SlideSleuth: An Integrative Deep Learning Framework to Classify Lung Adenocarcinoma From Whole Slide Images

Jackson Howe

May 1, 2023 – August 15, 2023

1. Introduction/Background

1.1 Lung Adenocarcinoma

Lung adenocarcinoma (LUAD) is the leading cause of cancer death among men and women in the western world (Wei et al. 2019). There are 5 main histological subtypes of LUAD – lepidic, acinar, papillary, micropapillary, and solid (Yang et al. 2022). Of the 5 main subtypes of LUAD, only the lepidic histological subtype is considered noninvasive, with the other 4 types considered invasive (Yang et al. 2022). The distinction between invasive and noninvasive subtypes of LUAD is important because an invasive subtype tumour is treated much more aggressively than a noninvasive counterpart, and the noninvasive tumour may not even be treated at all, if physicians deem the case to have low enough risk. Distinguishing LUAD subtypes is a challenging task for both physicians and intelligent machines (Fig. 1).

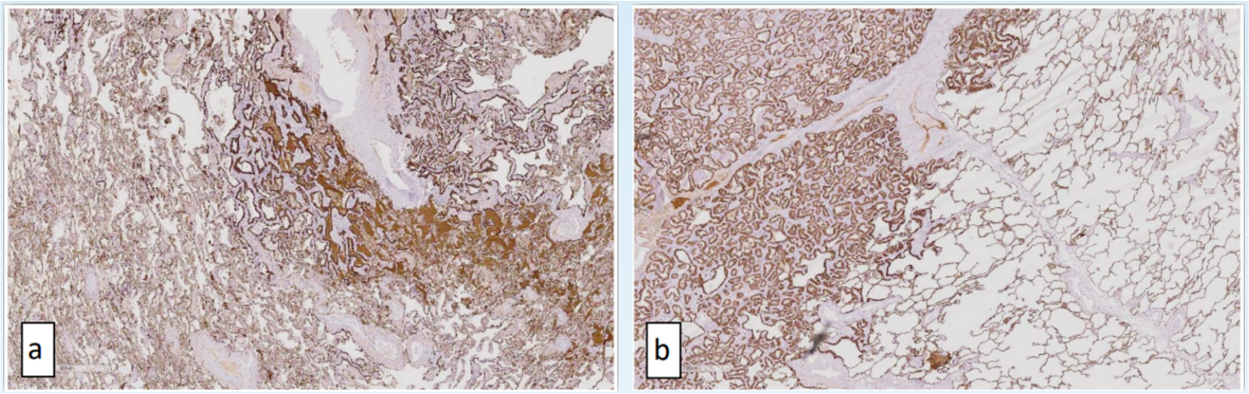


Figure 1. Spot the (lack thereof a) difference between: (a) invasive LUAD (b) non-invasive LUAD with CK7 stain

1.2 Deep Learning

Deep learning is a type of machine learning that uses additional, “hidden” layers in a neural network model to achieve a latent representation of a problem or observation, and thus deep learning is a type of representation learning. The most popular supervised deep learning model for analyzing images is a convolutional neural network (CNN). Rather than weights and biases between layers, a CNN includes filters between layers, which are small matrices that slide over an image and take the dot product at each successive subsection of the image, effectively downsampling the image in a creative manner. Much has been studied about CNNs since their introduction circa 2012, and I encourage those seeking more information about CNNs to view one of the numerous educational resources on this topic on the internet.

Another well-known deep learning model is the autoencoder (AE). An AE is an unsupervised learning model composed of two multilayer perceptrons, the encoder – which takes the input and reduces it to a lower-dimensional latent representation, and the decoder – which takes that lower-dimensional latent vector and recreates the original input. An augmented version of the autoencoder, the variational autoencoder (VAE), encodes a single point as a statistical distribution, then samples from that distribution to encode the input into the latent vector. VAEs have become very popular in the last few years as generative models, in which the latent model representation is used to generate data.

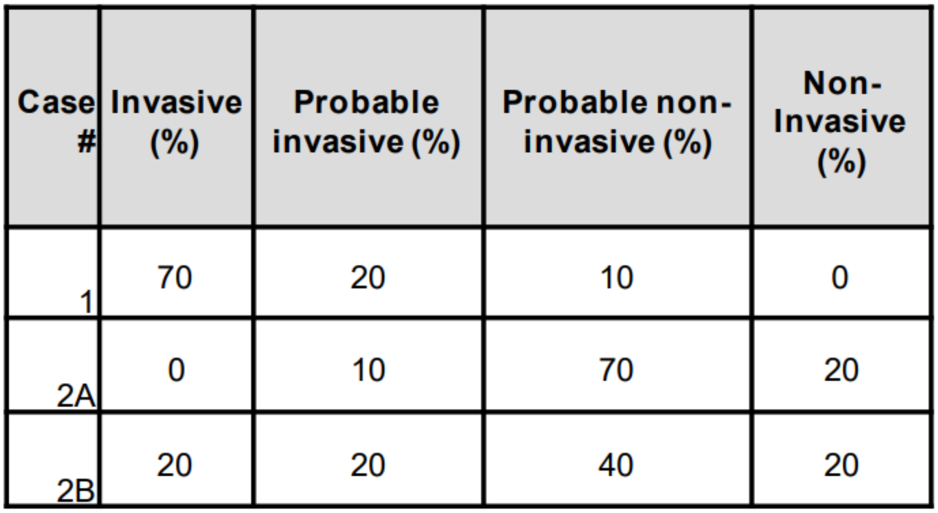
1.3 Past Work

Past work in the area of digital histopathology revolves around training a CNN classifier to identify classes of histopathological images, usually H&E stained whole slide images (WSIs). Coudray et al. classified WSIs as LUAD or lung squamous cell carcinoma (LUSC) with an area under the curve (AUC) of 0.97, and predicted commonly mutated genes with AUCs ranging from 0.733 to 0.856 (Coudray et al. 2018). Wei et al. classified the 5 histopathological subpatterns of LUAD with a Kappa score of 0.525, which was higher than the inter-pathologist score of 0.485 (Wei et al. 2019). Saednia et al. used a high-magnification network combined with a low-magnification network and attention mechanism to achieve an AUC score of 0.975 (Saednia et al. 2022). Chen et al. Achieved an AUC score of 0.959 using WSIs that had no annotations to identify tumorous regions with a dataset of 5915 WSIs (Chen et al. 2021). While much has been made of benchmarking the task of classifying WSIs, there is little research to explain what the model is doing whilst classifying images, and what features are important to the model during classification. Yang and Tsao curated a LUAD WSI dataset and expert thoracic pathologists manually classified it, achieving an intraclass correlation coefficient (ICC) of 0.5 (Yang et al. 2022). When the WSIs had a CK7 stain, the ICC increased to 0.6 (Yang et al. 2022).

2. Methods

2.1 Dataset

The dataset used for the LUAD classification and feature detection consists of 106 patients curated from University Health Network in Toronto, Ontario, Canada, 53 of which were diagnosed with lepidic-predominant pT1N0 lung adenocarcinoma, and 53 of which were diagnosed with acinar-predominant pT1N0 adenocarcinoma. Some patients had multiple WSIs, and thus the dataset consisted of 158 LUAD WSIs. There are two copies of each WSI, one with H&E stain, and one with CK7 stain. The WSIs were blindly scored by 2 thoracic pathologists, 1 fellow, and 1 resident for percentage of noninvasive, probable noninvasive, invasive, and probable invasive components (Fig. 2). The percentage scores were normalized to be on the interval [0,1]. For each case the scores were normalized so they were in the interval [0,1], then they were augmented so that a score of invasive was given a weighting of 1, a score of noninvasive was given a weighting of 0, and the two probable cases were given a weighting of 0.5. For example, for the case 1 below, the overall score would be 0.7*1 + 0.2*0.5 + 0.1*0.5 + 0*0 = 0.85. These final scores were then averaged over the 4 raters to come up with a label in [0,1] for the image, where 1 is highest confidence invasive, and 0 is lowest confidence invasive.



Case #	Invasive (%)	Probable invasive (%)	Probable non-invasive (%)	Non-Invasive (%)
1	70	20	10	0
2A	0	10	70	20
2B	20	20	40	20

Figure 2. Example of estimated percentage scoring by pathologists, fellow, and resident

The TCGA-PAAD dataset of diagnostic WSIs from GDC was used as further validation data for our model. The dataset contained 256 WSIs labeled with a TCGA label of either “Tumour Primary,” or “Solid Tissue Normal,” with there being 219 tumour primary, and 37 solid tissue normal slides. As the project was done on the UHN dataset, this dataset will not be discussed throughout the methods section, but all operations done with the UHN dataset were also done in validation with the TCGA dataset.

2.2 Data Processing and Augmentation

The 158 WSIs were tiled at 5x magnification into 224x224 pixel RGB tiles. Tiles were filtered out using Otsu’s method, where tiles with more than 50% background were not used in the dataset (Fig. 3). Tiles were assigned the label of their respective slide, which means that the dataset does have significant noise. For the classification task, we binned the labeled dataset into 3 bins: invasive, noninvasive, and undefined. We took any label in [0,0.4] to be noninvasive, any label in (0.4,0.6) to be undefined, and [0.6,1] was considered invasive. Overall, there were a total of 144,024 image tiles, with 64,448 invasive, 40,894 noninvasive, and 38,682 undefined.

When training the classifier, data was augmented to include a random rotation in the range of [-20, 20] degrees, a decision to horizontally flip the image at a probability of 0.5, a brightness range of [0.5x, 1.5x], a horizontal shift range of [-0.2, 0.2], and a shear range of [-20, 20] degrees. The data for the unsupervised autoencoder was also augmented with a decision to horizontally flip the image at a probability of 0.5, a rotation range of [-20, 20], a shear range of [-0.2, 0.2], and a width shift range of [-0.1, 0.1].

During the tiling process, some of the tiles become corrupt, usually only about 0.02% of tiles in the dataset. These corrupt tiles were removed from the dataset, and not used in model training or testing.

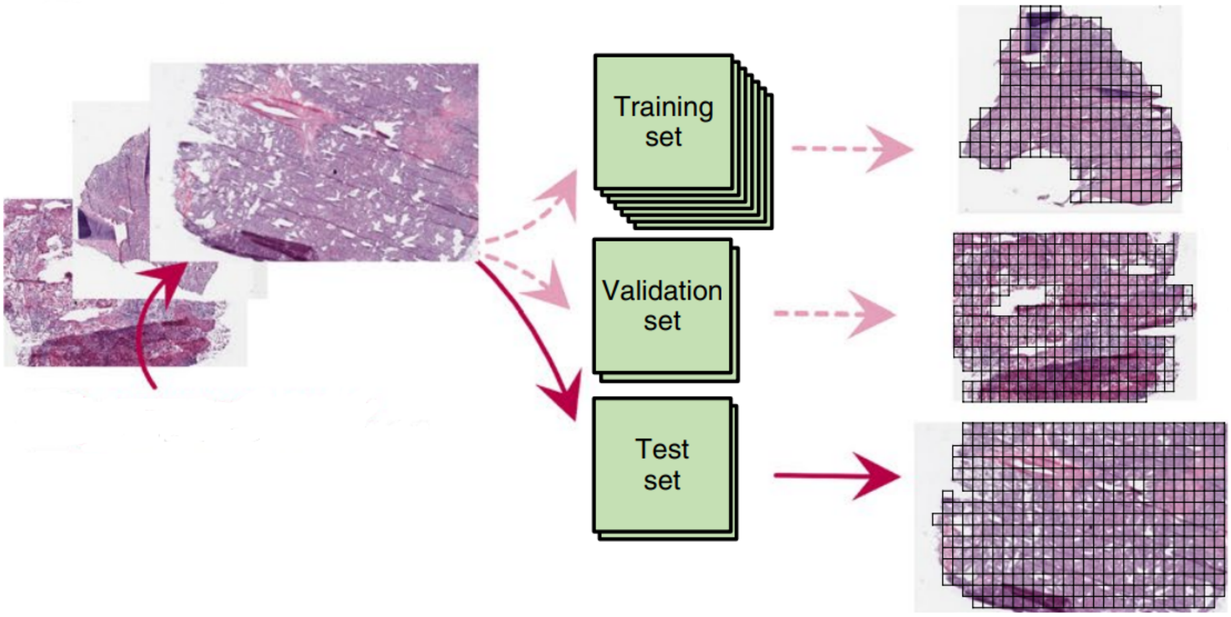


Figure 3. Illustration of data processing and tiling (Coudray et al. 2018)

2.3 Supervised Classification Model Workflow

We used a classification model with the following architecture: ResNet50 without the top classification layer, and a final output activation layer using the sigmoid activation function. We initialized the network parameters with the weights ResNet50 learned on ImageNet, and froze all but the last 15 layers. During training, we performed fine tuning on the model, and optimal weights were learned for the last 15 layers. The UHN CK7 dataset was separated into a training, testing, and validation dataset, with dataset proportions of 0.8, 0.1, and 0.1 respectively. The top activation layer has a single neuron for binary classification, and has a L2 kernel regularizer with a regularization factor of 0.001. We used the Adam optimizer, with a decaying learning rate starting with a learning rate of 0.0005 with the learning rate halved every 254,000 steps of model training. The loss function used was categorical cross entropy with weights of 0.74 for invasive, 1.24 for noninvasive, and 1.20 for undefined. The model was trained using Tensorflow equipped with an early stopping callback to stop training once no further improvement on the validation set was observed. The model was trained for 10 epochs with a batch size of 128. We also used two Tesla P100 GPUs from NVIDIA and utilized Tensorflow's distributed training library to evenly distribute the training over both GPUs.

2.4 Unsupervised Model Workflow

We used a convolutional variational autoencoder (CVAE) for the unsupervised model pipeline. The architecture of the encoder is as follows: an input layer, followed by 4 convolutional layers, then a flattening layer with two dense layers from it that represent the mean and log variance of the statistical distribution each point will be encoded as. The architecture of the decoder is as follows: an input layer, followed by a reshape layer, then 4 convolutional transpose layers, then a final sigmoid activation layer of the same input image size to the autoencoder, but grayscale (1 channel). Each convolutional layer is made up of a 2D convolutional layer, an activation layer using the ReLU function, and a batch normalization layer. The convolutional transpose layers follow the same architecture, except a 2D convolutional transpose layer instead of the convolutional layer. From the first to fourth layer, the convolutional filters are of size 32, 64, 64, and 64, and the convolutional kernels are all of size 3. The strides for the respective convolutional layers are 1, 2, 2, and 1, and the latent space dimension is 200. A learning rate of $1e^{-6}$ was used with the Adam optimizer. We used mean squared error for the reconstruction loss function, and the KL-Divergence as a regularization term for the loss function. The model was trained using Tensorflow equipped with an early stopping callback to stop training once the validation loss improved no further. The model was trained for 50 epochs with a batch size of 32, and we also used 1 Tesla P100 GPU for training.

3. Results

3.1 Supervised Classifier Model

As of right now, the supervised classifier model completes the WSI classification task of tumour and normal tissue with a test area under the receiver operator curve (AUC-ROC) of 0.52 (Fig. 4). Chance AUC-ROC is 0.50, so the model shows little evidence of learning the training data. The model displays similar performance on the UHN dataset. As shown by the loss, precision-recall curve, precision, and recall metrics, the model is not learning the training dataset (Fig. 5).

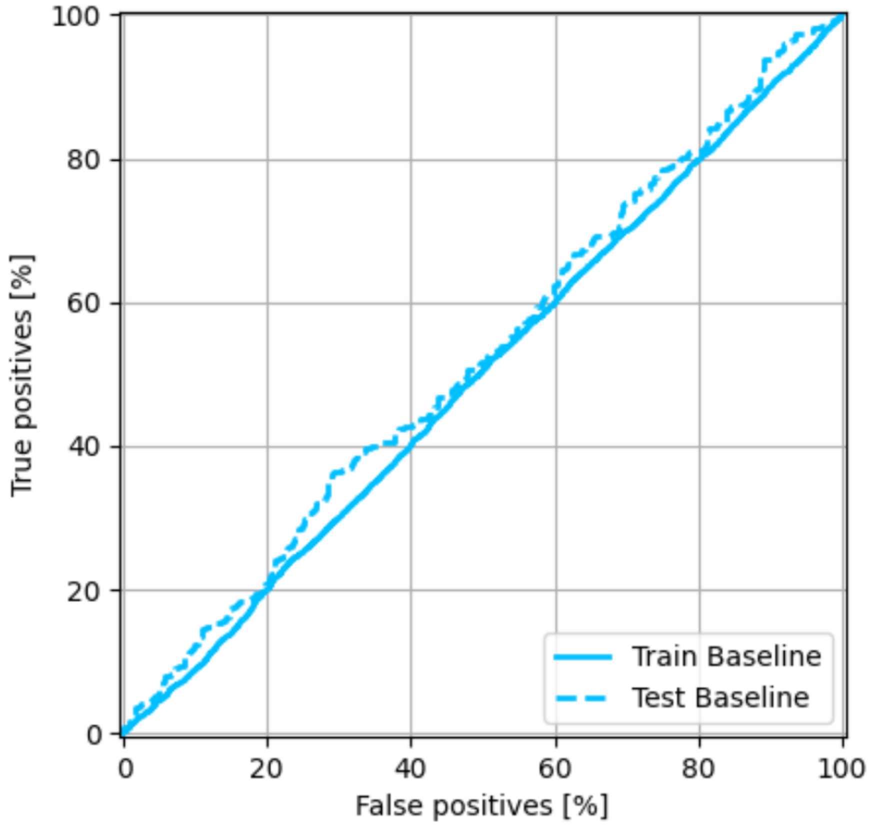


Figure 4. Train and test AUC-ROC for the classifier model on the TCGA-PAAD dataset

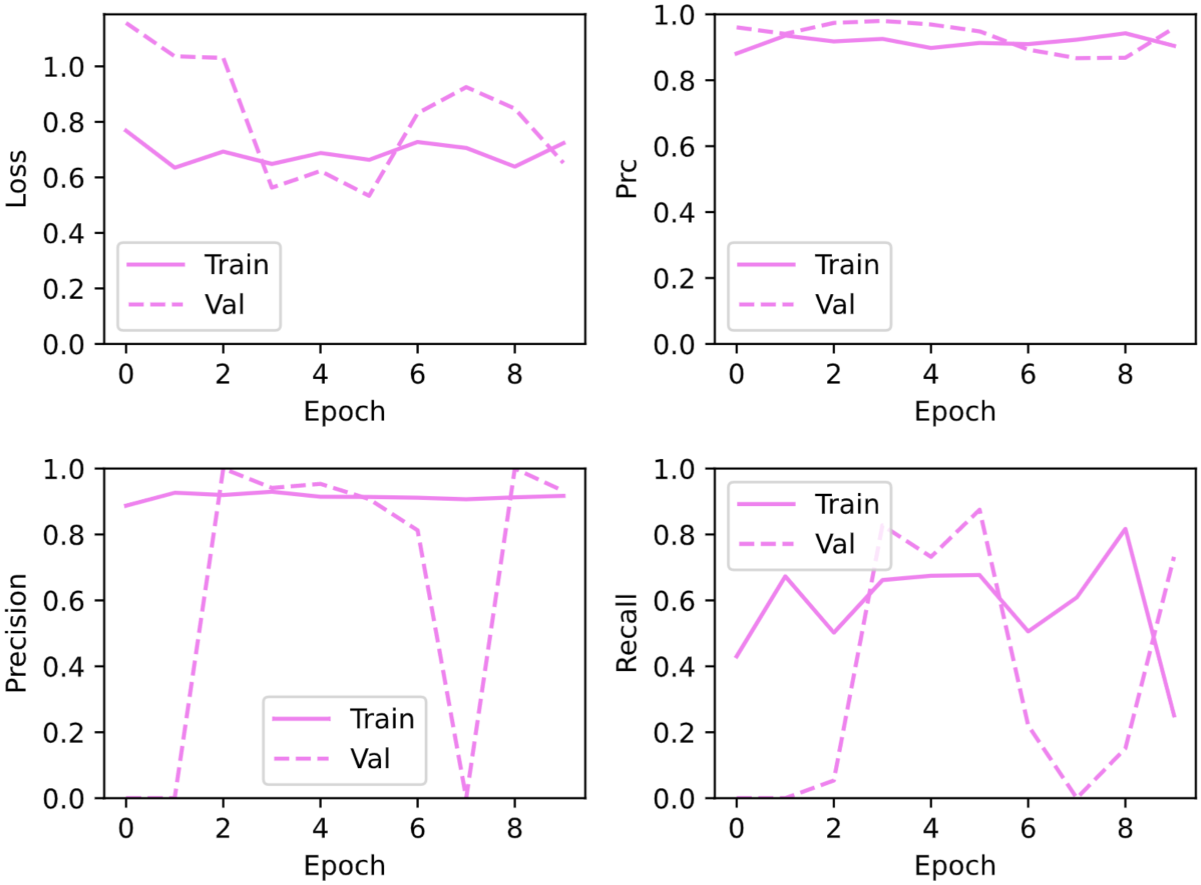


Figure 5. Training and validation metrics for the supervised classifier on the TCGA-PAAD dataset

3.2 Convolutional Variational Autoencoder

The variational autoencoder shows strong evidence of learning, achieving a validation reconstruction error (mean squared error) of 0.07 after 50 epochs with no early stopping (Fig. 6). In addition to the performance of the reconstruction error, the KL-Divergence term achieved a validation loss score of 3.60 (Fig. 7). The overall loss function achieved a score of 66.81 (Fig. 8).

Alongside the model performance statistics, a visual representation of a few sample images was generated (Fig. X). The model was able to capture very large scale image features, mainly those associated with colour or brightness, but was unable to capture features that were more specific in any way. Currently, it has been noted that there is no difference in model performance between images that have a CK7 stain and images that have a H&E stain.

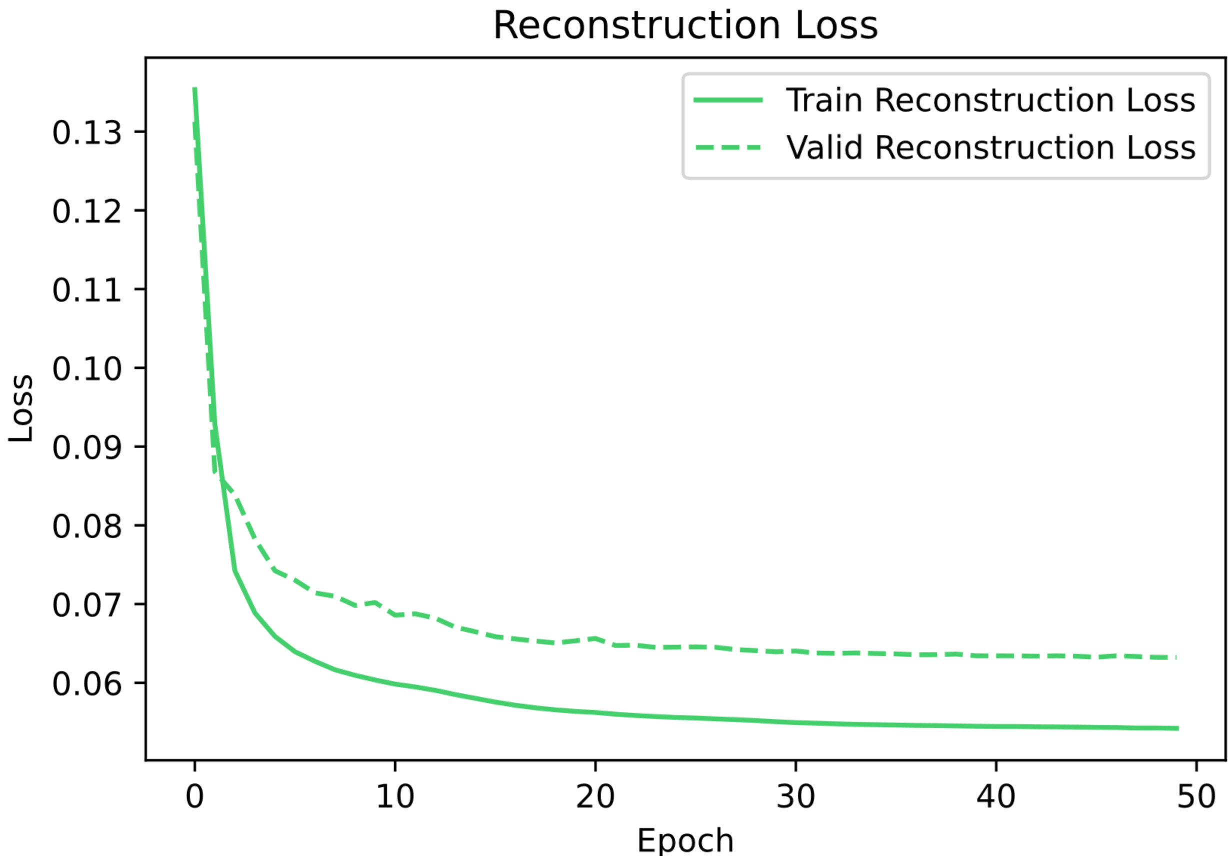


Figure 6. CVAE reconstruction loss over time

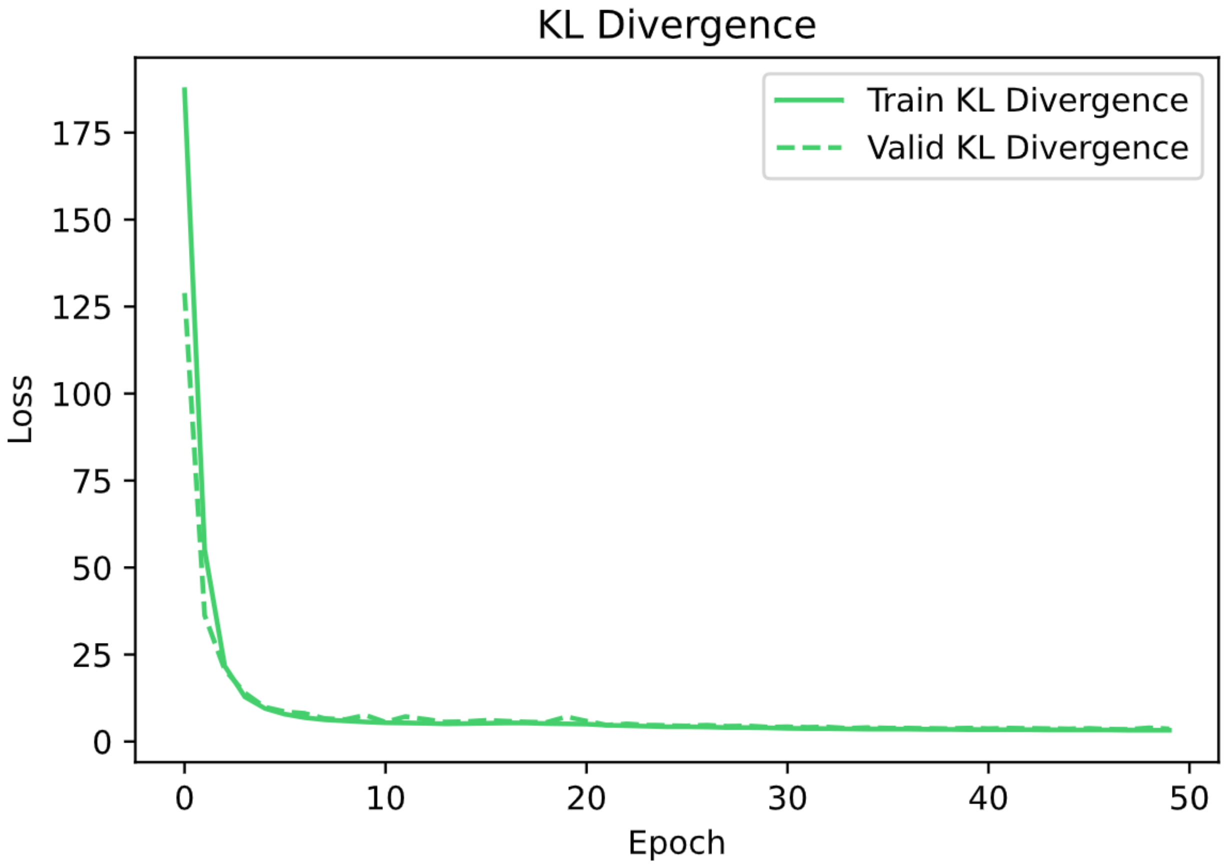


Figure 7. CVAE KL-Divergence over time

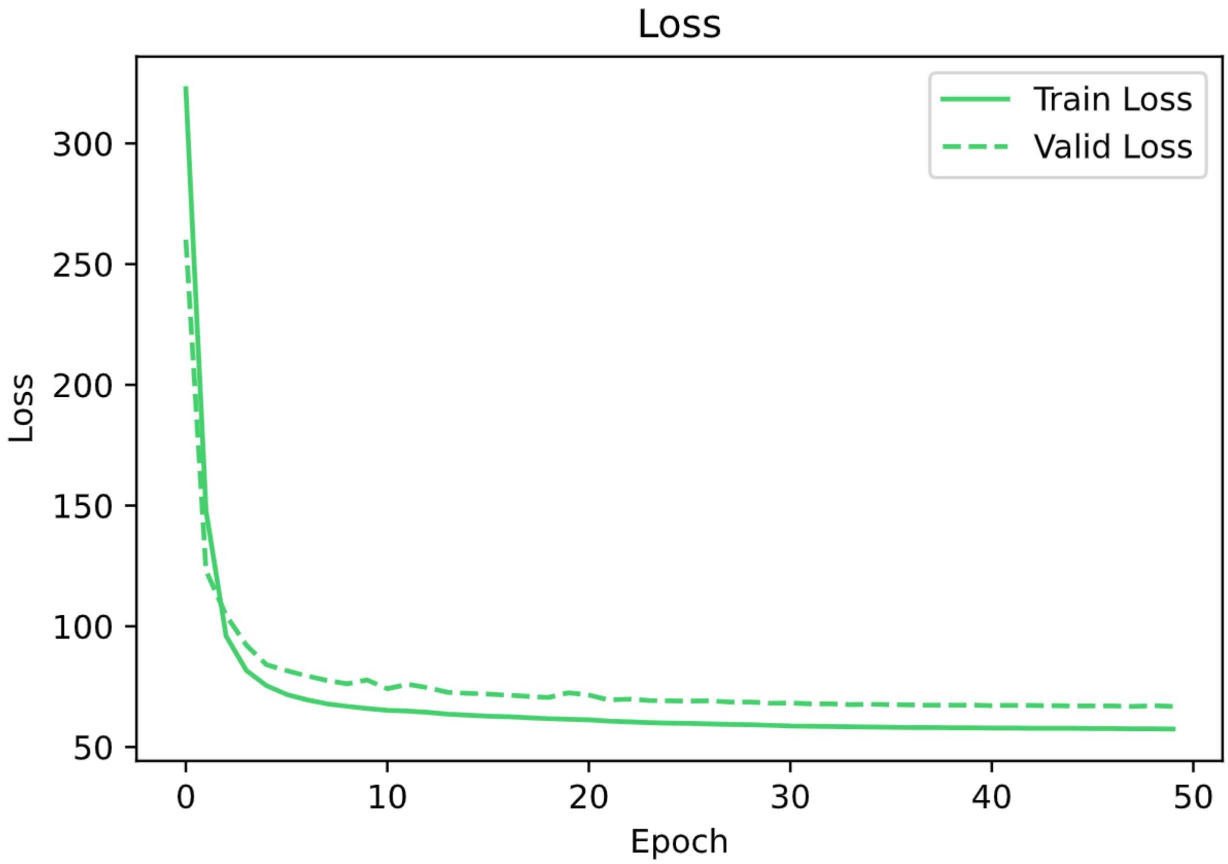


Figure 8. CVAE total loss over time

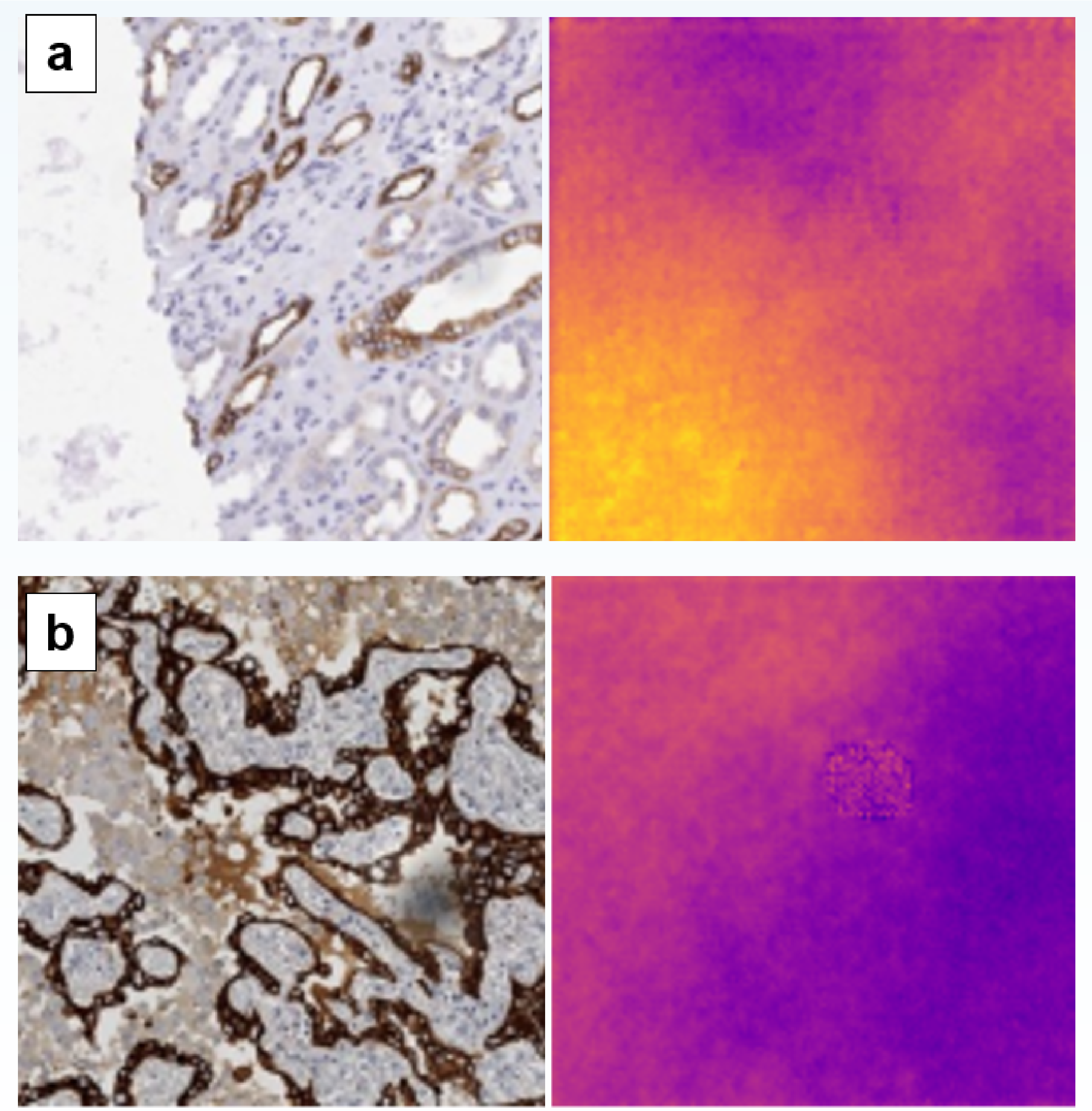


Figure 9. A sample of slide image crops (left) stained with CK7 stain, and the CVAE's reconstruction of the images (right). Large spatial features such as those in (a) are captured, but intricate details such as those in (b) are not.

4. Discussion

The supervised classifier showed no evidence of learning on either a test dataset from TCGA or the private whole-slide image dataset from UHN. We hypothesize that the amount of noise in the dataset is too great, and the image-to-noise ratio is thus too high.

The unsupervised workflow involving feature extraction learned large scale features of images. This demonstrates evidence that unsupervised deep learning models can learn image features of slide images. Our current model is not at a level with which there is confidence in features extracted by the model for analysis by senior researchers and expert pathologists. We hypothesize that the reason for such poor specificity in reconstruction is because the bottleneck dimension is 200, and the input dimension is 50,176. Increasing bottleneck dimension will most likely increase the specificity of the CVAE's reconstruction.

5. Next Steps

Next steps for the supervised model are to ask expert pathologists to segment the tumour images to provide more accurate labels. With tumour segmentation or a bounding box drawn, we will be able to assign tiles labels on a per-tile basis, rather than assign the tile with the slide image label. Past work in [4] shows that this method is successful, and is the current standard of supervised classification models in the field.

Next steps for the unsupervised workflow are to cluster the latent representations of the slide tiles into 2 or 3 informative clusters. Our hope is that the latent representations of similar status tiles (invasive vs. non-invasive) are similar, and thus the clustering will cluster images of similar status together, illuminating features in slide tiles that are similar. The pipeline for this has already been made and the results await the action of a collaborator.

6. References

1. Campanella, Gabriele, et al. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images." Nature medicine 25.8 (2019): 1301-1309.
2. Coudray, Nicolas, et al. "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning." Nature medicine 24.10 (2018): 1559-1567.
3. Saednia, Khadijeh, William T. Tran, and Ali Sadeghi-Naini. "Automatic characterization of breast lesions using multi-scale attention-guided deep learning of digital histology images." Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 11.1 (2023): 103-111.
4. Wei, Jason W., et al. "Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks." Scientific reports 9.1 (2019): 3358.
5. Yang, Ellen, et al. "Reproducibility in assessment of 'invasion' in lung adenocarcinoma with lepidic component: an interobserver concordance study with cytokeratin 7 stain." United States and Canadian Academy of Pathology Annual Meeting. 2022.
6. Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.