Bachelor Thesis Project Final Report

Debanjan Mondal

June 2020

Contents

1 Introduction

Metastasis is the phenomenon, in which cancer cells break away from the original (primary) tumor, travel through the blood or lymph system, and form a new tumor in other organs or tissues of the body.

Applying machine learning approaches in tumor diagnosis is an active research field. Detecting metastasizing tumor tissues with human eyes is tedious and often prone to errors.

Our work attempts to investigate the case, which involves identifying metastasizing and non-metastasizing tissue images.We also aim to single out the regions containing metastasizing cells.

2 Related Work

While dealing with the problem we found no other significant approaches to this specific task. But in the process we used some other approaches which have shown success in medical imaging.Our approach is based on Attention based Multiple Instance Learning [1]. They used an attention based approach to solve the Multiple Instance Learning problem.

2.1 Multiple Instance Learning

In the classical (binary) supervised learning problem one aims at finding a model that predicts a value of a target variable, $y \in \{0, 1\}$, for a given instance, $x \in R^D$. In the case of the MIL problem, however, instead of a single instance there is a bag of instances, $X = x_1, ..., x_K$. It consists of three steps:

- a transformation of instances to a low-dimensional embedding
- a permutation-invariant (symmetric) aggregation function
- a final transformation to the bag probability.

2.2 Attention Based Multiple Instance Learning

They proposed to use a weighted average of instances (low-dimensional embeddings) where weights are determined by a neural network. Additionally, the weights must sum to 1 invariant to the size of a bag. Let $H = h_1, ..., h_K$ be a bag of K embeddings, then they proposed the following MIL aggregating step:

$$\mathbf{z} = \sum_{k=1}^{K} a_k h_k$$

$$\mathbf{a}_k = \frac{exp\{\mathbf{w}^T tanh(\mathbf{V}\mathbf{h}_k^T)\}}{\sum_{j=1}^{K} exp\{\mathbf{w}^T tanh(\mathbf{V}\mathbf{h}_j^T)\}}$$

where $w \in \mathbb{R}^{L \times 1}$ and $V \in \mathbb{R}^{L \times M}$ are parameters. The hyperbolic tangent both positive and negative values for proper gradient flow.

2.3 Gated Attention

Furthermore, the tanh($\cdot$) non-linearity could be inefficient to learn complex relations, since tanh(x) is approximately linear for $x \in [1, 1]$, which could limit the final expressiveness of learned relations among instances. Therefore, they proposed an additional gating mechanism [2] together with tanh($\cdot$) non-linearity:

$$\mathbf{a}_k = \frac{exp\{w^T tanh(\mathbf{V}\mathbf{h}_k^T) \odot sigm(]\mathbf{U}\mathbf{h}_k^T\}}{\sum_{j=1}^{K} exp\{w^T tanh(\mathbf{V}\mathbf{h}_j^T) \odot sigm(\mathbf{U}\mathbf{h}_j^T)\}}$$

where $U \in \mathbb{R}^{L \times M}$ is a parameter

Fold No.	N0M0	Rest
1	82	62
2	83	60
3	81	61
Overall	**246**	**183**

Table 1: 3 fold cross-validation distribution

3 Dataset

In the previous semester I tried an approach to classify some slide images provided by **John Hopkins Dataset**. However, we needed to reproduce the results on other datasets. We decided to apply our model in **TCGA Basal** slide images. **TNM** labels were available along with the slide images. M0 meant no distant metastasis, whereas M1 or above implied the tumor tissue had distant metastasis. N0 But the images available were too big to be fed to our model. N0 meant no affected regional lymph nodes, N1 implied 1-3, and N2 indicated 4 or more regional lymph nodes. Initially we had 99 images, later I filtered out 90 valid images. Out of these 55 were N0M0 and the rest had other labels.

3.1 Data Annotation

The images available were too big to be fed to our model. So we had to manually crop interesting regions. We annotated IDC(**Invasive Ductal Carcinoma**) , DCIS(**Ductal Carcinoma in situ**) and Benign regions of these slide images. However, while training we used only the IDC regions. The annotated regions were cropped out programmatically. In most cases, multiple IDC regions were available from one slide images. We treated each such region as a separate entity while training. So now we had 246 images with label N0M0 and 183 images with other labels.

3.2 Cross Validation

We chose to do a 3-fold cross validation. TCGA has data from different institutions. Say the name of a image is TCGA-**A1**-A0SP-01Z-00-DX1.20D689C6-EFA5-4694-BE76-24475A89ACC0.svs. The highlighted letters denote the institution. Data obtained from a particular institution were put to one set only. Also the ratio between two labels were kept almost same across the sets. I wrote a program to check all permutations and find one that satisfies the above two criteria.

4 Approach 1 : Resnet Based Model

We started our task with a resnet[3] based approach with slight modifications. This was a preliminary approach and was aimed knowing whether something can be learnt from the dataset at all.

4.1 Model Architecture

A pretrained Resnet34 model was initialized. The final fully connected layer was replaced by a fully connected layer with 2 output channels.

4.2 Data Preparation

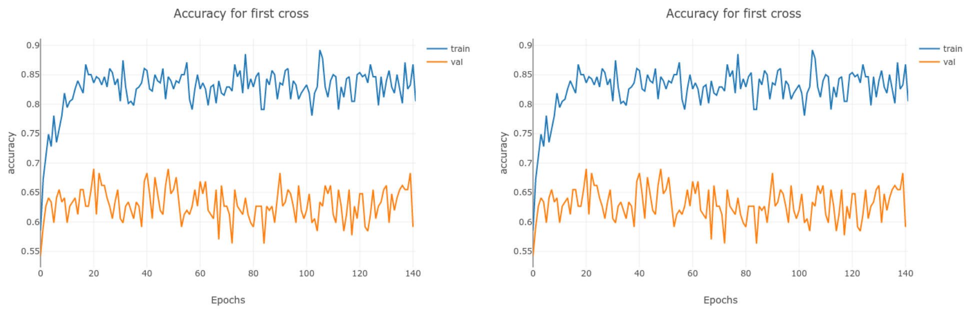
Due to limited size of the dataset, I used runtime augmentations while training. Random 450 × 450 patches were cropped from the tissue images. Below this size, most of the random patches were having too many black pixels. (All the cropped regions were not rectangular, so some cropped images had black pixels). I used three types of augmentations :

- Random Rotation upto 90 degrees
- Random Horizontal Flip
- Color Jitters

Patches which had more than 20 percent black pixels were discarded in the process.

4.3 Training

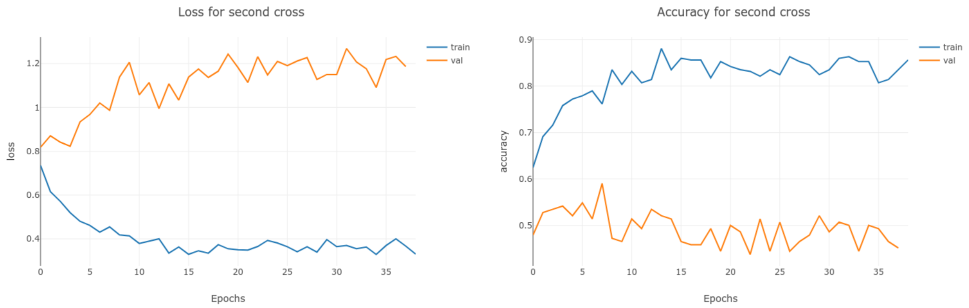
An **Adam** optimizer with learning rate 10^{-5} was used. A step scheduler with step size 10 and decay value 0.1 was employed.Weighted CrossEntropyLoss(with weights being inverse of class size) was optimized to train the model.



(a) Epoch vs Loss in first cross

(b) Epoch vs Accuracy in first cross

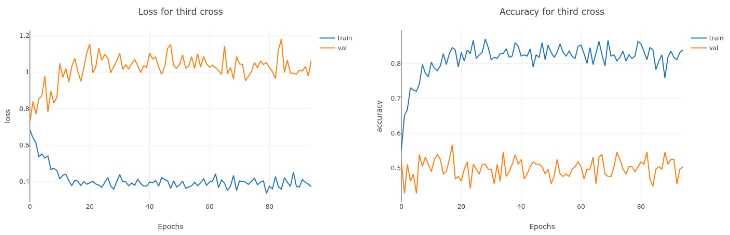
Figure 1: Loss and Accuracy for first cross



(a) Epoch vs Loss in second cross

(b) Epoch vs Accuracy in second cross

Figure 2: Loss and Accuracy for second cross



(a) Epoch vs Loss in Run3

(b) Epoch vs Accuracy in Run3

Figure 3: Loss and Accuracy for Third Cross

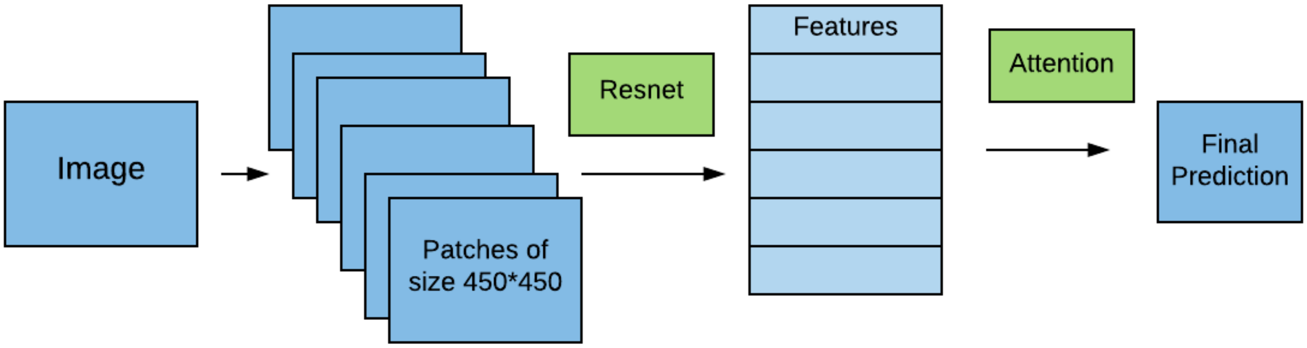


Figure 4: Model Architecture

4.4 Results

Due to small size of the dataset, I also used test time augmentations.I cropped random patches of size 450 × 450 from the slide images, added color jitters and used them in the test dataset. However the results were not satisfactory and widely varied across the three crosses.

4.5 Limitations

We used random patches from the image for the classification , which doesn't capture the essence of the whole image.Also there is a fundamental problem with this approach. A tissue image being classified as metastasizing doesn't imply all its patches contain metastasizing cells. Also we aimed at identifying regions of interest, which is clearly not possible in this approach.

5 Approach 2 : Attention + Resnet

This approach is based on the paper Attention Based Multiple Instance Learning**??**.We modified the architecture to some extent based on our problem.

5.1 Model Architecture

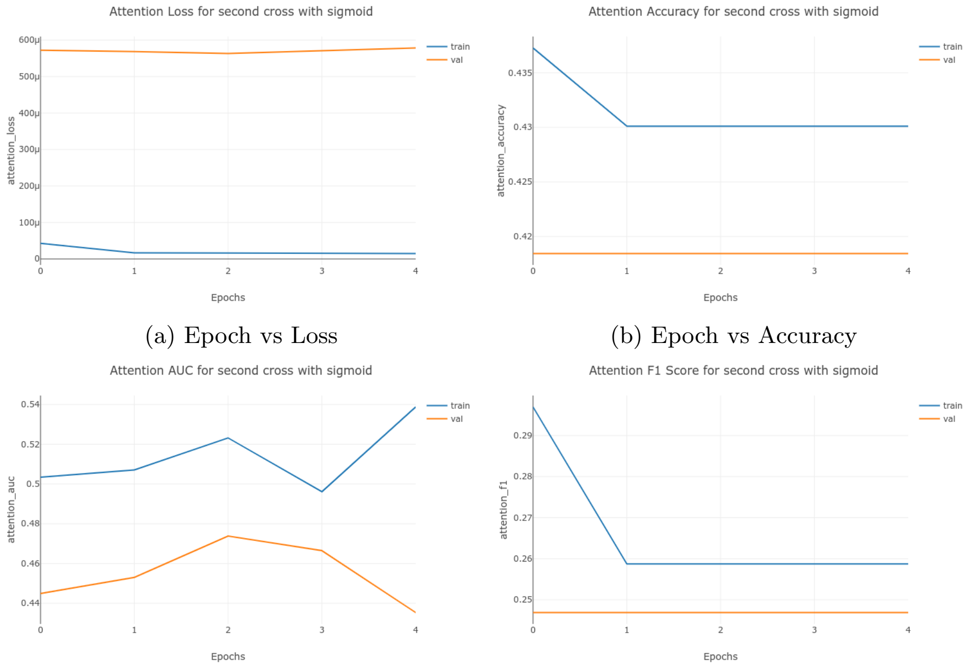
We collected all 450×450 patches from each image and treated each image as a bag of instances. Then we used pretrained Resnet34(trained in Approach 1 minus the final fully connected layer) to extract features from the patches. The Resnet weights were frozen during this training. The attention weights are calculated by passing the features through two fully connected layers as described in 2.2. In my implementation, the first linear layer has 128 output channels, followed by a tanh activation. The second fully connected layer has 1 output channel.

After calculating the weights the weighted feature(referenced as z in 2.2 is calculated. It is passed through a fully connected layer having 1 output channel and a sigmoid activation to obtain the probability of belonging to the positive class. In my implementation, I treated the metastasizing tissue containing class as positive class. An overview of the architecture can be seen in Figure 4

5.2 Data Augmentation and Training

Due to limited dataset, I added random colorjitters to each image and created 5 augmented images from each image. The dataset size was now 6 times the original size. An adam optimizer with learning rate 10^{-4} was used.A step scheduler with step size 10 and decay value 0.1 was employed. A weighted CrossEntropy loss function was optimized.

5.3 Results



(a) Epoch vs Loss

(b) Epoch vs Accuracy

(c) Epoch vs AUC

(d) Epoch vs F1 Score

Figure 5: Results with Sigmoid Approach

However this approach also didn't yield good results. It was not showing improvement over the epochs and was classifying all the images as non-N0M0.

6 Approach 3: ResNet + Attention Variation

This approach is basically a slight tweak of 5.

6.1 Model Architecture

Instead of the final fully connected layer with one output channel followed by a sigmoid, I used a fully connected layer with 2 output channels. The sigmoid layer was omitted. The outout obtained was a vector of size 2.

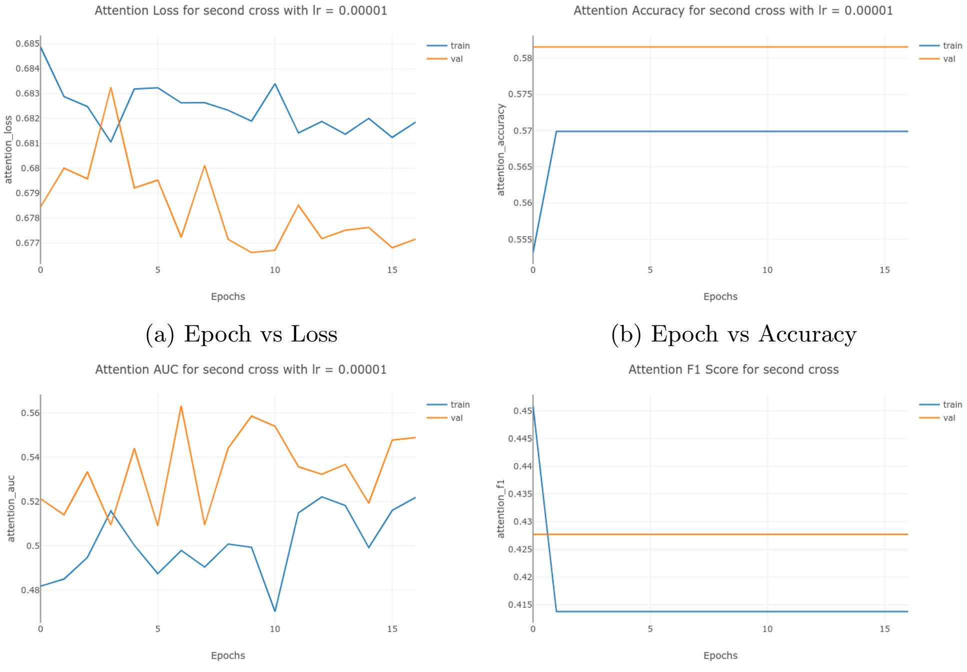
6.2 Data Augmentation and Training

The data augmentation and loss function used were same as 5.

6.3 Results

While testing also, I added random color jitters and random horizontal flip to each image and created 5 augmented images from each image. The test dataset size too was now 6 times the original size. An adam optimizer was used. A step scheduler with step size 10 and decay value 0.1 was employed. A weighted CrossEntropy loss function was optimized. With this architecture I ran some experiments, by tweaking the learning rate and using gated and non-gated attentions.

6.3.1 With Learning Rate = 0.00001 and Gated Attention



(a) Epoch vs Loss

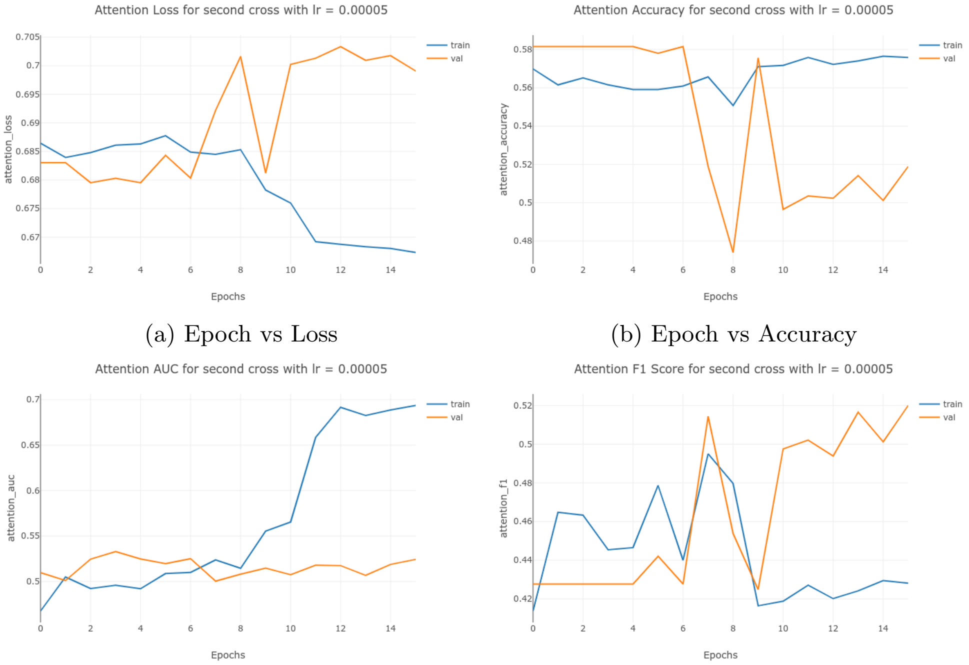
(b) Epoch vs Accuracy

(c) Epoch vs AUC

(d) Epoch vs F1 Score

Figure 6: Results with Learning Rate 0.00001 and Gated Attention

6.3.2 With Learning Rate = 0.00005 and Gated Attention



(a) Epoch vs Loss

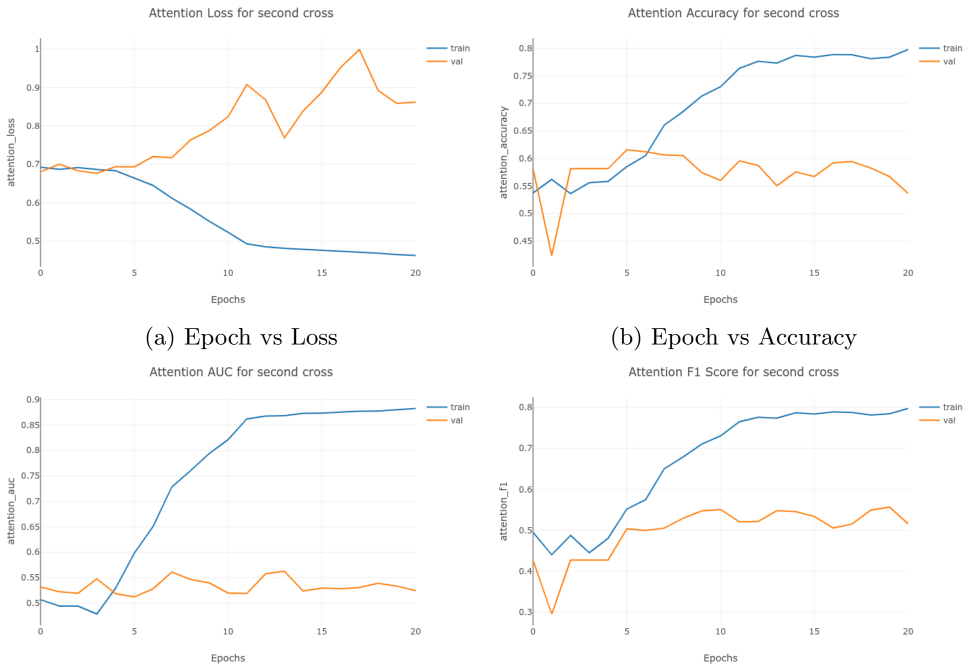
(b) Epoch vs Accuracy

(c) Epoch vs AUC

(d) Epoch vs F1 Score

Figure 7: Results with Learning Rate 0.00005 and Gated Attention

6.3.3 With Learning Rate = 0.0001 and Gated Attention



(a) Epoch vs Loss

(b) Epoch vs Accuracy

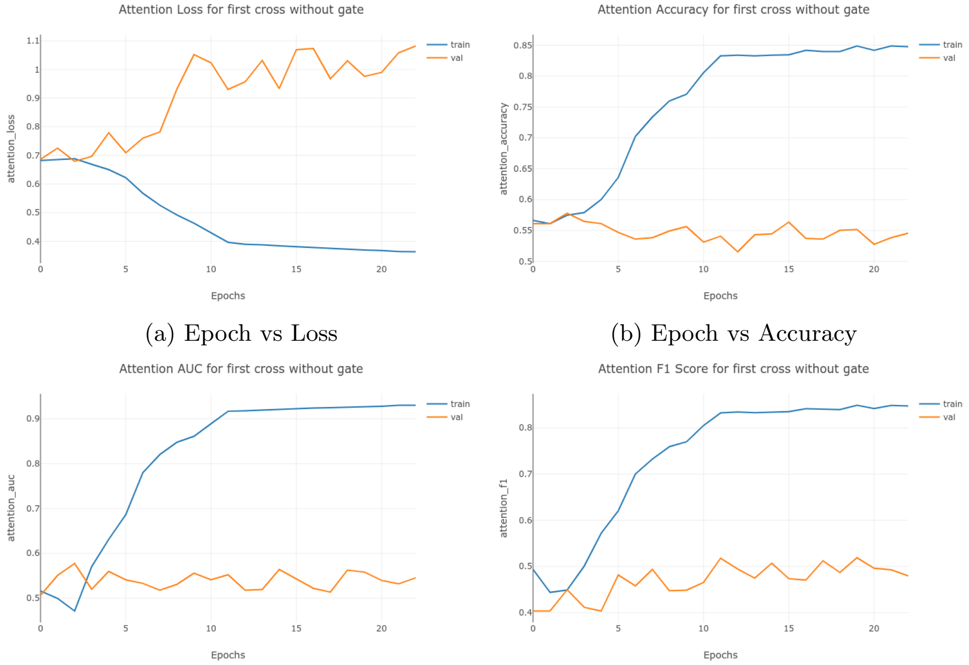
(c) Epoch vs AUC

(d) Epoch vs F1 Score

Figure 8: Results with Learning Rate 0.0001 and Gated Attention

6.3.4 With Learning Rate = 0.0001 and Without Gated Attention

This architecture gave the best result among all. I did a 3 fold cross validation with the folds 1 mentioned before. The Results 2 reported are based on this architecture.



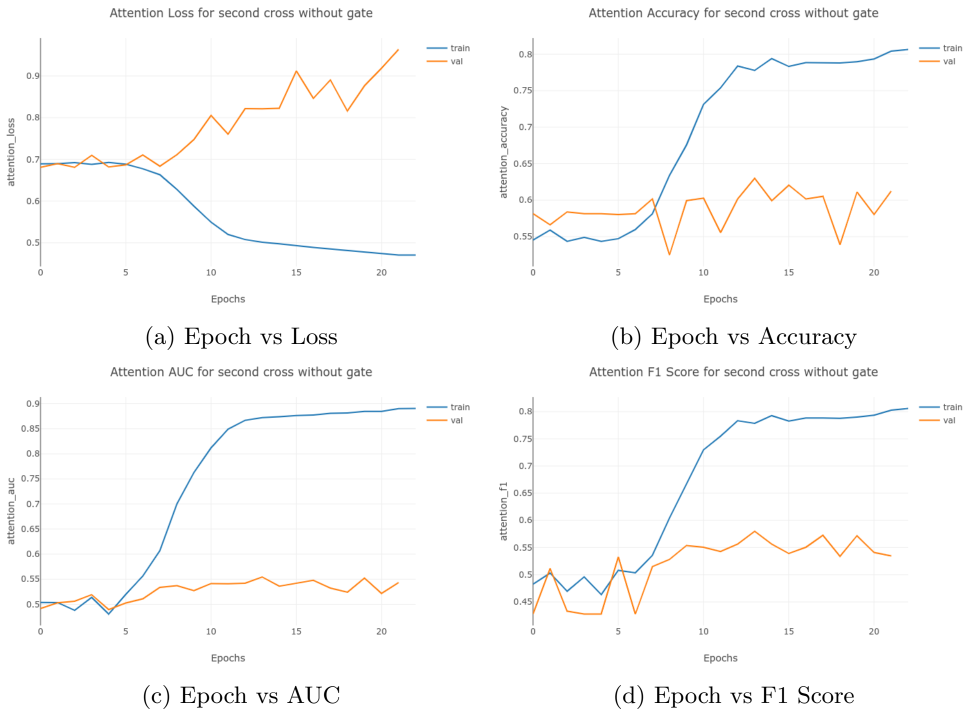
(a) Epoch vs Loss

(b) Epoch vs Accuracy

(c) Epoch vs AUC

(d) Epoch vs F1 Score

Figure 9: Results with Learning Rate 0.0001 and Without Gated Attention First Cross



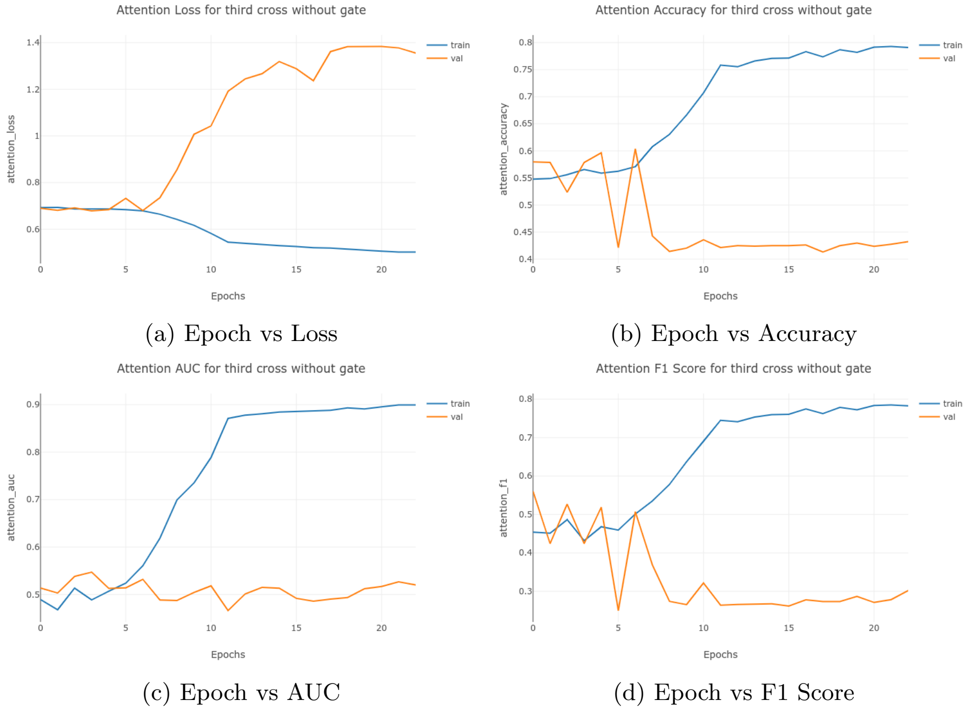
(a) Epoch vs Loss

(b) Epoch vs Accuracy

(c) Epoch vs AUC

(d) Epoch vs F1 Score

Figure 10: Results with Learning Rate 0.0001 and Without Gated Attention Second Cross



(a) Epoch vs Loss

(b) Epoch vs Accuracy

(c) Epoch vs AUC

(d) Epoch vs F1 Score

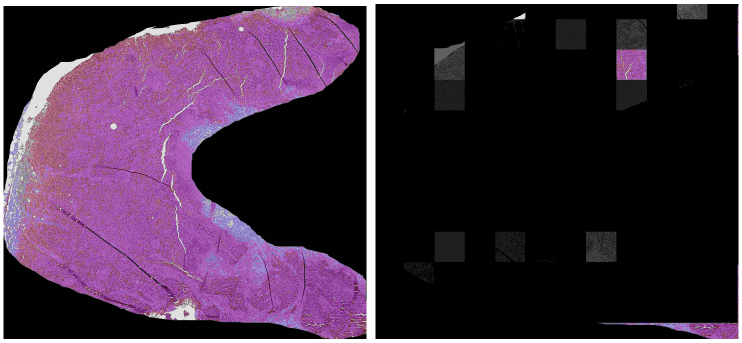
Figure 11: Results with Learning Rate 0.0001 and Without Gated Attention Third Cross

	Validation Accuracy	AUC Score	Patient Level AUC
First Cross	0.5779	0.5772	0.6296
Second Cross	0.6300	0.5545	0.5777
Third Cross	0.5785	0.5472	0.5950
Overall	**0.5945** ± 0.0244	**0.5596** ± 0.0128	**0.6008** ± 0.0216

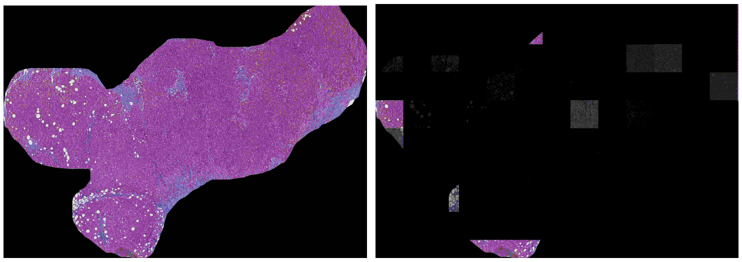
Table 2: 3 fold cross-validation results

6.4 Region Of Interests

While testing we also collected the attention weights of the patches corresponding to each image. We multiplied the original pixel values with the weights to get regions of interest.



(a) Actual Image (b) Regions Of Interest



(a) Actual Image (b) Regions Of Interest

7 Conclusion

Clearly better results were observed for ResNet + Attention model. Also we were able to identify regions containing metastasizing cells in this approach. This can be very convenient in medical domain, and will help in detecting harmful areas to apply treatment.

8 Limitations

I couldn't achieve high accuracy and AUC scores. This may be attributed to the small data available. This kind of problems require complex models , and the data provided were not sufficient to train such models. Pixel wise annotations were not available. So we didn't know which regions contain metastasizing cells.Also there was a class imbalance, smaller number of metastasizing tissue images were available, which was clearly not conducive for our approach to this problem.

References

[1] *Attention-based Deep Multiple Instance Learning*, Maximilian Ilse, Jakub M. Tomczak, Max Welling , ICML 2018

[2] *Language modeling with gated convolutional networks*, Dauphin, Yann N, Fan, Angela, Auli, Michael, and Grangier, https://arxiv.org/abs/1612.08083

[3] *Deep Residual Learning for Image Recognition*, Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun https://arxiv.org/abs/1512.03385