GIP Lab Project – 234329
Final Report

Weakly Supervised Learning Methods for Analysis of Histological Whole Slide Images

Tal Neoran
Supervised by Gil Shamai
Semesters: Spring 2021 – Winter 2021/22

1. Introduction

In the field of histopathology, slices of tissue samples are typically stained and examined under a microscope for diagnosis of cancer and other diseases, and for analysis of different properties of the diseases. In recent years, the use of whole slide imaging technology for digital scanning of glass slides has rapidly increased, allowing the employment of artificial intelligence (AI) methods to improve the analysis and prediction abilities in computational histopathology. Scanning of a glass slide results in a scan known as a whole slide image (WSI), a gigapixel image of a tissue sample constructed from multiple different magnification resolutions. Performing analysis of WSIs using deep learning algorithms poses multiple challenges, for example due to the large size of the images, and due to the lack of availability of local annotations.

In this project, we implemented and assessed the potential of methods in weakly supervised and self-supervised deep learning to perform analysis and predict clinical outcomes from Hematoxylin and Eosin (H&E) stained WSIs of patients with breast cancer. We proposed and evaluated ways to deal with the challenges that arise when working with WSIs and applied our methods to WSI dataset for classification of hormone receptor status information.

Throughout the report, we will provide a background of related work in the field of computational histopathology, describe the methods we implemented, discuss the experiments performed and their results, and finally suggest possible limitations in the suggested solutions and propose improvements or possible continuation directions.

1.1 Background and Related Work

Previous works have shown the potential of using AI to predict clinical information from histopathology images [1], [2], [3]. Typically, a feature extractor CNN is learned in a supervised fashion for the classification of the desired attribute of the cancer, such as hormone receptor status information of ER, PR, and Her2. However, this supervision is only possible when local labels are available, i.e., a label for each image that is passed through the network. In the case of WSIs, usually only a single slide-level label is available, while the WSI cannot be processed in its entirety by the feature extractor due to its size. In addition, since the adoption of digital scanning technology in pathology is still ongoing, there is often a lack of labeled data in general for many different clinical tasks, while the existing data also varies between sources and may be affected by external factors.

In recent years, many different architectures were proposed to deal with the challenges of WSIs. Examples for suggested methods can be found in [2], [3], [4], [5]. Researches employ various normalization and augmentation methods to deal with variability in the data and suggest weakly supervised or unsupervised components such as clustering to improve results

on WSI classification tasks. Other clinical tasks such as cancer survival prediction have also been of interest in computational histopathology [6], [7], [8].

During the first part of the project, we implemented a method of attention-based multiple instance learning (MIL) which will be expanded on in the following section. The MIL formulation has recently become a common approach to the WSI setting [9], [10]. As will be explained, when applying MIL to our problem, an aggregation mechanism for instance features or scores is required. Researchers in the field suggest different aggregation mechanisms, such as pooling or attention-based aggregations, which serve as the inspiration for this part. Solving the problem of feature or prediction aggregation in WSIs has been a focus of many recent works. It is an important step towards achieving slide-level information extraction at a pathologist's level. The attention mechanism and the MIL formulation are both popular approaches to dealing with this issue, and we will reflect on the results of them in this project to suggest future improvements.

During the second part of the project, we examined the potential of a self-supervised learning (SSL) algorithm, using a contrastive learning approach. The goal of SSL in our case is to create a generic feature extractor for histopathology images, that can be used for multiple different downstream tasks, in an unsupervised manner. SSL is a rapidly growing field of machine learning, in which supervised pre-text tasks are created to allow training with unlabeled data. We focused on self-supervised contrastive learning approaches, where positive and negative sample pairs are chosen, and the pretext-task is defined as learning an embedding space in which positive samples are close to each other, while negative samples are far away. Relevant architectures in self-supervised contrastive learning include SimCLR [11] and MoCo [12], [13]. The application of SSL to histopathology is also an active research field and shows promising results in dealing with the challenges of WSIs [14], [15].

2. Methods

A general pipeline for WSI processing and analysis typically contains the following stages: Foreground segmentation, patch extraction, patch feature extraction, and patch feature/classification aggregation to the slide level. We describe our architecture and algorithm choices for theses stages. For the feature extraction and classification tasks, our work in the project can be divided into two parts, where we implemented and evaluated two different approaches to deep learning for analysis of histopathology WSIs, each suggested to deal with different challenges of the input data.

2.1 Preprocessing

An average WSI contains an order of $100{,}000 \times 100{,}000$ pixels. In addition, large portions of the image are background areas that contain no tissue information. To apply common deep learning algorithms to WSI data, some preprocessing must be done.

The first stage of preprocessing is foreground segmentation, where a segmentation algorithm is applied on the thumbnail image of the slide to find the areas containing tissue information and discard the background. This stage is done by thresholding the image and simply discarding the areas where the pixel values are under or above the threshold, although the choice of the threshold can be dynamic based on the image statistics, such as in Otsu's method. We tried a few different options and parameters for thresholding, but small differences in the segmentation procedure are insignificant when compared to the scale of a WSI, and a few background patches should not affect the performance.

The second stage is patch extraction, where a grid is constructed over the segmented foreground of the WSI, and image patches at a resolution suitable for a convolutional neural network (CNN) are extracted. We constructed a grid of 256×256 patches trivially, and extracted patches from the coordinates in the grid, allowing small random variations in the exact coordinate choice for the patches as a form of regularization, and with the goal of using as much of the data as possible. The patch extraction is done dynamically during training, and a collection of random patches is chosen from every slide in the dataset. A single slide can contain hundreds or thousands of extracted patches, although usually only a small portion are used in a single training epoch.

There are further considerations when doing preprocessing of WSI data, but for the sake of the project we chose to use a dataset that has already been examined and reviewed, and as such there are less issues with artifacts in the images and unusable samples. Despite this, a large part of the work involved dealing with the technical challenges of WSIs and handling the data.

2.2 Attention Multiple Instance Learning

In the first part of the project, the method we implemented and evaluated is an attention-based MIL method. It was originally proposed in [9] and applied to histopathology WSIs in [8] and [10]. MIL is a form of weakly supervised learning, where the goal is to classify a label that is available for a set of multiple samples, referred to as a bag, instead of a label for every single sample.

Formally, in the case of MIL for binary classification, training samples are a bag of instances, where M is the bag size, that are drawn i.i.d. from the instance distribution. Each instance x_i has a label $y_i \in \{0,1\}$ associated with it, and a bag label Y is defined for bag X such that it is positive if there exists at least one positive instance in the bag. This definition is equivalent to defining Y as the maximum label in the bag. Different approaches to MIL suggest ways to learn with bag labels. A trivial approach is simply assigning the input bag label to all samples in the bag and learning an instance level classifier using the assigned labels, followed by applying an aggregation mechanism to get a bag-level prediction. This trivial approach is also the baseline approach for WSIs that we compare our methods to.

We claim that the MIL problem formulation fits the WSI setting, in which a slide is a bag of patches (instances), and the slide-level label is chosen to be the label associated with a bag of patches extracted from the slide. This is the assumption of the method we implemented. We employ an instance level feature extractor, a ResNet-50, to extract a feature space representation of input patches. The output features are then aggregated using one of multiple aggregation mechanisms, and passed through a classifier MLP head to produce a bag-level prediction of the slide label.

There are several options for the choice of aggregation mechanism. In most previous works, pooling-based methods such as maximum or average pooling are used. These methods used a fixed aggregation and may not represent the correct combination of instance embeddings in a bag. Instead, we use an attention-based approach, where the bag aggregation is a weighted average:

$$z = \sum_{i=1}^{M} a_k h_k$$

Where h_k are the embeddings in the bag, and a_k are the normalized weights parametrized by a MLP attention head:

$$a_k = \frac{\exp\left(w^T \cdot tanh\left(V h_k^T\right)\right)}{\sum_{j=1}^{M} \exp\left(w^T \cdot tanh\left(V h_j^T\right)\right)}$$

Where w and V are learned parameters. This approach allows weights attributed to embeddings in a bag to be learned during training. It also improves the interpretability of the model by allowing us to view the weights assigned to patches in a slide and determining which patches contained significant information to the prediction of the slide label.

When applying attention MIL to a WSIs, there exists the consideration of constructing the bags. We implemented two main alternatives for this, a single-bag method, where a fixed number of patches from each slide is chosen at each epoch and passed through the network as a single bag, and a multi-bag method, where multiple bags can be extracted from the same slide. It is also possible to change the bag size dynamically, however this was not necessary in our case, and we found that it is generally enough to sample a relatively small number (order of 10) of patches using the single-bag method from each slide, and coverage of the slides is achieved by randomization of patch choices between epochs.

Another improvement to the MIL architecture that we implemented was the replacing of the conventional Cross-Entropy classification loss with Focal Loss, which attempts to deal with imbalanced data and improve stability. We evaluated the use of Focal Loss for our tasks.

2.3 Self-Supervised Learning

While MIL is a possible solution to the slide-level labels issue. We look to the domain of SSL as a solution for cases where not enough data is available, to overcome variability, and to make use of unlabeled WSIs or ones that are annotated for different clinical tasks.

We attempted to overcome further challenges of histopathology WSIs using ideas from contrastive self-supervised learning. The adoption of technology for digital analysis of tissue samples is still ongoing, and as such data and specifically labeled data are often scarce. In addition, there are evident differences in colors, image quality, and scan formats, between WSIs obtained from different medical centers or using different digital scanners. SSL has been shown to achieve state-of-the-art results on image processing tasks in the recent years, and as such we suggest this approach as a solution to these issues.

The application of SSL to our problem can be divided into 2 stages. First, the pre-training stage, where a feature extractor, which in our case is a ResNet-50, is trained on a dataset of WSIs, or a combination of such datasets, to perform a contrastive pre-text task that makes no use of the slide labels. The pre-training stage usually requires many training samples, but this is not an issue as they do not need to be labeled. We therefore combine multiple datasets of WSIs, potentially with different labels for multiple classification tasks, discard the labels, and train on patches extracted from the slides. Next, supervised classifier training is performed with the pre-trained feature extractor, on a relatively small dataset for the target downstream task.

When training the classifier on the downstream task, there are multiple options for how the pre-trained feature extractor is used. One possibility, which is the common choice in SSL research, is linear classifier fine-tuning, when the weights of the pre-trained feature extractor are frozen, and a linear classifier layer is trained with full supervision for the downstream task. Another option is using the weights simply as an initialization for the downstream training, and propagating gradient weight updates to the entire network. We provide a comparison between these options, as well as regular supervised training from scratch on the target dataset in the results section.

We implemented the MoCo-v2 contrastive SSL architecture [12], [13]. In this architecture, as well in other architectures such as SimCLR [11], the contrastive learning pretext-task is defined as follows. For each training sample (in our case, a patch from a WSI), two separate views are generated, using random augmentations on the image. The two views are then considered a positive sample pair. In contrast, the views of different samples are considered negative. Each training batch then contains an anchor sample (the first view of the input image), its positive pair (the second view), and K negative samples (views of other images in the batch). The network is trained using one of a variety of contrastive loss functions, to minimize the embedding distance between the original image to its positive second view, while maximizing the distance to negative samples.

In MoCo, encoded samples are stored in a dictionary managed as a queue and used as negative samples in future passes through the network the contrastive task. In addition, the contrastive loss function that is used is InfoNCE, which is given by:

$$\mathcal{L}_q = -\log \frac{\exp\left(q \cdot \frac{k_+}{\tau}\right)}{\sum_{i=0}^{K} \exp\left(k \cdot \frac{k_i}{\tau}\right)}$$

Where τ is a temperature hyper-parameter, q is an encoded sample regarded as a query for the dictionary, k_+ is the encoded positive sample, regarded as a positive key, and k_i are the encoded negative samples, also regarded as keys for the dictionary lookup. This loss function can be interpreted as the loss of a classifier with the goal of classifying the query as the positive key, as opposed to the negative keys.

To achieve this formulation in practice, the MoCo implementation is comprised of two identical CNNs, f_q and f_k, serving as the query and the key encoders respectively. The key encoder network is updated from the query encoder parameters using a momentum update: $\theta_k \leftarrow m\theta_k + (1-m)\theta_q$, where m is a hyper-parameter. The momentum update enables the encoded keys in the queue to remain relatively consistent for as long as they remain in the queue, due to the parameters of their encoder evolving in a smoother manner.

Further technical details of the architecture are described in the original papers. In our method, we used most of the practices suggested by the authors, and mainly worked on the adaptation

of the architecture to our tasks of interest. This includes changes in the data processing pipeline during training, variations of the backbone CNN model, different choices for the augmentations applied to input samples (which in the original paper were suggested for natural image classification), and further adjustments.

3. Experiments, Statistical Evaluation and Results

The MIL experiments were conducted on the HEROHE dataset. HEROHE is a dataset that was published as part of a digital pathology challenge. It contains 350 train and 160 test WSIs, annotated for binary classification at the slide-level of the expression of the HER2 breast cancer hormone receptor status. We trained our network architecture on the training set and evaluated it on the test set.

For measuring performance, we used many statistical metrics throughout the project. Since the problem is binary classification, the main metrics we used were patch classification accuracy, as well as the slide-level classification AUC value and the associated ROC curve. Depending on the experiment, we performed further statistical evaluations such as balanced accuracy, and Positive Predictive Value (PPV) and Negative Predictive Value (NPV) graphs.

We conducted experiments comparing different variations of the attention MIL architecture. Variations include the bag size, the number of bags extracted from each slide, the loss function, architecture hyperparameters, and more. We compared the variations of the architecture between themselves, as well as evaluated the performance of MIL when compared with the naïve supervised learning approach of attributing the slide-level labels to each patch in the slide. The following is a sample of some of the results from the experiments.

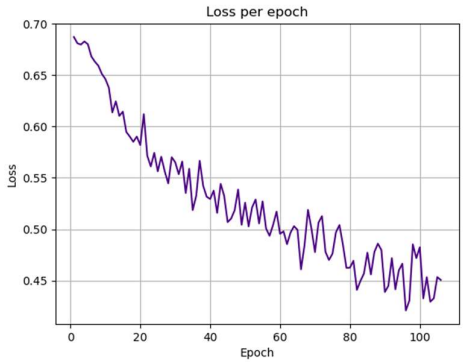

Figure 1: Attention MIL training loss convergence over 100 epochs.

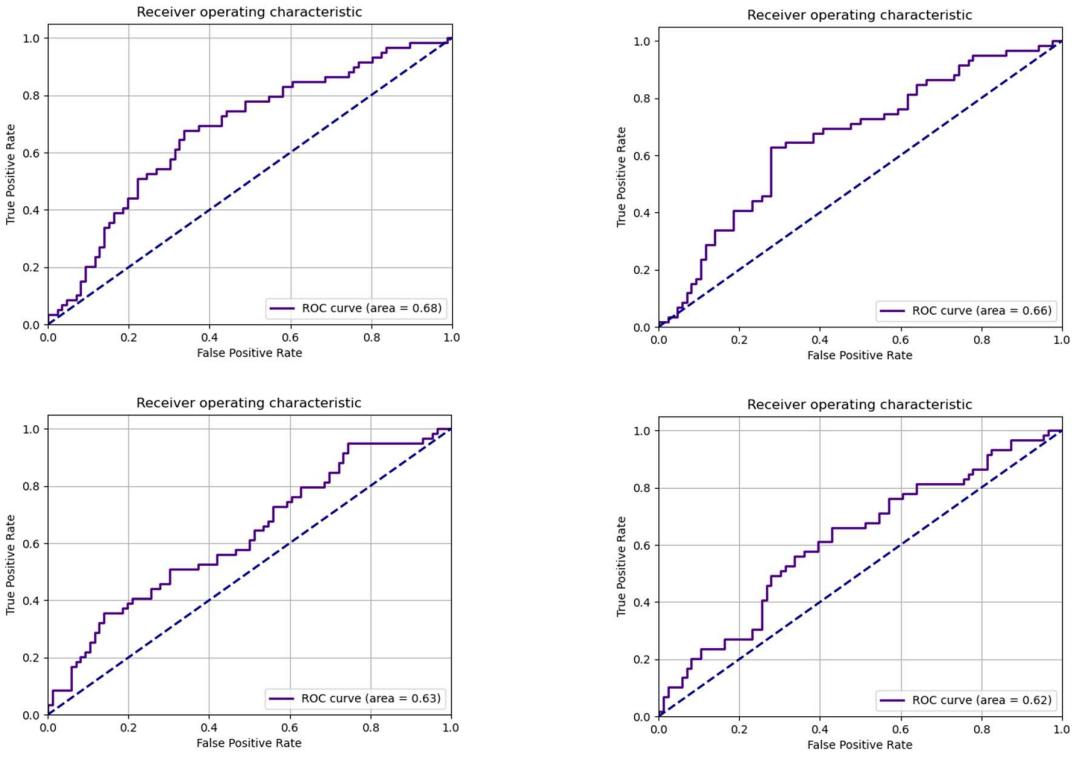

Figure 2: Slide-level classification ROC curve and AUC of MIL variations. Top left to bottom right – Single bag size 50 max pooling with class weighting, Single bag size 50 max pooling with different first conv stride, Single bag size 50 max pooling, Single bag size 50 avg pooling.

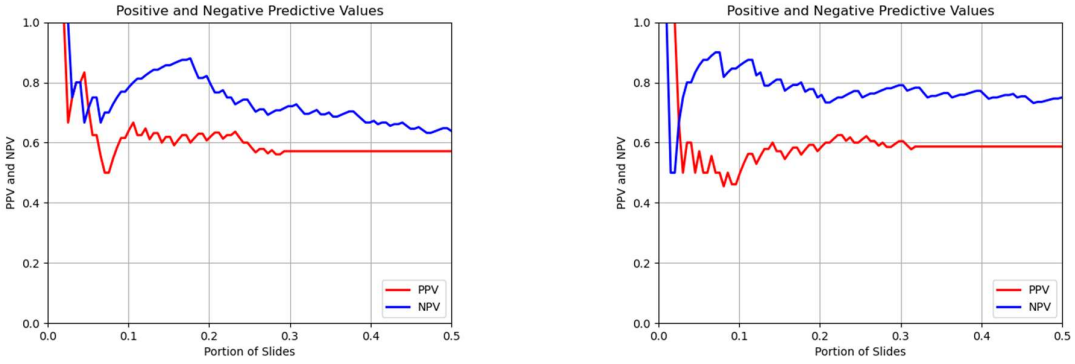

Figure 3: PPV-NPV plots for single bag size 50 with max pooling. Balanced class weighting and unweighted.

In terms of classification accuracy, the better experiments resulted in a classification accuracy in the range between 65% and 70%.

Unfortunately, while some attention MIL variations performed better than others, the overall performance of end-to-end training of MIL were sub-par compared to the performance of supervised training. We attribute the poor performance to reasons such as instability of the bags, the i.i.d. assumption, overfitting, and difficulties relating to the optimization and computation process. Further discussion of the problems, limitations and possible improvements of MIL is featured in the next section.

Although the results of the MIL architecture were not ideal, we are able to make multiple conclusions. First, the use of focal loss did not significantly improve results over standard cross-entropy. In addition, using a relatively small bag size, of 10-50 instances, and selecting only a single bag per slide, seems to be a better and sufficient choice for the dataset. Selecting multiple bags per slide proved to be even less stable and lead to poor performance. Other conclusions include max pooling aggregation performing better than mean pooling, as well as balanced class weighting improving performance.

Using the SSL MoCo implementation, we performed unsupervised pre-training of the feature extractor on a combination of datasets of WSI from patients with breast cancer available in the lab. After the pre-training stage, we performed supervised fine-tuning of a classifier using two different strategies. The first strategy is regular fine-tuning, where the feature extractor weights are frozen, and only a classifier MLP head is trained on the target dataset. The second strategy is using the pre-trained weights as an initialization for supervised learning. The fine-tuning was performed on two target datasets, HEROHE as explained above, and another small dataset of an order of 150 WSIs for classification of the Oncotype breast cancer score. We compared the performance of both fine-tuning strategies to supervised training from scratch on the target dataset.

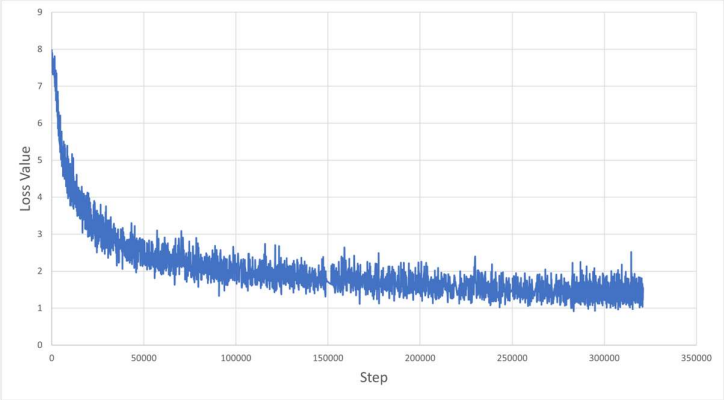

Figure 4: MoCo v2 self-supervised pre-training on a combination of WSI datasets. Contrastive loss convergence over 1000 epochs.

Task	Supervised	SSL Fine-Tuning	SSL Initialization
HEROHE Her2 fold 1	0.783	0.748	0.806
HEROHE Her2 fold 2	0.788	0.821	0.830
Oncotype fold 1	0.631	0.626	0.639
Oncotype fold 2	0.688	0.704	0.756

Table 1: Slide classification AUC score comparison of using the pre-trained feature extractor from MoCo as opposed to performing supervised training from scratch.

As can be observed, while the performance of fine-tuning still has room for improvement, the experiments show the potential of contrastive SSL for use on WSI analysis tasks. We note that since fine-tuning only involves training a small classifier over the learned features from pre-training, the results show that the learned features are capable of generalizing to multiple different downstream tasks. In addition, using the features from SSL pre-training often outperforms simple supervised training from scratch, further suggesting the potential of this approach for further use. Both target datasets are relatively small when compared to the dataset used for unsupervised pre-training, meaning that SSL also allows improving performance on smaller dataset using a larger, unlabeled one.

In total, while there are still improvements to be made in both research directions we explored in the project, we conclude that the results indicate potential for using both for WSI analysis. As a follow up to this project, we intend to work on multiple improvements and other suggestions for architectures inspired by our work here, as discussed in the following section.

4. Limitations and Future Work

As observed in our experiments, end-to-end training of the attention-based MIL architecture did not achieve our expected results and was often outperformed by supervised training from scratch on HEROHE. We suggest several explanations for this. First, it is possible that the choice of bag size and sampling algorithm for patches in a bag has a more significant effect than we attributed to it, meaning that further experiments and modifications to these aspects can still improve results. Second, and likely more importantly, we propose that the MIL assumption, in its original form, raises multiple problems when applied to WSIs.

In the MIL formulation, as previously explained, instances in a bag are assumed to be sampled i.i.d. from the instance distribution, and a positive bag label implies the existence of a positive instance in the bag. However, both assumptions are not necessarily the case in our setting. When pathologists examine WSIs, their diagnosis is usually affected by a combination of attributes in different areas of the slide, in such a way that the conclusion made from a single patch might depend on the contents of other patches. This means that patches in a WSI are inherently correlated when making a slide-level prediction, in contradiction to the i.i.d. assumption. In addition, we speculate that randomly sampling an order of 10 patches from a WSI, may result in a bag that does not contain a single patch that is of any indication of the slide label. Such sampling can then lead to noisy and unstable bags during training, and eventually hurt performance. This problem is not easily solvable by simply increasing the bag size, as too many instances in a bag can also encourage instability and increase the complexity of the data.

We suggest two possible research directions to improve on the original attention MIL architecture, both of which we are now actively working on. The first is reducing the use of the attention MIL mechanism to a fine-tuning stage. We can perform supervised training from scratch and then fine-tune the classifier portion of the model using the attention-based aggregation trained with the MIL assumption. This has already shown initial results of improving performance outside the scope of the project. The second improvement is replacing the attention mechanism with self-attention and making use of the Transformer architecture. Self-attention does not require the i.i.d. assumption, as the correlations between instances in the bags can be learned during training and considered when making predictions. Self-attention also possesses many desirable properties that are likely to improve results as has been shown in recent research papers.

Regarding the SSL part of the project, we see the outcome as a good initial result suggesting the potential of using SSL to improve performance for multiple tasks in histopathology. We propose that there are multiple improvements to be made in both the pre-training and fine-tuning stage of the architecture. Such improvements include further examining of the augmentations used on histopathology data, modifications to the selection process of positive and negative samples during pre-training, for example by making use of the existing partial labels and metadata, and another potential idea of combining MIL and contrastive learning

by attempting to learn similarity of slide-level representations. The last suggestion can also be combined with the use of self-attention as mentioned in the previous paragraph.

5. Conclusion

Throughout the project, we implemented multiple methods and architectures in the attempt of dealing with the difficult challenges of using deep learning to analyze histopathology WSIs. We evaluated our proposed methods for prediction of clinical outcomes from WSIs, suggested and examined possible improvements to our solutions, and analyzed the results. Finally, we reflected on the results of our experiments to find various limitations to our methods and suggested possible future research directions. We intend to use conclusions and experience from this project in our current and future research, in the hopes of further progressing the capabilities of digital WSI analysis.

References

[1] G. Shamai, Y. Binenbaum, R. Slossberg, I. Duek, Z. Gil and R. Kimmel, "Artificial intelligence algorithms to assess hormonal status from tissue microarrays in patients with breast cancer," *JAMA network open,* vol. 2, p. e197700–e197700, 2019.

[2] J. D. Ianni, R. E. Soans, S. Sankarapandian, R. V. Chamarthi, D. Ayyagari, T. G. Olsen, M. J. Bonham, C. C. Stavish, K. Motaparthi, C. J. Cockerell and others, "Tailored for real-world: a whole slide image classification system validated on uncurated multi-site data emulating the prospective pathology workload," *Scientific reports,* vol. 10, p. 1–12, 2020.

[3] L. Pantanowitz, G. M. Quiroga-Garza, L. Bien, R. Heled, D. Laifenfeld, C. Linhart, J. Sandbank, A. A. Shach, V. Shalev, M. Vecsler and others, "An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study," *The Lancet Digital Health,* vol. 2, p. e407–e416, 2020.

[4] R. R. Rawat, I. Ortega, P. Roy, F. Sha, D. Shibata, D. Ruderman and D. B. Agus, "Deep learned tissue “fingerprints” classify breast cancers by ER/PR/Her2 status from H&E images," *Scientific reports,* vol. 10, p. 1–13, 2020.

[5] C. Xie, H. Muhammad, C. M. Vanderbilt, R. Caso, D. V. K. Yarlagadda, G. Campanella and T. J. Fuchs, "Beyond classification: Whole slide tissue histopathology analysis by end-to-end part learning," in *Medical Imaging with Deep Learning*, 2020.

[6] X. Zhu, J. Yao, F. Zhu and J. Huang, "Wsisa: Making survival prediction from whole slide histopathological images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[7] E. Wulczyn, D. F. Steiner, Z. Xu, A. Sadhwani, H. Wang, I. Flament-Auvigne, C. H. Mermel, P.-H. C. Chen, Y. Liu and M. C. Stumpe, "Deep learning-based survival prediction for multiple cancer types using histopathology images," *PloS one,* vol. 15, p. e0233678, 2020.

[8] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins and J. Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Medical Image Analysis,* vol. 65, p. 101789, 2020.

[9] M. Ilse, J. Tomczak and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*, 2018.

[10] N. Naik, A. Madani, A. Esteva, N. S. Keskar, M. F. Press, D. Ruderman, D. B. Agus and R. Socher, "Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains," *Nature communications,* vol. 11, p. 1–8, 2020.

[11] T. Chen, S. Kornblith, M. Norouzi and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," *arXiv preprint arXiv:2002.05709,* 2020.

[12] K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[13] X. Chen, H. Fan, R. Girshick and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297,* 2020.

[14] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen and others, "Big self-supervised models advance medical image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[15] O. Ciga, T. Xu and A. L. Martel, "Self supervised contrastive learning for digital histopathology," *Machine Learning with Applications,* vol. 7, p. 100198, 2022.