A comparison of histopathology imaging comprehension algorithms based on multiple instance learning

Adam Saunders[a], Sajal Dash[b], Aristeidis Tsaris[b], and Hong-Jun Yoon[b]

[a]Department of Electrical and Computer Engineering, University of Dayton, Dayton, OH, USA
[b]Oak Ridge National Laboratory, Oak Ridge, TN, USA

ABSTRACT

Whole slide imaging (WSI), also called digital virtual microscopy, is a new imaging modality. It allows for the application of AI and machine learning methods to cancer pathology to help establish a means for the automatic diagnosis of cancer cases. However, designing machine-learning models for WSI is computationally challenging due to its required ultra-high resolution. The current state-of-the-art models use multiple instance learning (MIL). MIL is a weakly-supervised learning method in which the model uses an array of inferences from many smaller instances to make a final classification about the entire set. In the context of WSI, researchers divide the ultra-high-resolution image into many patches. The model then classifies the slide based on an array of inferences from the patches. Among several ways of making the final classification, attention-based mechanisms have resulted in superb accuracy scores. The Transformer, one attention-based algorithm, has reported substantial improvements for WSI comprehension tasks. In this project, we studied and compared several WSI comprehension algorithms. We used the following three datasets: CAMELYON16+17, TCGA-Lung, and TCGA-Kidney. We found that attention-based MIL algorithms performed better than standard MIL algorithms for classifying WSI images, achieving a higher mean accuracy and AUC. However, none of the attention-based algorithms performed significantly better than the others, reporting accuracy scores that varied widely. Presumably, it is due to the limited availability of training samples in the data corpus. Since it is not easy to increase the samples from human subjects, some machine learning techniques like transfer learning could help mitigate this issue.

Keywords: Transformers, multiple instance learning, whole slide imaging, histopathology imaging, deep learning

1. INTRODUCTION

Cancer pathology involves studying tissue samples to make diagnostic decisions. Usually, cancer pathologists must perform this task by hand on glass slides, a process that can be tedious and prone to errors. However, with the introduction of whole-slide imaging (WSI), this process has become digitized. Using digital scans of tissue samples, WSI allows for methods like deep learning that can automate the analysis and diagnosis of cancers. Unfortunately, deep learning with WSI can be difficult. The datasets can be small, and the images are often gigapixels in size, which makes analysis computationally expensive.[1]

One state-of-the-art method for WSI classification is to split the image into many smaller patches and analyze the patches together using multiple instance learning (MIL). MIL is a weakly-supervised method where we group

This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).
Adam Saunders: E-mail: saundersa5@udayton.edu
Sajal Dash: E-mail: dashs@ornl.gov
Aristeidis Tsaris: E-mail: tsarisa@ornl.gov
Hong-Jun Yoon (Corresponding author): E-mail: yoonh@ornl.gov, Telephone: (865) 241-2626

instances of data into sets called bags. Labels exist only for the bags, not the individual instances. Then, we can train a model to infer a classification for the bag based on the instances. Usually, we extract features from the instances instead of working with the instances directly.[2] For WSI, we extract features from the patches to form the instances, and the slide is the bag. With this method, we do not need patch-level annotations for the slide, which are often difficult to obtain.[1]

When using MIL for WSI classification, we wish to have a flexible and interpretable model so we can understand the output. Attention is one way we can accomplish this goal. Attention is a deep learning technique where the model assigns a numerical weight to an item in a sequence based on its relative importance. The model is flexible because it learns what items are important, and the results are interpretable because we can look at which items have a high attention weight.[3]

There are many ways to implement attention. There can be one single "branch" or "head" that calculates the attention weight for an item in a sequence. Alternatively, there can be multiple, parallel branches that calculate attention weights based on different linear projections of the item. Self-attention is a variation that calculates the attention weight based only on the items in the sequence and not on any task-dependent information.[3] The Transformer is a popular algorithm that uses self-attention, but it can be computationally complex.[4]

In regards to MIL, we usually combine instance labels with a simple max-pooling operation to determine a label for the bag. However, we can replace this aggregation operation with an attention-based operation to determine which instances are the most important.[5] For WSI classification, we can determine which patches contribute the most to a classification. This interpretability is highly desirable for medical imaging tasks, as it allows us to create heatmaps that highlight areas of interest on an image.

In this paper, we aim to compare several MIL algorithms for WSI classification, including attention-based algorithms. Our goal is to find an optimal WSI classification algorithm using MIL. We also wish to determine if Transformer-based algorithms can achieve higher performance than other attention-based algorithms despite the large number of trainable parameters and limited training time.

2. METHODS

2.1 MIL Algorithms

To classify a WSI image, the MIL algorithms studied here begin with the process described in Ref. 6. First, we segmented the WSI image based on a saturation threshold in HSV color space. We applied a black mask to all parts of the image that do not contain tissue. Then, we divided the unmasked parts of the image into patches of 256 × 256 pixels. We used a ResNet50 model pretrained for feature extraction to extract 1024-dimensional features from each patch. Each group of features extracted from the patches from a WSI slide formed a bag for the MIL algorithm.

We investigated five different MIL algorithms. The algorithms differ in how they combine the features to classify the slide. We refer to the standard MIL algorithm that uses max-pooling simply as "MIL." The other algorithms were attention-based. CLAM SB uses a single-branch of attention to combine information from the patches, while CLAM MB uses multiple branches of attention, the same as the number of classes used for classification.[6] We also study a model that uses two Transformer layers, which we will refer to as "Transformer." TransMIL uses two Transformers connected by a pyramid position encoding generator (PPEG), which uses convolutional layers to relate spatial information between the patches.[7] Figure 1 shows diagrams for all of the algorithms that we study here. Table 1 summarizes the different algorithms and demonstrates the massive number of trainable parameters for Transformer-based algorithms.

2.2 Datasets

We trained and tested our models on several openly available datasets. CAMELYON16 consists of lymph node tissue samples from breast cancer patients with binary labels indicating if the sample contains metastasized tissue. CAMELYON17 is a larger dataset containing similar tissue samples with a more detailed pN-level staging.[8] We simplified the pN-level staging for CAMELYON17 by treating normal and isolated tumor cell (ITC) as negative and all others as positive. The other datasets consist of data collected from The Cancer Genome Atlas (TCGA)

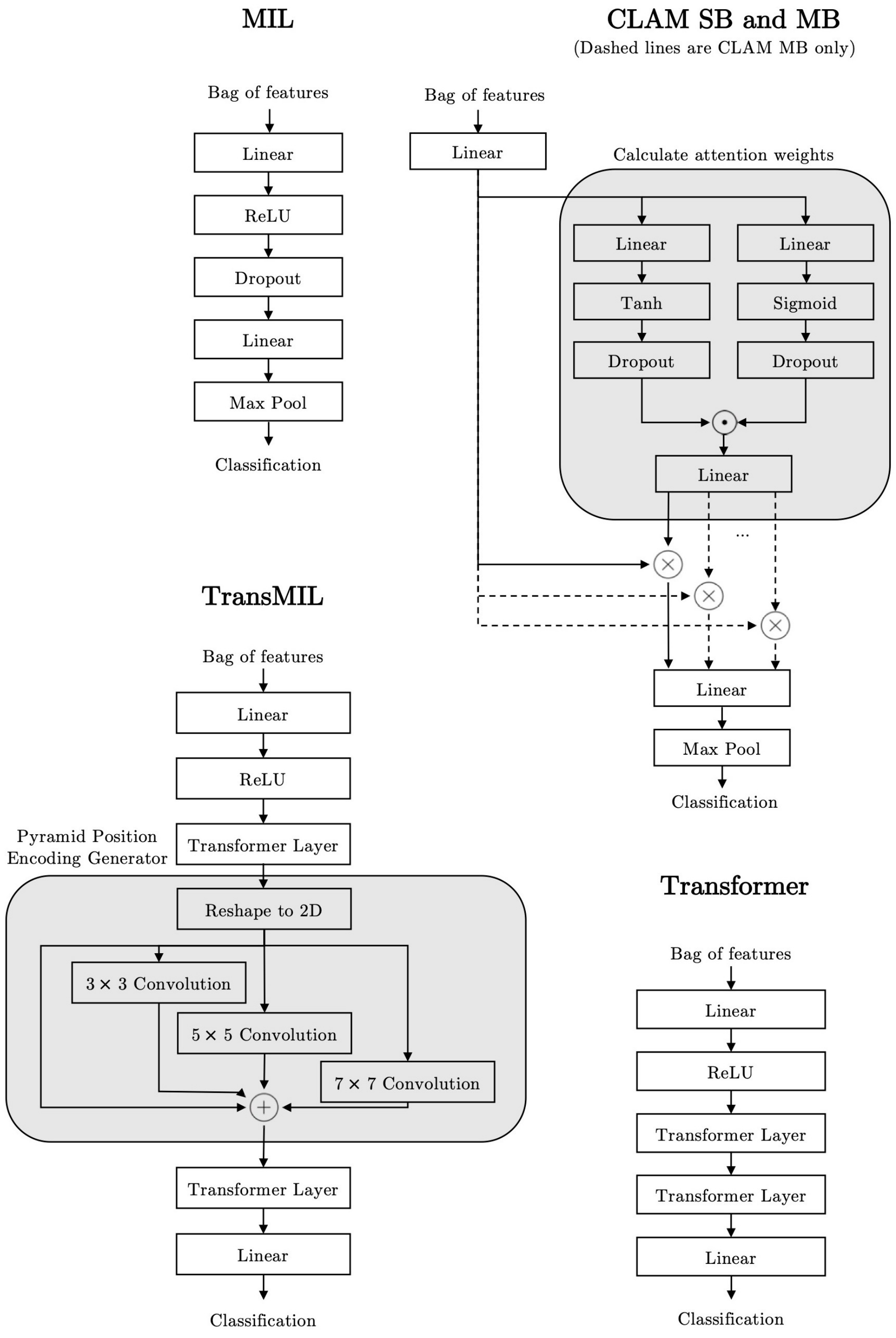

Figure 1: Diagrams for MIL algorithms. Note that $\odot$ represents element-wise multiplication, $\otimes$ represents matrix multiplication, and $\oplus$ represents element-wise addition.

Table 1: Summary of MIL algorithms. Note that the number of trainable parameters was calculated for binary classification problems.

Model	Aggregation operation	Number of trainable parameters
MIL	max-pooling	525,826
CLAM SB	single-branch attention	790,791
CLAM MB	multiple-branch attention	791,084
TransMIL	Transformer with positional encoding	2,672,146
Transformer	Transformer	2,628,114

from the National Cancer Institute. TCGA-Lung is a combination of data from TCGA-LUAD and TCGA-LUSC with labels for classifying tumorous tissues into lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). Finally, TCGA-Kidney is a combination of data from TCGA-KICH, TCGA-KIRC, and TCGA-KIRP with labels to subtype tumorous tissues into kidney chromophobe (KICH), kidney renal papillary cell carcinoma (KIRP), and kidney renal clear cell carcinoma (KIRC).

For CAMELYON16 and the combined CAMELYON16+17, we split the provided training dataset into 90% training data and 10% validation data. We tested both using the CAMELYON16 testing dataset since CAMELYON17 does not provide any labeled testing data. For TCGA-Lung and TCGA-Kidney, we split the dataset into 80% training data, 10% validation data, and 10% testing data.

2.3 Experimental Setup

For each MIL algorithm, we trained 200 models on each of the datasets using the Summit supercomputer at the Oak Ridge Leadership Computing facility. We used a randomized split of data for each model. We calculated the testing accuracy, the area-under-curve (AUC) for the receiver operating characteristic (ROC) curve, and the training time. Since the randomized splits are samples of the training dataset, we used bootstrapping to estimate the mean accuracy, AUC, and training time for all models trained on these datasets.

In addition, we also trained an ensemble of 50 models for each algorithm using the same split of data on each dataset. We used two different methods to aggregate the results from the ensembles: majority voting, where we choose a classification based on the classification chosen by a majority of the models, and soft voting, where we sum the ensemble's softmaxes to determine a classification. We compare these ensemble methods with the mean accuracy and AUC of the individual models. For the ensemble methods, we use bootstrapping to create 200 samples using 65% of the testing dataset, and we generate confidence intervals using the mean accuracy and AUC from these samples.

We trained using the recommended hyperparameters and optimizers for each model. For MIL, CLAM SB, and CLAM MB, we used the Adam optimizer with a learning rate of 2×10^{-4} and a weight decay of 1×10^{-5}. We trained for at least 50 epochs and at most 200 epochs with a validation patience of 20 epochs for early stopping. For Transformer and TransMIL, we used the Lookahead optimizer with a learning rate of 2×10^{-4} and a weight decay of 1×10^{-5}. We trained for at most 200 epochs with a validation patience of 10 epochs.

3. RESULTS

3.1 Comparison of Algorithms

3.1.1 Accuracy and AUC

Figure 2 shows the mean accuracy, AUC, and training time for the models for each dataset, while Tab. 2 shows the mean values and confidence intervals for these values. Notice that for all datasets, the standard MIL model has a lower mean accuracy and mean AUC than the attention-based models.

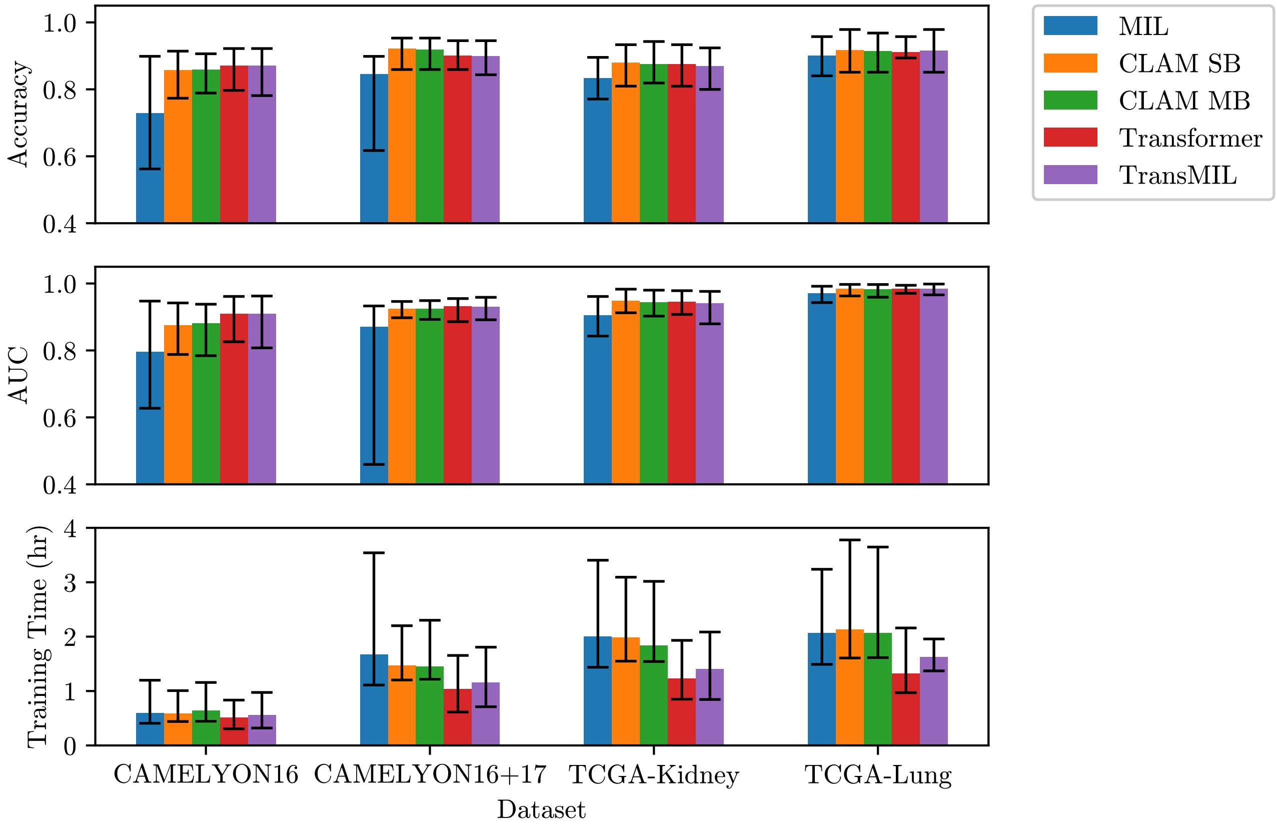

Figure 2: Mean accuracy, AUC, and training time for the models for each dataset ($N = 200$). Error bars represent 95% confidence intervals.

Table 2: Mean values and confidence intervals for accuracy, AUC, and training time of the models.

(a) Mean accuracy and 95% confidence intervals ($N = 200$).

	CAMELYON 16	CAMELYON16+17	TCGA-Lung	TCGA-Kidney
MIL	0.729 (0.562, 0.899)	0.846 (0.617, 0.898)	0.901 (0.840, 0.957)	0.834 (0.771, 0.896)
CLAM SB	0.857 (0.773, 0.914)	**0.921** (0.859, 0.953)	**0.917** (0.851, 0.979)	**0.879** (0.810, 0.933)
CLAM MB	0.858 (0.789, 0.906)	0.919 (0.859, 0.953)	0.913 (0.851, 0.968)	0.875 (0.819, 0.943)
Transformer	**0.870** (0.797, 0.922)	0.901 (0.859, 0.945)	0.911 (0.894, 0.957)	0.875 (0.809, 0.933)
TransMIL	0.870 (0.781, 0.923)	0.898 (0.844, 0.945)	0.915 (0.851, 0.979)	0.869 (0.800, 0.924)

(b) Mean AUC and 95% confidence intervals ($N = 200$).

	CAMELYON16	CAMELYON16+17	TCGA-Lung	TCGA-Kidney
MIL	0.796 (0.627, 0.947)	0.871 (0.459, 0.933)	0.970 (0.943, 0.992)	0.905 (0.843, 0.961)
CLAM SB	0.876 (0.788, 0.942)	0.925 (0.898, 0.946)	0.984 (0.963, 0.997)	**0.948** (0.913, 0.983)
CLAM MB	0.880 (0.784, 0.938)	0.924 (0.893, 0.949)	0.983 (0.959, 0.997)	0.944 (0.903, 0.980)
Transformer	**0.910** (0.826, 0.961)	**0.932** (0.886, 0.955)	**0.984** (0.971, 0.995)	0.946 (0.908, 0.979)
TransMIL	0.909 (0.808, 0.963)	0.930 (0.892, 0.959)	0.984 (0.966, 0.998)	0.941 (0.879, 0.976)

(c) Mean training time (mm:ss) and 95% confidence intervals ($N = 200$).

	CAMELYON16	CAMELYON16+17	TCGA-Lung	TCGA-Kidney
CLAM SB	35:16 (26:23, 60:28)	88:14 (72:12, 132:03)	127:48 (96:28, 226:40)	118:50 (92:59, 185:35)
CLAM MB	38:15 (26:41, 69:28)	87:07 (73:03, 138:05)	124:00 (96:51, 218:50)	109:53 (92:34, 181:00)
MIL	35:30 (24:28, 71:55)	100:11 (66:38, 212:28)	123:56 (89:26, 194:19)	120:03 (86:16, 204:17)
Transformer	**30:38** (18:21, 50:01)	**62:20** (36:49, 99:16)	**79:02** (58:10, 129:35)	**73:40** (51:02, 115:54)
TransMIL	33:07 (19:13, 58:27)	69:17 (42:41, 108:24)	97:27 (82:14, 117:26)	84:17 (50:41, 125:06)

Next, to compare the attention-based algorithms (CLAM SB, CLAM MB, Transformer, and TransMIL), notice that no attention-based model consistently achieves a higher mean accuracy or mean AUC than the other models across all of the datasets. For example, the Transformer-based models (Transformer and TransMIL) achieve a higher mean accuracy on the CAMELYON16 dataset, but both models perform worse than CLAM SB and CLAM MB on the CAMELYON16+17 dataset.

Transformer and TransMIL perform similarly in terms of accuracy and AUC, despite the fact that TransMIL contains the PPEG layer for positional encoding, and Transformer does not.

Figure 3 shows an example ROC curve for the calculation of AUC. Notice that the Transformer and TransMIL both achieve a true positive rate of 100% at a lower threshold than the other models. Therefore, these models may be able to detect true positive cases (i.e., tumorous tissue) with less false positives than other models. However, we should note that we cannot assume that the model has been optimized with the correct threshold, especially if the amount of validation data is limited.[2]

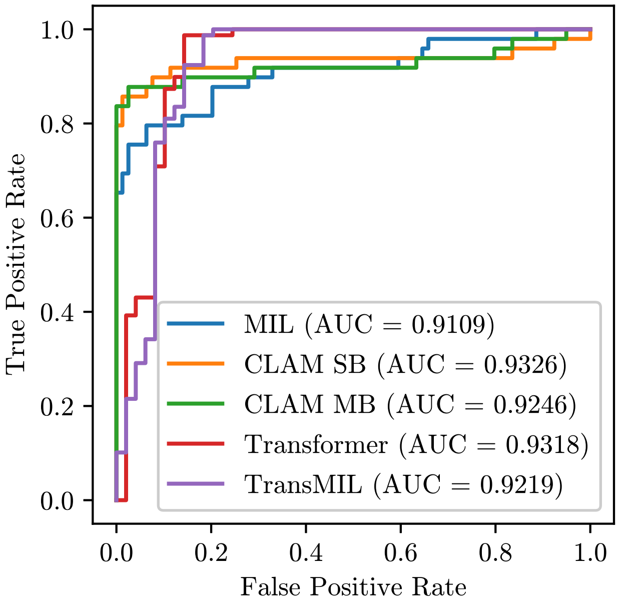

Figure 3: Example ROC curve for CAMELYON16+17 dataset.

3.1.2 Training Time and Memory Usage

MIL, CLAM SB, and CLAM MB all trained in about the same amount of time. On average, Transformer and TransMIL both trained in less time than the other models. However, the validation patience for Transformer and TransMIL was 10 epochs less than the for the other models, so we expect the models to train in less time.

In terms of memory usage during training, most of the models trained on GPUs with 16 GB of RAM. However, for the TCGA-Kidney dataset, Transformer and TransMIL both required GPUs with 32 GB of RAM.

3.2 Ensemble Methods

Figure 4 shows a comparison of the accuracy and AUC for the ensemble methods, and Tab. 3 shows this data in tabular form. We include mean accuracy and AUC of the individual models for comparison. We notice that majority voting and soft voting generally resulted in an increase in the accuracy and AUC for all algorithms. Additionally, soft voting generally produced a higher increase in the accuracy and AUC than majority voting.

We note that the accuracy for Transformer algorithm trained on the TCGA-Lung dataset is particularly low when we used majority voting. This low accuracy shows how majority voting methods can fail to achieve

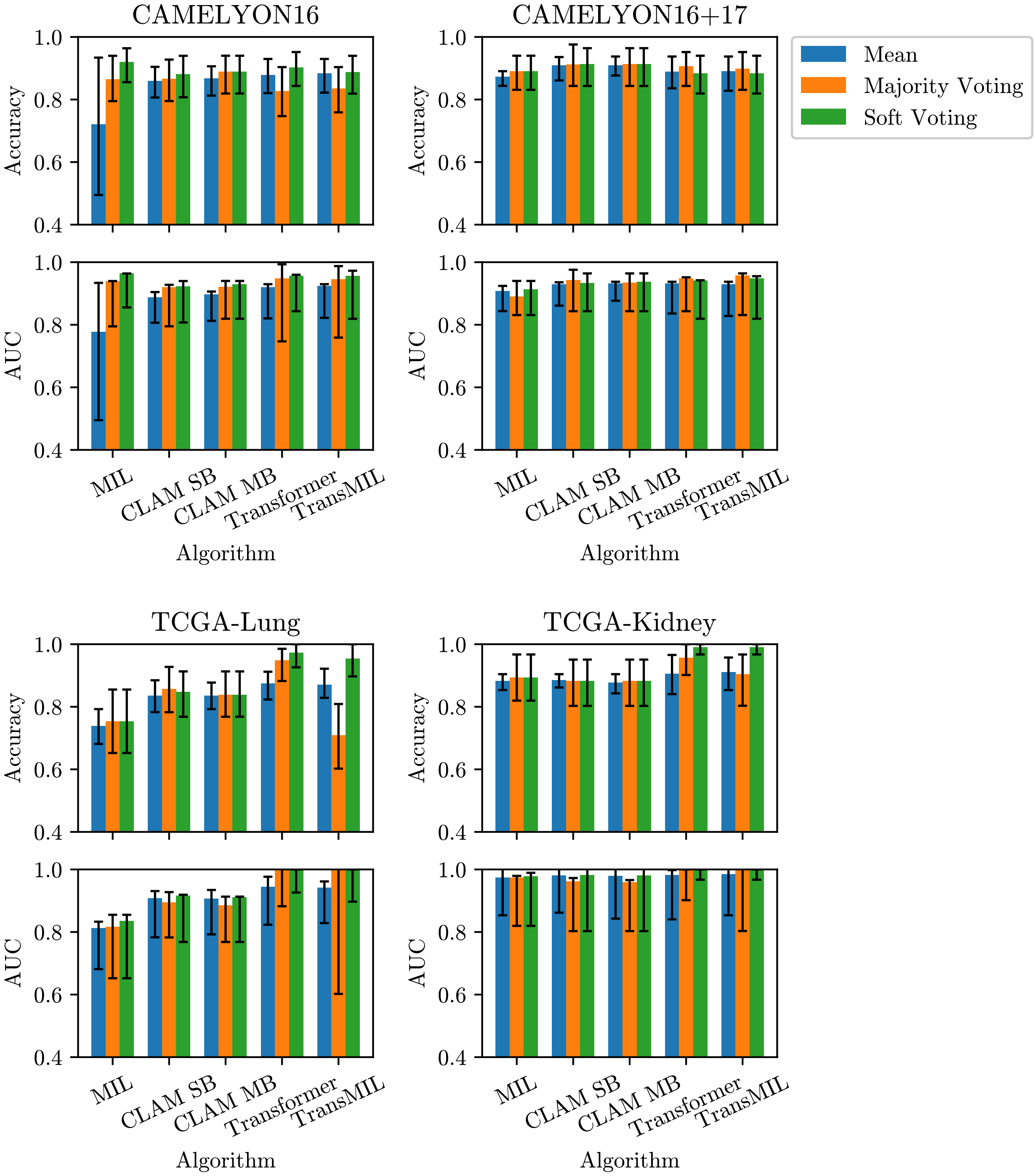

Figure 4: Comparison of accuracy and AUC for ensemble methods. Error bars represent 95% confidence intervals.

Table 3: Accuracy and AUC for ensemble methods.

(a) Accuracy of ensemble methods ($N = 50$).

	CAMELYON16			CAMELYON16+17			TCGA-Lung			TCGA-Kidney		
	Mean	Maj.	Soft	Mean	Maj.	Soft	Mean	Maj.	Soft	Mean	Maj.	Soft
MIL	0.720	0.865	0.920	0.873	0.891	0.891	0.738	0.753	0.753	0.882	0.893	0.893
CLAM SB	0.859	0.866	0.881	0.909	0.913	0.913	0.835	0.856	0.847	0.885	0.882	0.882
CLAM MB	0.868	0.889	0.889	0.909	0.913	0.913	0.835	0.838	0.838	0.877	0.882	0.882
Transformer	0.879	0.827	0.903	0.890	0.906	0.883	0.874	0.948	0.972	0.905	0.957	0.990
TransMIL	0.883	0.835	0.888	0.890	0.898	0.883	0.870	0.709	0.954	0.911	0.904	0.990

(b) AUC of ensemble methods ($N = 50$).

	CAMELYON16			CAMELYON16+17			TCGA-Lung			TCGA-Kidney		
	Mean	Maj.	Soft	Mean	Maj.	Soft	Mean	Maj.	Soft	Mean	Maj.	Soft
MIL	0.777	0.939	0.964	0.907	0.890	0.913	0.813	0.817	0.836	0.973	0.973	0.978
CLAM SB	0.887	0.920	0.923	0.930	0.943	0.934	0.908	0.895	0.916	0.981	0.962	0.981
CLAM MB	0.897	0.921	0.929	0.932	0.934	0.937	0.906	0.885	0.911	0.979	0.958	0.980
Transformer	0.920	0.948	0.956	0.933	0.948	0.941	0.944	0.996	0.997	0.982	0.999	1.000
TransMIL	0.924	0.946	0.956	0.930	0.958	0.948	0.942	0.996	0.996	0.984	1.000	1.000

a higher accuracy than the individual models. For majority voting, this issue can occur even if the individual models have a low error rate.[9]

4. CONCLUSION

This work provides an in-depth comparison of several attention-based MIL algorithms. Previous work has focused on comparing models based on only a few folds,[6,7] but here, we compared the models on a much larger scale, using bootstrapping on hundreds of samples of the available dataset. We obtained a much fairer comparison between the models. We found that the positional encoding layer in TransMIL did not contribute to a significantly higher accuracy or AUC as previously reported.[7] In addition, we found that using ensemble methods like soft voting can increase the accuracy and AUC.

All attention-based MIL algorithms performed WSI classification with satisfactory results. We verified that attention-based MIL algorithms perform better than the standard max-pooling MIL algorithm. We were unable to determine an optimal algorithm for WSI classification since the attention-based MIL algorithms performed similarly, and the accuracies for each algorithm varied widely around the mean. This variation may be due to the limited amount of training data available. Due to the difficulty in obtaining data from human subjects, future research should look into machine learning techniques like transfer learning.

ACKNOWLEDGMENTS

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the DOE Office of Science under Contract No. DE-AC05-00OR22725.

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships program. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

REFERENCES

[1] Dimitriou, N., Arandjelović, O., and Caie, P. D., "Deep learning for whole slide image analysis: An overview," *Frontiers in Medicine* **6** (2019).

[2] Carbonneau, M.-A., Cheplygina, V., Granger, E., and Gagnon, G., "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition* **77**, 329–353 (2018).

[3] Niu, Z., Zhong, G., and Yu, H., "A review on the attention mechanism of deep learning," *Neurocomputing* **452**, 48–62 (2021).

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., "Attention is all you need," in [*Proceedings of the 31st International Conference on Neural Information Processing Systems*], (2017).

[5] Ilse, M., Tomczak, J., and Welling, M., "Attention-based deep multiple instance learning," in [*Proceedings of the 35th International Conference on Machine Learning*], Dy, J. and Krause, A., eds., *Proceedings of Machine Learning Research* **80**, 2127–2136, PMLR (10–15 Jul 2018).

[6] Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F., "Data efficient and weakly supervised computational pathology on whole slide images," *Nat. Biom. Eng.* **5**, 555–570 (2021).

[7] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., and Zhang, Y., "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in [*Advances in Neural Information Processing Systems*], Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., eds., **34**, 2136–2147, Curran Associates, Inc. (2021).

[8] Litjens, G., Bandi, P., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., de Loo, R., Vogels, R., Manson, Q., Stathonikos, N., Baidoshvili, A., Diest, P., Wauters, C., Dlijk, M., and Laak, J., "Supporting data for 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset," (2018).

[9] Vardeman, S. B. and Morris, M. D., "Majority voting by independent classifiers can increase error rates," *The American Statistician* **67**(2), 94–96 (2013).