# Distilling Knowledge from Ensembles of Cluster-Constrained-Attention Multiple-Instance Learners for Whole Slide Image Classification

Folami Alamudun
*Oak Ridge National Laboratory*
Oak Ridge, USA
alamudunft@ornl.gov

Jacob Hinkle
*Oak Ridge National Laboratory*
Oak Ridge, USA
hinklejd@ornl.gov

Sajal Dash
*Oak Ridge National Laboratory*
Oak Ridge, USA
dashs@ornl.gov

Benjamín Hernández
*Oak Ridge National Laboratory*
Oak Ridge, USA
hernandezarb@ornl.gov

Aristeidis Tsaris
*Oak Ridge National Laboratory*
Oak Ridge, USA
tsarisa@ornl.gov

Hong-Jun Yoon
*Oak Ridge National Laboratory*
Oak Ridge, USA
yoonh@ornl.gov

***Abstract*—The peculiar nature of whole slide imaging (WSI), digitizing conventional glass slides to obtain multiple high resolution images which capture microscopic details of a patient's histopathological features, has garnered increased interest from the computer vision research community over the last two decades. Given the unique computational space and time complexity inherent to gigapixel-size whole slide image data, researchers have proposed novel machine learning algorithms to aid in the performance of diagnostic tasks in clinical pathology. One effective algorithm represents a Whole slide image as a bag of smaller image patches, which can be represented as low-dimension image patch embeddings. Weakly supervised deep-learning methods, such as cluster-constrained-attention multiple instance learning (CLAM), have shown promising results when combined with image patch embeddings. While traditional ensemble classifiers yield improved task performance, such methods come with a steep cost in model complexity. Through knowledge distillation, it is possible to retain some performance improvements from an ensemble, while minimizing costs to model complexity. In this work, we implement a weakly supervised ensemble using clustering-constrained-attention multiple-instance learners (CLAM), which uses attention and instance-level clustering to identify task salient regions and feature extraction in whole slides. By applying logit-based and attention-based knowledge distillation, we show it is possible to retain some performance improvements resulting from the ensemble at zero cost to model complexity.**

***Index Terms*—whole slide imaging, pathology, deep learning, ensemble, knowledge distillation, weak supervision, multiple instance learning, model compression, clam, attention, logits, resnet50**



## I. Introduction

Combined, advances in compute capability and artificial intelligence enable machine analysis of gigapixel whole-slide images (WSIs) for a range of tasks in clinical pathology including diagnosis, prognosis, and prediction of therapeutic-response [1]–[3]. Through algorithmic breakthroughs in machine vision, we have witnessed significant advances medical imaging research, notably classification and prediction [2], [4]. Inherent challenges in computational pathology stemming from: *(i)* spatial complexity and memory footprint to process gigapixel-size whole slide images; and *(ii)* availability of manually-annotated gigapixel WSIs. In recent years, these two problems combined constitute the bulk of computational pathology research. Generating image patches is one solution to address spatial complexity of gigapixel WSIs. Patch generation applies a divide-and-conquer strategy, where single gigapixel WSIs are decomposed into a collection of smaller images, which lend themselves to sequential or parallel processing algorithms [5].

Machine learning algorithms for computational pathology, such as deep learning, is dependent on availability of sufficient manually annotated gigapixel WSIs for supervised learning, or large corpora of slide-level labels for weakly supervised learning. However, as only a disproportionately small region within a given gigapixel WSI corresponds to the slide-level label, a majority of learning algorithms rely on pixel or patch-level annotations, or annotated regions-of-interest (ROIs) [6]. The results obtained through such methods reflect performance decline resulting from both noise generated by applying a single label to all patches obtained from a WSI [7], and bias from sub-sampling tissue regions with lower task salience.

Given the high cost of obtaining manually annotated WSIs sufficient for training a deep neural network, recent work gives evidence to clinical grade performance using weak supervision

with slide-level labels on binary classifiers. General weak supervision methods, such as multiple-instance learning (MIL), generate labeled data for supervised training and modeling. However, aggregation functions commonly used for ROI slide-level or patch-level predictions in weakly supervised whole-slide classification methods under-perform in binary and multi-class tissue sub-typing problems, and do not generalize on tests sets in which images are generated by a different device [8].

In [9], Lu et. al. developed clustering-constrained-attention multiple-instance learning (CLAM), a high-throughput deep-learning framework that achieves high performance across different whole-slide-level classification tasks using a limited number of training labels. The initial layers in the CLAM framework combine transfer learning and convolutional neural network (CNN) encoders for dimensionality reduction. In the subsequent layers in the framework apply attention-based multiple-instance aggregation [10] on encoded outputs from the previous layer. This combination allows CLAM to be generalized to multi-class classification and tumor subtpying classification problems in addition to tumor versus normal binary classification tasks.

Ensemble learning algorithms are considered the state-of-the-art for a number of machine learning tasks. The term *ensemble learner* refers to a general class of methods that combine the predictive output multiple base learners to make a final decision; primarily in supervised machine learning tasks [11]. A base learner, is an machine learning algorithm that learns to discriminate a set of labeled examples as input, the end result of which is a a generalizable classification or regression model. By utilizing models generated through base learners, predictions can be drawn for new unlabeled examples.

While the aggregation of base learners in an ensemble yields an improvement in task performance for almost any machine learning algorithm, they can be cumbersome to implement, and come with a steep increase in spatial and computational complexity. These increases are more pronounced when deep neural networks are utilized as base learners in an ensemble; making the latter unsuitable for deployment in many real world applications. Bucilu˘ et al. [12] initially proposed a method for compressing the knowledge encoded in an ensemble of models into a single model.

Recently, much research effort is focused on achieving same or similar accuracy while compressing deep neural networks. These include pruning [13], quantization [14], efficient neural network families [15], [16], and knowledge distillation [17], [18]. In [17], Hington et al. developed a knowledge distillation method by combining a single compressed model, and fewer specialized models. The idea behind knowledge distillation is the augmentation of available class labels through the use of soft probabilities from a larger, fully trained *teacher network* to supervise the training of a smaller *student* network.

In this paper, we designed a study that applies two knowledge distillation methods to compress a bootstrap ensemble of CLAM base learners, trained on the CAMELYON16 [19] dataset. In Section , we describe the bootstrap aggregation sampling applied to the CAMELYON16 dataset. We also describe the logit-based and attention-based knowledge distillation methods used in our experiments. Section presents the results of our experimental; and Section covers a brief discussion, and directions for future work.

## II. Methods

### A. Whole Slide Image Dataset

Assessing the extent of cancer spread by histopathological analysis of sentinel axillary lymph nodes (SLNs) is an important part of breast cancer staging. To encourage the development of diagnostic machine learning algorithms, the CAMELYON16 dataset, for detecting lymph node metastases, was published in November 2016. The CAMELYON16 dataset consists of a total of 399 whole-slide images and corresponding glass slides of SLNs from Radboud University Medical Center (RUMC) and University Medical Center Utrecht (UMCU) in the Neatherlands. RUMC images were produced with a digital slide scanner (Pannoramic 250 Flash II; 3DHISTECH) with a 20x objective lens (specimen-level pixel size, $0.243\mu m \times 0.243\mu m$). UMCU images were produced using a digital slide scanner with a 40x objective lens (specimen-level pixel size, $0.226\mu m \times 0.226\mu m$).

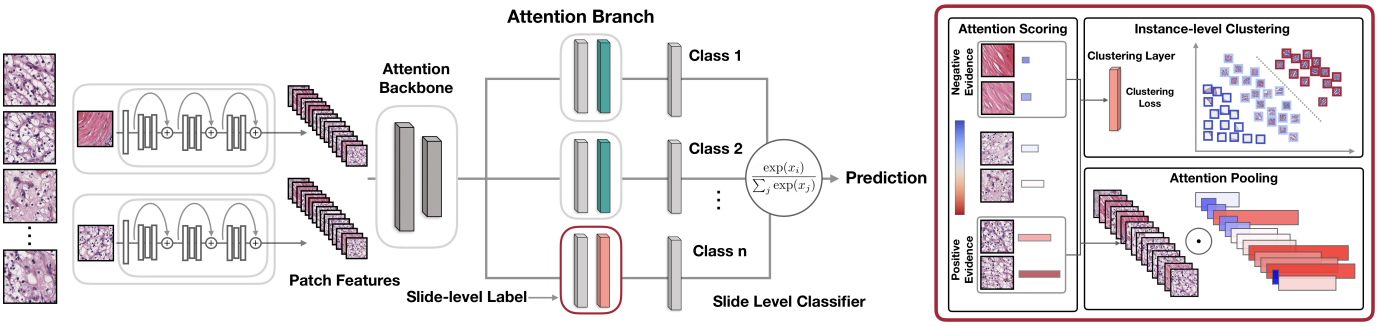
For algorithm development, the dataset includes a training and validation set consisting of (n=110) and without (n=160) nodal metastases verified by immunohistochemical staining. For algorithm performance evaluation, the dataset includes an independent test set of 129 whole-slide images (49 with and 80 without metastases). As a baseline for human performance, the same test set of corresponding glass slides was evaluated by a panel of 11 pathologists with time constraint (WTC) from the Netherlands to ascertain likelihood of nodal metastases for each slide in a flexible 2-hour session, simulating routine pathology workflow, and by 1 pathologist without time constraint (WOTC) [19].

### B. Cluster-Constrained-Attention Multiple-Instance Learning

CLAM is a high-throughput and interpretable framework for data efficient whole slide image (WSI) classification utilizing slide-level labels alone. Not requiring ROI extraction or patch-level annotations, the CLAM framework is capable of modeling multi-class tumor subtyping tasks. In [9], Lu et al. demonstrate the efficacy of the CLAM framework on three WSI datasets. Models trained on each dataset were adapted to independent test cohorts of WSI resections and biopsies as well as smartphone microscopy images (photomicrographs).

### C. Ensemble Learning

Bootstrap aggregation ensemble algorithm, known as bagging, creates a collection of identical base learners trained on sub-samples randomly drawn from an underlying dataset. Samples from the underlying dataset are drawn through bootstrap sampling, where data instances are drawn with replacement. The resulting sampled dataset will have multiple instances of some original samples and no instances of others.




(a)

Fig. 1: Cluster-constrained-attention multiple-instance learning (CLAM) framework for whole slide image classification tasks. WSI images are decomposed to uniform sized patches, compressed using a pretrained CNN-based DNN, and fed through a cluster-constrained attention model to obtain output predictions as illustrated in [9].

The final prediction of the bagging ensemble is obtained by aggregating the predictive outputs of constituent learners. Thus, instance errors for each individual learner are compensated for by other learners in the ensemble. Intuitively, ensemble methodology stems from human nature and our tendency to gather different individualized data informed by unique perspective; individually weighing and combining data points to form opinions and make complex decisions. The main idea is that weighing and aggregating several individualized opinions will be better than aggregating or choosing individual decisions [20]–[22].

### *D. Model Compression*

The basic concept of model compression [12] is to train a miniaturized model to approximate the function learned by a larger, more complex model. First, model compression generates samples by passing unlabeled data through a pretrained, larger or more complex model (*teacher model*). The outputs or scores obtained from the larger model serve as synthetic labels for the corresponding sample. These synthetically labeled data form the training corpus for a smaller and less complex model (*student model*). The mimic model is not trained on the original labels it is trained to learn the function that was learned by the larger model. If the compressed model learns to mimic the large model perfectly it makes exactly the same predictions and mistakes as the complex model. The process of compression suggests the complexity of a model limits its ability to learn a complex function from training data, but the complex function can be approximated through an intermediate, more complex model capable of learning such functions from training data.

In this work, we apply two compression methods on an ensemble of models generated using CLAM framework. The first compression method uses log probability values, also known as logits, obtained before the softmax activation. Training on logarithms of predicted probabilities are shown to improve the learning process for the student model by placing an equal emphasis on the functions learned by the teacher model across all targets. In [18], Ba and Caruana noted that the logits of two samples may differ greatly, showing the teacher model is capable of discriminating between two samples of the same class. However, with a softmax operation, this information is lost as both outputs represent the same class. By training the student model directly on the logits, the student is better able to learn the internal model learned by the teacher, without suffering from the information loss that occurs from passing through logits to probability space.

The second compression method employs activation-based attention transfer, by using spatial attention maps to transfer information from teacher model to student model. In [23], Zagoruyko et al. take the absolute value of a hidden neuron activation as a measure of importance of that neuron w.r.t that specific input, and compute the statistics of these values across the channel dimension to construct class-discriminative spatial attention maps. Defining attention as gradient w.r.t input, can be viewed as an input sensitivity map as described in [24], attention to a spatial location within the input encodes how sensitive the output prediction is w.r.t. changes at that input location. To obtain gradient of loss w.r.t input for both student and teacher model, we recall from [23]:

$$J_S = \frac{\partial}{\partial x}\mathcal{L}(\mathbf{W}_S, x), J_T = \frac{\partial}{\partial x}\mathcal{L}(\mathbf{W}_T, x) \tag{1}$$

Then, using $l_2$ distance, we can minimize the distance between the student gradient attention and teacher attention:

$$\mathcal{L}_{AT}(\mathbf{W}_S, \mathbf{W}_T, x) = \mathcal{L}(\mathbf{W}_S, x) + \frac{\beta}{2}\left\|J_S - J_T\right\|_2 \tag{2}$$

Given $\mathbf{W}_T$ and $x$, the derivative $\frac{\partial}{\partial \mathbf{W}_S}\mathcal{L}_{AT}$ can be obtained as:

$$\frac{\partial}{\partial \mathbf{W}_S}\mathcal{L}\left(\mathbf{W}_S, x\right) + \beta\left(J_S - J_T\right)\frac{\partial^2}{\partial \mathbf{W}_S \partial x}\mathcal{L}\left(\mathbf{W}_S, x\right) \tag{3}$$

To enforce horizontal flip invariance on gradient attention maps, horizontally flipped images as well as originals are propagated, then backpropagated and gradient attention maps are flipped back, as proposed in *Group Equivariant CNN* proposed by Cohen & Welling in [25]. We then add l2 losses on the obtained attentions and outputs, and do second backpropagation:

$$\mathcal{L}\left(\mathbf{W}, x\right) + \frac{\beta}{2}\left\|\frac{\partial}{\partial x}\mathcal{L}\left(\mathbf{W}, x\right) - flip(\frac{\partial}{\partial x}\mathcal{L}\left(\mathbf{W}, flip(x)\right))\right\|_2 \tag{4}$$

## III. Experimental

In the following section, we compare the performance of knowledge distillation using logit-based distillation and attention-based distillation on an ensemble of CLAM learners trained on the *CAMELYON16* dataset. We split this section into two parts, first we describe the training and model selection process, which is uniform across all experiments. Then we include the experimental results for the base CLAM model ($CLAM_B$, and results for both the CLAM ensemble used as *teacher model* ($\text{CLAM}_T$) and both the logit-based ($\text{CLAM}_{SL}$) and the attention-based ($\text{CLAM}_{SA}$) distilled student models.

### A. Training and Model Selection

All our experiments were run using WSIs dataset partitions drawn from the *CAMELYON16* training dataset. Using the CLAM pipeline, we segmented the tissue region in each WSI image and extracted $256 \times 256$ pixel-sized patches. Each patch was then compressed using a deep CNN to convert tissue image patches to a low-dimensional image embedding. For this experiment, we used a deep residual network model [26] pre-trained on ImageNet, and applied adaptive mean-spatial pooling after the third residual block, converting each $256 \times 256$ patch into a $1024$ dimension feature vector.

During training, for each slide, we apply a cross-entropy loss function on scores of patch selected through max-pooling; model parameters are optimized via stochastic gradient descent using a batch size of one and the Adam optimizer with the same hyperparameters as CLAM. Namely, we use a learning rate of $2 \times 10^4$, a weight decay of $1 \times 10^5$, with $\beta 1 = 0.9$, $\beta 2 = 0.999$ and $\epsilon = 1 \times 10^8$.

The $\text{CLAM}_B$ model and both student models were evaluated using a k-fold cross-validation ($k = 10$) scheme by creating ten unique partitions. For each fold, we reserve a partition as hold-out, while the rest of the data is split using $90\%$ for model training, and $10\%$ for validation (5%). To create the ensemble, we trained $100$ $\text{CLAM}_B$ models as base-learners. The bootstrap aggregation ensemble algorithm first creates 100 partitions. Each partition is formed by uniformly sampling 95% of the *CAMELYON16* without replacement. Bootstrap samples ($n = 270$) are drawn uniformly with replacement from each partition to from model training (90%) and validation (10%) splits.

All models are trained for at least 50 epochs and up to a maximum of 100 epochs if the early stopping criterion is not met. Validation loss is monitored each epoch and when it has not decreased from the previous low for over 20 consecutive epochs, early stopping is used. The saved model, which has the lowest validation loss, is then tested on the *CAMELYON16* test set ($n = 129$).

### B. Binary Tumor Versus Normal Classification Results

On the CAMELYON16 dataset for breast-cancer-metastasis detection in axillary lymph nodes, the ($\text{CLAM}_B$) model achieved an average test AUC of $0.904$, 95% CI $[0.899, 0.908]$. The ensemble teacher model, ($\text{CLAM}_T$), achieved a test AUC of $0.935$. The logit-based student model ($\text{CLAM}_{SL}$) achieved an average test AUC of $0.916$, 95% CI $[0.892, 0.925]$, while the attention-based student model ($\text{CLAM}_{SA}$), achieved an average test AUC of $0.906$, 95% CI $[0.882, 0.920]$

## IV. Discussion

In this paper, we briefly reviewed the application of computer vision algorithms in clinical pathology to address challenges inherent in digitized whole slide imaging. We also reviewed the cluster-constrained-attention multi-instance-learning framework developed by Lu et al. [9], and described how we can leverage ensemble learning and knowledge distillation to achieve improvements in task performance while minimizing the increased computational costs of deploying ensemble algorithms. As our results on the CAMELYON16 dataset for breast-cancer-metastasis detection in axillary lymph node suggest, we are able to preserve some of the performance improvements obtained through ensemble learning, while entirely eliminating the computational costs, producing a student model of similar complexity to the base-learner (CLAM) but better task performance.


## Acknowledgements

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] M. K. K. Niazi, A. V. Parwani, and M. N. Gurcan, "Digital pathology and artificial intelligence," *The lancet oncology*, vol. 20, no. 5, pp. e253–e261, 2019.

[2] K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi, "Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology," *Nature reviews Clinical oncology*, vol. 16, no. 11, pp. 703–715, 2019.

[3] S. Dash, B. Hernández, A. Tsaris, F. T. Alamudun, H.-J. Yoon, and F. Wang, "A scalable pipeline for gigapixel whole slide imaging analysis on leadership class hpc systems," in *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2022, pp. 1266–1274.

[4] D. Tellez, G. Litjens, J. van der Laak, and F. Ciompi, "Neural image compression for gigapixel histopathology image analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 567–578, 2019.

[5] A. Cruz-Roa, A. Basavanhally, F. González, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," in *Medical Imaging 2014: Digital Pathology*, vol. 9041. SPIE, 2014, p. 904103.

[6] S. Wang, Y. Zhu, L. Yu, H. Chen, H. Lin, X. Wan, X. Fan, and P.-A. Heng, "Rmdl: Recalibrated multi-instance deep learning for whole slide gastric image classification," *Medical image analysis*, vol. 58, p. 101549, 2019.

[7] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature medicine*, vol. 24, no. 10, pp. 1559–1567, 2018.

[8] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[9] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature biomedical engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[10] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.

[11] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[12] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 535–541.

[13] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[14] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, "Quantized convolutional neural networks for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4820–4828.

[15] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4794–4802.

[16] A. Gordon, E. Eban, O. Nachum, B. Chen, H. Wu, T.-J. Yang, and E. Choi, "Morphnet: Fast & simple resource-constrained structure learning of deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1586–1595.

[17] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[18] J. Ba and R. Caruana, "Do deep nets really need to be deep?" *Advances in neural information processing systems*, vol. 27, 2014.

[19] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[20] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[21] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 241–258, 2020.

[22] X. Zhou, J. He, and C. Yang, "An ensemble learning method based on deep neural network and group decision making," *Knowledge-Based Systems*, vol. 239, p. 107801, 2022.

[23] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[24] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[25] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *International conference on machine learning*. PMLR, 2016, pp. 2990–2999.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.