whole slide image, we select multiple 224 x 224 pixel patches from regions of maximum tumor probability, and train our model to predict gene expression from these patches. Using the heat map data to generate these 224×224 patches of tumorous tissue was an unexpected challenge (see data pre-processing for details). Beyond this obstacle, our data appears to be presented in very usable formats. We are grateful to our mentor for this support.

a

b



c

id	NT	GZMB	WARS2	SFRP4	CDX1		log(1+a)	GZMB	WARS2	SFRP4	CDX1
TCGA-3M-AB46-01A-11R-A414-31	1	39135.63	192558.1076	547087.9	1096822			15.25623	17.55494	19.06142	20.0649
TCGA-3M-AB47-01A-22R-A414-31	1	13130.24	176891.6975	3834845	4980274			13.68072	17.43251	21.87074	22.24779
TCGA-B7-5818-01A-11R-1602-13	1	647612.9	67065.58522	106325.4	3343.863			19.30477	16.03331	16.69814	11.70773
TCGA-B7-A5TI-01A-11R-A31P-31	1	121190.6	140369.1188	789702.3	102768.3			16.88693	17.09888	19.59095	16.64905
TCGA-B7-A5TJ-01A-11R-A31P-31	1	24483.3	113836.9466	1002333	2040894			14.57957	16.79662	19.93493	20.96077
TCGA-B7-A5TK-01A-12R-A36D-31	1	2952078	76129.68692	564872.4	8402.729			21.4933	16.21619	19.10757	13.03681
TCGA-B7-A5TN-01A-21R-A31P-31	1	75762.67	97657.67238	2691767	56997.93			16.20922	16.57546	21.36012	15.79865

Fig. 1. **a.** Heat map identifying regions of the whole slide image that are likely tumorous. **b.** H&E-stained whole slide image of gastric tissue biopsy. **c.** Gene expression data for the CDX1, GZMB, SFRP4, and WARS genes; these genes were previously shown to be predictive of chemotherapy response in gastric cancer patients.

4 Approach

Our baseline model and overall methodology is heavily based on Schmauch et al.'s implementation [9], mentioned above. Their predicted gene expressions had significant correlations with true gene expression in 8/28 cancer types, but not gastric cancer, our cancer of interest. We modified several parts of Schmauch et al.'s workflow and model for our project. Thus, we improved upon their methods in a few ways:

- **Sampling tumorous patches from the image:** Schmauch et al. sampled 8000 tumor patches from each image, requiring that each patch contained tissue rather than solely white space. However, they did not sample patches that were likely tumorous from the tissue. For our approach, we sample 10 patches from each image, each of which is likely tumorous. We do so by using a heat map that signals where the highly tumorous tissue areas are. We then use a K-Means clustering model, fit to a wide array of 224×224 patches, to filter out potentially non-tissue patches that were artifacts of the heat maps. We hope that using likely-tumorous patches will yield a stronger signal in predicting gene expression from the tumor tissue.

- **Computing gene expression with more biologically-grounded methods:** Their computation of whole slide gene expression by averaging the top K tiles for some random value K does not appear to have biological grounding, and was performed to improve the model's robustness. After sampling the likely-tumorous patches, we extract ResNet features for each patch, average these features, then predict one gene expression value from this averaged vector. Thus, all patches used in our gene expression prediction are likely tumorous, providing our approach with more biological grounding.

4.1 Sampling tumorous patches from the image

First, we used the heat map data to generate 10 224×224 patches of tumorous tissue from each patient's whole slide image (Fig. 2). Not all patients with whole slide image data (451 patients) had corresponding heat maps (404 patients), so we were left with a total of 404 whole slide image-heat map pairs. To extract the tumorous patches, we first select a region of high tumor probability from the heat map, starting by looking at the top left corner of the image. We then extracted a 224 $\times$ 224 patch from the corresponding location on the whole slide image and ran the sampled patch through an unsupervised K-means ($K = 3$) clustering algorithm. This algorithm determines whether the extracted patch contains meaningful tissue. We do this because many samples have annotations on the image which are usually mapped to high tumor probability but do not actually represent a tissue area in the sample (Fig. 3). By running this clustering, we ensure that the sampled image is significant by comparing it to other previously sampled significant patches. If the sampled patch is significant, we min-max normalize the patch by subtracting the minimum pixel intensity from each pixel value and dividing each value by the maximum pixel intensity, then save the patch for later use. If it is not significant, we sample a new patch.

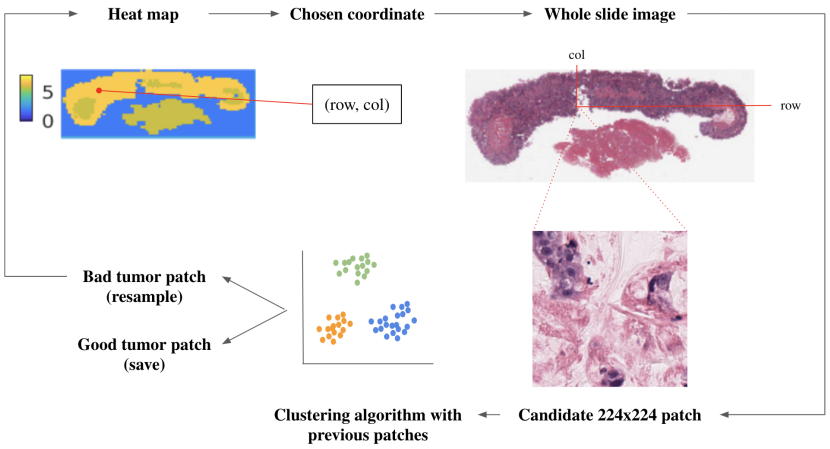


Fig. 2. Sampling tumorous patches from whole slide image workflow for a single patch.

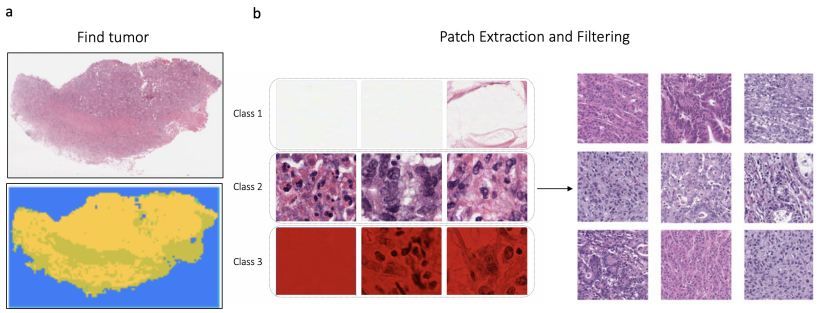


Fig. 3. 3a: H&E image and its corresponding tumor probability heatmap, 3b: K-means (K=3) clustering for patch filtering: in Class 1, we have white patches (background); in Class 2, we have the patches that represent tissue areas; in Class 3, we have the patches with artifacts. We only select the patches in Class 2 for model training.

4.2 Models

4.2.1 Model Inputs

Our predictive model takes the extracted 224 $\times$ 224 patches and gene expression tuples as input for training, and the tumor patch alone as input for testing. The label is a tuple consisting of 4 gene

expressions in FPKM-UQ units (an estimation of gene expression from RNA-seq data). Since the raw gene expression values cover several orders of magnitude, we normalize the values by performing a $\log(1+\alpha)$ transformation. After removing the duplicated patient ids and matching each label to the H&E image with the same patient id, we have 350 H&E images left, where each image has multiple 224×224 patches. Thus, the total data used to train and test the model consists of 350 samples, each with an image (with multiple 224×224 patches) and a label (tuple consisting of 4 gene expressions as floating numbers). We split the data into training (60%), validation (20%) and test (20%) sets.

4.2.2 Model architectures

- **Baseline Model.** The model takes one 224×224 patch from each H&E image as input and extracts 2048-features from each image patch from the global average pooling layer in ResNet-50, pretained on ImageNet. This feature vector is then passed to a multi-layer CNN model with input_dim = (# patches, 1, # features) and output_dim equals to 4 given that we are interested in predicting the 4 relative gene expressions. This baseline model architecture is shown in Fig. 4.

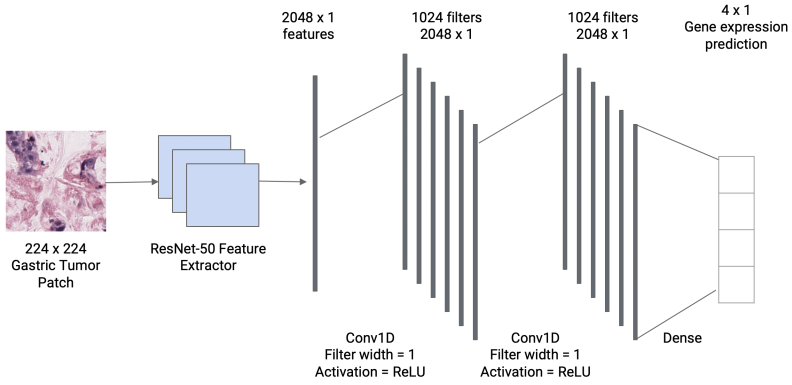


Fig. 4. Model architecture: Baseline Model

- **Simplified Version Model.** The simplified version model is almost equivalent to the baseline model. Instead of extracting the features from pre-trained ResNet-50 model, we stack the pre-trained ResNet-50 layers excluding the fully connected layers with a global average pooling layer, followed by multiple fully connected layers for retraining to predict the four gene expressions.
- **Patch-based Model.** The patch-based prediction model takes $N > 1$ patches from each H&E image, which are converted to $2048 \times N$ features using the ResNet-50 feature extractor. After that, we perform a patch aggregation using both max and average aggregation methods across all the patches to obtain $2048 \times N$ features. This is followed by the same multi-layer CNN model we used in our baseline model to obtain a 4×1 predicted gene expressions. The patch-based model framework is shown in Fig. 5.

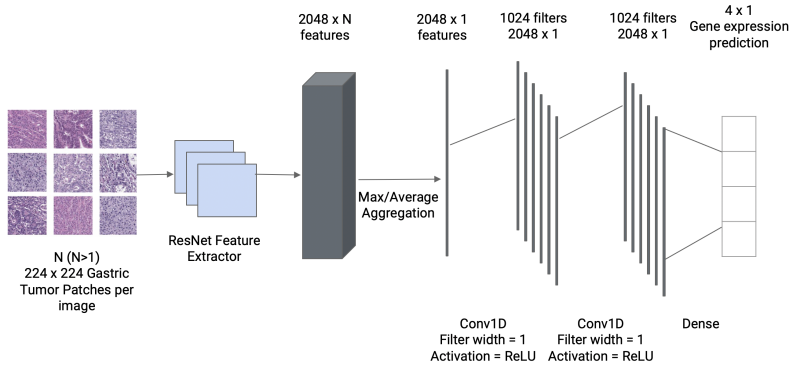


Fig. 5. Model architecture: Patch-based Model

- **Multi-task Model.** The multi-task model is built based upon the baseline and patch-based model. Instead of using one fully connected layer for four genes prediction, we use four separate fully connected layers for each gene prediction. This is inspired by the idea that each of the four gene expression predictions can be considered as separate tasks which have something in common in their parameters. The patch-based model framework is shown in Fig. 6.

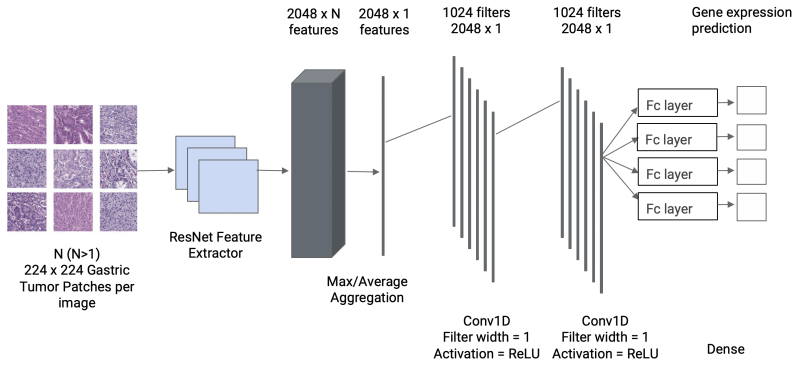


Fig. 6. Model architecture: Multi-task Model

4.2.3 Model outputs

The model outputs a tuple consisting of 4 gene expression predictions as floats. Since the predicted gene expressions are continuous variables, mean squared error (MSE) is applied as loss during model training. For each gene, we plot the predicted vs true expression as shown below (Fig. 7).

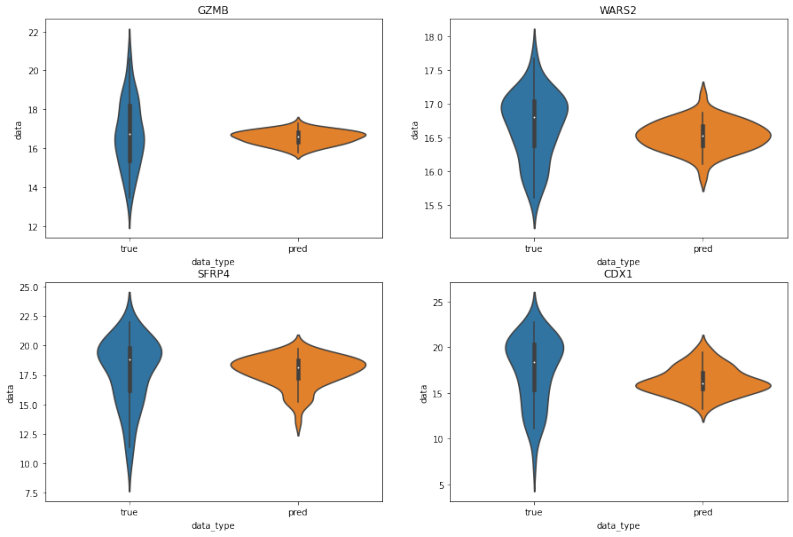


Fig. 7. Sample true vs predicted expression violin plots (Baseline model, N=1 patches)

5 Experiments

Our project aims to build a model to predict gene expression from whole slide images (WSI) for gastric adenocarcinoma (STAD). In particular, we are interested in predicting the gene expression values for the following 4 genes: CDX1, GZMB, SFRP4 and WARS. Since the predicted gene expressions are continuous variables, we use the mean squared error (MSE) to compute the loss during model training. Furthermore, to evaluate the accuracy of our gene expressions prediction, we apply the Pearson correlation $r =$ [9][12], a test statistic that measures the statistical relationship

between two continuous variables:

$$\text{Pearson Correlation } r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}},$$

where x_i is the true gene expression values, $\bar{x}$ is the mean of the true values, y_i is the predicted gene expression values and $\bar{y}$ is the mean of the predicted values. We compute the Pearson correlation between the true gene expressions and predicted gene expressions for the four genes. We run the experiments with three models with two patch aggregation approaches (maxpooling and avgpooling) and vary number of patches per image ($N \in \{1, 10\}$) described in Section 4.2.2 and the results are shown in Table 1. The patch-based model in multitask-setting with 10 patches per image achieved the highest Pearson correlation, which is 0.3122, among all experiments. This shows a statistically significant relationship between the predicted and true values.

	Pearson Correlation r	
Model	num_patches (N) = 1	num_patches (N) = 10
Simplified Version Model (M1)	0.2629	N/A for $N > 1$
Baseline Model (M2)	0.2635	N/A for $N > 1$
Patch-based Model + Average Patch Aggregation (M3)	0.2635 (same as M2 for $N = 1$)	0.3043
Patch-based Model + Maximum Patch Aggregation (M4)	0.2635 (same as M2 for $N = 1$)	0.3119
M3 + Multi-task	0.2751	0.3042
M4 + Multi-task	**0.2751**	**0.3122**

Table 1: Model results

5.1 Effect of changing number of patches

According to Table 1, when we sample more patches from each image and perform patch aggregation, the model gives higher Pearson correlation and thus achieves better performance. Sampling only one patch from each image may be biased and may not be representative of the whole H&E image, while sampling and aggregating more patches allows us to get a better representation of tumor tissue in predicting gene expression. Due to the computational intensity of additional patch processing, we ran the experiments with only num_patches_per_image $= 1$ and num_patches_per_image $= 10$. For the next step, we will sample more patches, say 100 patches, from each WSI to better examine the effect of number of patches on the prediction results.

5.2 Effect of changing model architecture

According to Table 1, the Pearson correlation for M1 and M2 are very close to each other because these two architectures are almost equivalent to each other. For M1, we freeze the initial layers in ResNet-50 and modify the fully connected layers for gene expressions prediction. For M2, we extract the features from the global average pooling layer in the pre-trained ResNet-50 model. In addition, we found that applying maximum aggregation to the extracted features gives better results compared to Average aggregation method. Moreover, we observed that the model performs the best in multi-task setting where we use separate fully-connected layers to predict the gene expression for each of the four genes.

6 Conclusion and Next Steps

Our results suggest that sampling more tumor patches from each patient results in better predictions of gene expression. This is likely because the additional tumor patches provide more information

regarding the state of each patient's tumor. Further, our results suggest that different patch aggregation methods (average aggregation, max aggregation) and multi-task learning approaches do not result in substantially different predictions in gene expression.

For the next step, we will experiment with sampling more patches for each image to get more tumor context for each prediction. We will then connect our predicted gene expressions with Cheong et al.'s rule-based classifier of chemotherapy response and perform a similar survival analysis [13] in order to predict whether a patient will respond to chemotherapy from the patient's HE whole-slide image. To do so, we plan to binarize our output predictions and optimize the binarizing threshold to maximally separate responders from non-responders. We will then validate our response predictor on an external dataset to test the model's ability to generalize to non-TCGA populations.

7 Contributions

VV, RY and JH contributed to the writing and review of the abstract. VV took lead on writing of the introduction. VV, RY and JH contributed to finding previous related work, VV and JH took lead on writing of the related works section. Data was provided by our mentor YJ, RY took lead on writing the data section. JH, RY and VV took part in designing the approach, JH took lead on writing the model design subsection, RY and VV took lead on writing the evaluation subsection. JH, RY and VV took part on designing experiments. JH took the lead on building patch sampling, VV and RY took lead on writing the patch sampling section. RY took lead on building the model, VV and RY took lead on writing the baseline model section. JH, RY and VV took part on experiments planning, JH took lead on writing the experiments section. VV, JH and RY took part on drafting the conclusion of the project. Overall, JH, RY and VV equally contributed to the revision and editing of the paper. YJ, our mentor, gave us guidance in all sections. We are very grateful for his advice.

References

[1] Jakob Nikolas Kather, Alexander T Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature medicine*, 25(7):1054–1056, 2019.

[2] Jae-Ho Cheong, Han-Kwang Yang, Hyunki Kim, Woo Ho Kim, Young-Woo Kim, Myeong-Cherl Kook, Young-Kyu Park, Hyung-Ho Kim, Hye Seung Lee, Kyung Hee Lee, et al. Predictive test for chemotherapy response in resectable gastric cancer: a multi-cohort, retrospective analysis. *The Lancet Oncology*, 19(5):629–638, 2018.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567, 2018.

[5] Javad Noorbakhsh, Saman Farahmand, Mohammad Soltanieh-ha, Sandeep Namburi, Kourosh Zarringhalam, and Jeff Chuang. Pan-cancer classifications of tumor histological images using deep learning. *bioRxiv*, 2019.

[6] Hongming Xu, Sunho Park, Jean René Clemenceau, Nathan Radakovich, Sung Hak Lee, and Tae Hyun Hwang. Deep transfer learning approach to predict tumor mutation burden (tmb) and delineate spatial heterogeneity of tmb within tumors from whole slide images. *bioRxiv*, 2020.

[7] Andrew J. Schaumberg, Mark A. Rubin, and Thomas J. Fuchs. H&e-stained whole slide image deep learning predicts spop mutation state in prostate cancer. *bioRxiv*, 2018.

[8] P Chang, J Grinband, BD Weinberg, M Bardis, M Khy, G Cadena, M-Y Su, S Cha, CG Filippi, D Bota, et al. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *American Journal of Neuroradiology*, 39(7):1201–1207, 2018.

[9] Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, et al. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nature communications*, 11(1):1–15, 2020.

[10] Heij L.R. Grabsch H.I. et al. Kather, J.N. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature cancer*, 1(1):789–799, 2020.

[11] Ramon Viñas Torne Santiago Gonzalez Harald Vöhringer Mercedes Jimenez-Linan Luiza Moore Moritz Gerstung Yu Fu, Alexander W Jung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature cancer*, 1(1):800–810, 2020.

[12] Anna V Mikhaylova and Timothy A Thornton. Accuracy of gene expression prediction from genotype data with predixcan varies across and within continental populations. *Frontiers in genetics*, 10:261, 2019.

[13] Yuming Jiang, Cheng Jin, Heng Yu, Jia Wu, Chuanli Chen, Qingyu Yuan, Weicai Huang, Yanfeng Hu, Yikai Xu, Zhiwei Zhou, et al. Development and validation of a deep learning ct signature to predict survival and chemotherapy benefit in gastric cancer: A multicenter, retrospective study. *Annals of Surgery*, 2020.