Domain-Specific Pre-training Improves Confidence in Whole Slide Image Classification

Soham Rohit Chitnis[1], Sidong Liu[2,*], Tirtharaj Dash[3], Tanmay Tulsidas Verlekar[1],
Antonio Di Ieva[2,*], Shlomo Berkovsky[2,*], Lovekesh Vig[4], Ashwin Srinivasan[1]

Abstract— Whole Slide Images (WSIs) or histopathology images are used in digital pathology. WSIs pose great challenges to deep learning models for clinical diagnosis, owing to their size and lack of pixel-level annotations. With the recent advancements in computational pathology, newer multiple-instance learning-based models have been proposed. Multiple-instance learning for WSIs necessitates creating patches and uses the encoding of these patches for diagnosis. These models use generic pre-trained models (ResNet-50 pre-trained on ImageNet) for patch encoding. The recently proposed KimiaNet, a DenseNet121 model pre-trained on TCGA slides, is a domain-specific pre-trained model. This paper shows the effect of domain-specific pre-training on WSI classification. To investigate the effect of domain-specific pre-training, we considered the current state-of-the-art multiple-instance learning models, 1) CLAM, an attention-based model, and 2) TransMIL, a self-attention-based model, and evaluated the models' confidence and predictive performance in detecting primary brain tumors - gliomas. Domain-specific pre-training improves the confidence of the models and also achieves a new state-of-the-art performance of WSI-based glioma subtype classification, showing a high clinical applicability in assisting glioma diagnosis. We will publicly share our code and experimental results at **https://github.com/soham-chitnis10/WSI-domain-specific**.

Keywords— Domain-specific Pre-training, Whole Slide Image Classification, Multiple Instance Learning, Brain Tumor

I. INTRODUCTION

Currently, histopathology is the clinical gold standard for tumor assessment and diagnosis. With advancements in digital pathology and artificial intelligence (AI), deep learning techniques have shown great potential in assisting tumor diagnosis. However, deep learning-based approaches require either manually annotated whole slide images (WSIs) or large datasets with slide-level labels in a weakly supervised setting. The slide-level labels may correspond to small regions from the gigapixel WSIs. Therefore, most of the approaches rely on pixel, patch, or regions-of-interest (ROI) level annotations [1], [2], [3], [4]. There have been attempts to assign the same label to each patch [5], but these approaches suffer from noisy training and are not applicable when the slide-level label corresponds to tiny regions from the image.

Recently proposed multiple-instance learning models use feature embedding [6], [7], [8], [9], [10], [11]. Clustering-constrained Attention Multiple instance learning (CLAM) [10] is a deep learning-based digital pathology framework for weakly supervised WSI classification using attention-based instance-level clustering. The CLAM algorithm does not require ROI extraction, or pixel or tile or patch-level annotations. The algorithm works based on the attention scores assigned to the tiles or patches during the training, where the highest attention scores are considered positive evidence of the class. A ResNet50 network [12] pre-trained on the ImageNet dataset is used for extracting features from tiles or patches obtained from the segmented tissue region. CLAM produces interpretable heatmaps that allow medical practitioners to visualize the tissue regions that contribute to the predictions of the model. These heatmaps displaying the morphological features can be verified by trained pathologists whether the decisions made by the model align with the manual diagnostic determinations. TransMIL [11] is a recently proposed model which is based on a transformer model. Just like CLAM, TransMIL only requires a slide-level label and also shows great interpretability.

Training a deep neural network from scratch for specific problem statements may not be feasible due to limited available data, or training may take a very long time. Therefore, many researchers exploit pre-trained models. In domain-specific pre-training, the models are first pre-trained on a large dataset, followed by a fine-tuning over the domain-specific dataset. Generally, in computer vision, models pre-trained on ImageNet data have been used for domain-specific downstream tasks. We call a model pre-trained on ImageNet as generic pre-trained, and the model fine-tuned on a domain-specific dataset as domain-specific pre-trained.

Current state-of-the-art WSI classification models, such as CLAM and TransMIL, are based on features extracted by generic pre-trained models, e.g., ResNet50 pre-trained on ImageNet samples, which differ from WSIs in terms of color, texture, morphological, and geometric representations. However, the impact of domain-specific pre-training on WSI classification has not yet been investigated. Our main contributions are as follows:

- We introduce Confidence as an evaluation metric for WSI classification.
- We investigate the impact of domain-specific pre-

training on the model's performance and confidence.

II. EMPIRICAL EVALUATION

A. Aims

We conjecture that domain-specific pre-training improves not just the accuracy, but also the *confidence* in prediction of the models for WSI classification. We will test this conjecture using the following pre-training methods (a) a generic pre-trained model, DenseNet121 (pre-trained on ImageNet Data), (b) a domain-specific pre-trained model, KimiaNet (pre-trained on WSIs). Generic and domain-specific pre-trained models are used as inputs to two different approaches for constructing models for WSI: (a) CLAM and (b) TransMIL.

B. Datasets

Two datasets are used for pre-training feature extractors, including ImageNet and The Cancer Genome Atlas (TCGA). ImageNet dataset is a classical collection of natural images for the visual recognition task [13]. ResNet50 and DenseNet121 Networks are pre-trained on this dataset. TCGA is the large domain-specific dataset on which KimiaNet was trained. This is a publicly available repository [14], [15], [16] with 30,072 WSIs. KimiaNet was trained on a subset of this dataset with 11,579 WSIs with permanent hematoxylin and eosin (H&E) sections. Further details are available in [17].

The Digital Brain Tumor Atlas (DBTA) dataset [18] is used to train and test the classification models. DBTA dataset consists of 3,115 slides of 126 brain tumor types. A subset of this dataset contains a total of 866 histopathology slides of 791 patients with the 5 primary brain tumor types with molecular subtypes based on the 2021 WHO classification of central nervous system tumors [19], including 117 astrocytoma-IDH mutant, 66 astrocytoma-IDH wildtype, 176 oligodendroglioma-IDH wildtype, 34 glioblastoma-IDH mutant, and 473 glioblastoma-IDH wildtype WSIs. The 70:15:15 split is used for the training, validation, and test sets. As several patients have multiple slides, we ensure that different slides of a patient do not exist in both the training and test sets.

C. Models

We focus on the recently developed state-of-the-art models, CLAM [10], and TransMIL [11], both of which are multiple-instance learning models.

CLAM: We consider two variants of CLAM, namely single branch (CLAM-SB) and multiple branches (CLAM-MB). CLAM uses attention-based pooling [6] to aggregate the slide-level representations from patch-level representations. CLAM has N multi-class attention branches for the N multi-classification problem, which are used to score the class-specific slide-level representations. During training, it has the instance clustering level, which can learn the class-specific features; therefore, CLAM uses a weighted loss of the slide-level classification, and the instance-level clustering loss is used. The slide-level classification is the standard cross-entropy loss, and the instance clustering loss is a smooth SVM loss. For the instance clustering level, we consider $B = 8$, which is 8 positive and negative samples.

TransMIL It uses a self-attention-based mechanism to model the correlation between the patches. TransMIL has two transformer layers for aggregating the morphological information. TransMIL has adopted the approximation for self-attention proposed in the Nyström method [20]. It uses the Pyramid Positional Encoding Generator (PPEG) for positional encoding, which can encode global and context information. TransMIL uses the standard cross-entropy loss for training the model.

D. Evaluation Metrics

The state-of-the-art models have used accuracy and Area Under the Receiver Operating Characteristic Curve (AUC) as the evaluation metrics for benchmarking. We also evaluate the confidence of models, which is the average class probability of the predicted class. Confidence is given by the equation (1), where y_{pred} is the predicted class-label, and x is a sample from the test set X.

$$Confidence = \frac{\sum_x P(y_{pred}|x)}{|X|} \quad (1)$$

E. Algorithms and Machines

We train all models on NVIDIA Tesla V100-SXM2 GPU with 32GB memory. The code has been implemented using PyTorch Deep Learning Framework, and Python *net:cal* library [21] is used for evaluation of the confidence of the model.

III. METHOD

A. Pre-Processing

The gigantic size of WSIs poses challenges in training a model in an end-to-end fashion and would require expensive computing power; therefore, multiple-instance learning models are proposed. In the context of WSI classification, we consider the slide as the bag and the patches of the slides as the instances. Figure 1 shows the key steps of the WSI pre-processing pipeline.

Segmentation and patching For the segmentation and patching of each slide, we use the highly efficient method proposed by CLAM [10]. Patches of size 256 x 256 are cropped from the segmented foreground contours. DBTA has the slide image captured at 20x and 40x resolution. The higher the magnification, the higher the number of patches, leading to higher computational requirements. Therefore all models have been trained at 20x magnification.

Feature extraction The MIL-based models require creating patches from the slides. CLAM [10] and TransMIL [11] methods require a feature extraction process. They extract the features using a generic pre-trained model,
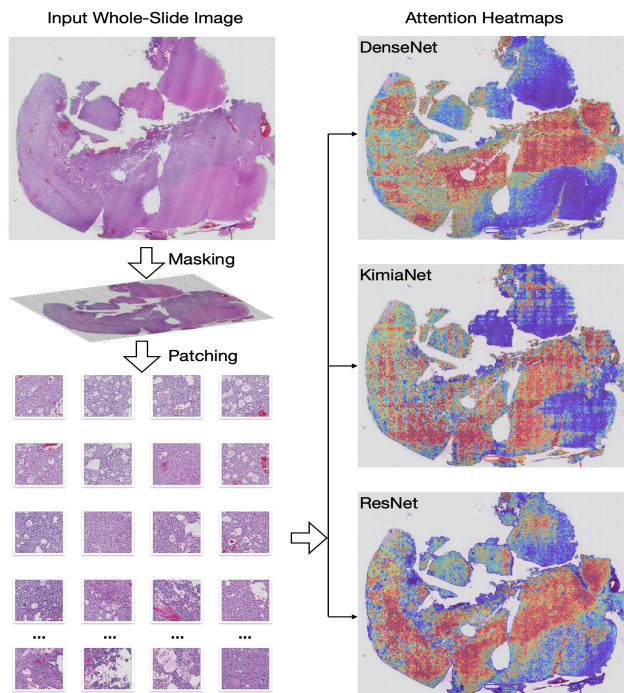
Fig. 1. The WSI pre-processing pipeline.

ResNet50 network pre-trained of the ImageNet data. Therefore, ResNet50 is considered the baseline. DenseNet architecture has shown better results on the ImageNet dataset due to the dense connections between all layers with matching feature-maps sizes [22]. Instead of the generic pre-trained model (DenseNet121), we propose to use the domain-specific pre-trained network called KimiaNet [17], which is a DenseNet121 network pre-trained on TCGA data. We will consider the DenseNet121 network, pre-trained on ImageNet, for comparison with KimiaNet, whose structure is the same.

B. Experimental Procedure

Our procedure to test the experimental conjecture in Sec. II-A is straightforward:

1) Let a slide be denoted by s, where $s \in S$, the set of slides
2) For each $s \in S$, segment and create patches; the set of all patches of s is denoted by p
3) For each pre-training method in {ResNet50, DenseNet121, KimiaNet}
   - For each p, extract features
   - Let D denote an array of data seeds, and M denote an array of model seeds
   - For each $data$-$seed$ in D:
     – Split the dataset into train/validation/test sets
     – For each $model$-$seed$ in M:
       * Use the selected features to train each classification model in {CLAM-SB, CLAM-MB, TransMIL}
       * Estimate the test accuracy, AUC, and confidence of the classification model

The following additional details are relevant:
- $data$-$seed$ is responsible for the train/validation/test split, and $model$-$seed$ is responsible for the initialization of model parameters. The number of $data$-$seed$ is 5 and the number of $model$-$seed$ is 3 in this study, so there are 15 combinations in total. In all cases, average values over all 15 experiments of accuracy, AUC, and confidence are used for comparison.
- We train models for 200 epochs in each run and use 'early stopping' with patience of 20 epochs. We train the model for at least 50 epochs to ensure the model has converged. The learning rate is set to 2e-4 with weight decay of 1e-5. The batch size is set to 1.
- Adam optimizer [23] is used for training CLAM, and Lookahead optimizer [24] is used for training TransMIL.
- Multinomial sampling is used to mitigate the class imbalance problem in the training set. The multinomial sampling probabilities are inversely proportional to the frequency of the ground truth.

IV. RESULTS

The main results from our experiments comparing generic and domain-specific pre-training are presented in Table I. The most interesting observation is that domain-specific pre-training results in higher average confidence in prediction (by 0.3% to 1.3%), which appears to be less well-known. The impact of domain-specific pre-training on the model's test accuracy and AUC are not conclusive. For example, when using CLAM as the classification model, KimiaNet results in higher accuracy, but DenseNet results in higher AUC, and it is the opposite when using TransMIL. We also note that, unlike the results in [11], which showed a clear advantage of using the transformer-based approach, we find results for TransMIL are mixed. This may be due to differences in the dataset we have used, which is different to the benchmarks used, or due to differences in hyper-parameter tuning.

TABLE I
EVALUATION OF DOMAIN-SPECIFIC PRE-TRAINING

Models	AUC		Accuracy		Confidence	
	DenseNet	KimiaNet	DenseNet	KimiaNet	DenseNet	KimiaNet
CLAM-SB	**95.86**	95.69	88.24	**89.84**	92.64	**93.51**
	(± **1.80**)	(± 1.99)	(± 3.74)	(± **2.60**)	(± 3.37)	(± **2.52**)
CLAM-MB	**96.53**	96.27	88.74	**89.72**	93.08	**94.37**
	(± **1.73**)	(± 1.51)	(± 2.64)	(± **4.03**)	(± 2.67)	(± **4.18**)
TransMIL	95.52	**96.85**	**90.58**	89.36	93.28	**93.57**
	(± 2.76)	(± **1.20**)	(± **1.93**)	(± 2.30)	(± 4.98)	(± **3.43**)

Table II shows a comparison between ResNet50 and DenseNet121. We note that the structure of the generic pre-trained model also plays an important role in improving both predictive accuracy (by 1.1% to 4.6%) and confidence (by 1.6% to 9.6%). Since the DenseNet structure is utilized by KimiaNet, the gains made by using the DenseNet structure are incorporated in the results with KimiaNet. In combination, we believe the results provide evidence for the improved performance on model's confidence with domain-specific pre-training, irrespective of the CLAM or TransMIL models used for classification.

Classification of glioma types (i.e., astrocytoma, oligoden-droglioma, and glioblastoma) is generally a difficult task; the current state-of-the-art performance is accuracy of 0.861 and AUC of 0.961 [25]. Classification of glioma molecular subtypes is also challenging; for instance, the best binary classification performance of the IDH-gene molecular sub-type is accuracy of 0.882, and AUC of 0.931 [26]. We test our models on the classification of both glioma types and molecular subtypes, which is a more challenging task, yet the models that use KimiaNet or DenseNet with TransMIL outperform this prior art.

TABLE II

EVALUATION OF THE STRUCTURE OF FEATURE EXTRACTOR

	AUC		Accuracy		Confidence	
Models	ResNet	DenseNet	ResNet	DenseNet	ResNet	DenseNet
CLAM-SB	95.43 (± 1.35)	**95.86** (**± 1.80**)	83.67 (± 4.23)	**88.24** (**± 3.74**)	83.05 (± 4.85)	**92.64** (**± 3.37**)
CLAM-MB	**96.70** (**± 1.03**)	96.53 (± 1.73)	87.65 (± 2.87)	**88.74** (**± 2.64**)	89.07 (± 4.18)	**93.08** (**± 2.67**)
TransMIL	95.40 (± 2.53)	**95.52** (**± 2.76**)	88.18 (± 3.53)	**90.58** (**± 1.93**)	91.69 (± 3.92)	**93.28** (**± 4.98**)

V. CONCLUSIONS

This work is the first study that investigates the impact of domain-specific pre-training in WSI classification and uses confidence as a metric to evaluate the model's performance. We first consider the DenseNet model pre-trained on ImageNet and assess its accuracy, AUC and confidence when combined with CLAM and TransMIL to classify brain tumor WSIs. We then explore KimiaNet, a DenseNet121 model pre-trained on TCGA slides to tackle the same problem. We propose domain-specific pre-training of KimiaNet improves the classification accuracy, AUC and confidence of CLAM and TransMIL. The evaluation of our proposal on the DBTA dataset suggests that KimiaNet is a better feature extractor than generic DenseNet and ResNet, as it improves CLAM and TransMIL's confidence. Additionally, using TransMIL with KimiaNet achieves a new state-of-the-art performance of WSI-based glioma classification, which shows high clinical applicability of the model in assisting glioma diagnosis, particularly in predicting the molecular subtypes. This study also sheds light on the impact of the feature extractor's structure on the prediction performance. In our future work, we will investigate if advanced convolutional structures can further benefit multiple-instance learning models and improve their classification performance.

REFERENCES

[1] S. Wang, Y. Zhu *et al.*, "Rmdl: Recalibrated multi-instance deep learning for whole slide gastric image classification," *Medical Image Analysis*, vol. 58, p. 101549, 12 2019.

[2] P.-H. C. Chen, K. Gadepalli *et al.*, "An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis," *Nature Medicine*, vol. 25, pp. 1453–1457, 9 2019.

[3] K. Nagpal, D. Foote *et al.*, "Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer," *npj Digital Medicine*, vol. 2, p. 48, 6 2019.

[4] B. E. Bejnordi, M. Veta *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, p. 2199, 12 2017.

[5] N. Coudray, P. S. Ocampo *et al.*, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature Medicine*, vol. 24, pp. 1559–1567, 10 2018.

[6] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 2127–2136. [Online]. Available: https://proceedings.mlr.press/v80/ilse18a.html

[7] N. Tomita, B. Abdollahi, J. Wei, B. Ren, A. Suriawinata, and S. Hassanpour, "Attention-Based Deep Neural Networks for Detection of Cancerous and Precancerous Esophagus Tissue on Histopathological Slides," *JAMA Network Open*, vol. 2, no. 11, pp. e1 914 645–e1 914 645, 11 2019. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2019.14645

[8] N. Hashimoto, D. Fukushima *et al.*, "Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with non-annotated histopathological images," *CoRR*, vol. abs/2001.01599, 2020.

[9] N. Naik, A. Madani *et al.*, "Deep learning-enabled breast cancer hormonal receptor status determination from base-level h&e stains." *Nature communications*, vol. 11, p. 5727, 11 2020.

[10] M. Y. Lu, D. F. K. 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, pp. 555–570, 3 2021.

[11] Z. Shao, H. Bian *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classication," in *Neural Information Processing Systems*, 2021.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[13] O. Russakovsky, J. Deng *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 12 2015.

[14] D. A. Gutman, J. Cobb *et al.*, "Cancer digital slide archive: an informatics resource to support integrated in silico analysis of tcga pathology data." *Journal of the American Medical Informatics Association : JAMIA*, vol. 20, pp. 1091–8, 2013.

[15] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (tcga): an immeasurable source of knowledge." *Contemporary oncology (Poznan, Poland)*, vol. 19, pp. A68–77, 2015.

[16] L. A. Cooper, E. G. Demicco, J. H. Saltz, R. T. Powell, A. Rao, and A. J. Lazar, "Pancancer insights from the cancer genome atlas: the pathologist's perspective." *The Journal of pathology*, vol. 244, pp. 512–524, 4 2018.

[17] A. Riasatian, M. Babaie *et al.*, "Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides," *Medical Image Analysis*, vol. 70, p. 102032, 5 2021.

[18] T. Roetzer-Pejrimovsky, A.-C. Moser *et al.*, "The digital brain tumour atlas, an open histopathology resource," *Scientific Data*, vol. 9, p. 55, 2 2022.

[19] D. N. Louis, A. Perry *et al.*, "The 2021 who classification of tumors of the central nervous system: a summary," *Neuro-Oncology*, vol. 23, pp. 1231–1251, 8 2021.

[20] Y. Xiong, Z. Zeng *et al.*, "Nyströmformer: A nyström-based algorithm for approximating self-attention," *CoRR*, vol. abs/2102.03902, 2021.

[21] F. Küppers, J. Kronenberger, A. Shantia, and A. Haselhoff, "Multivariate confidence calibration for object detection," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[22] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[24] M. R. Zhang, J. Lucas, G. Hinton, and J. Ba, *Lookahead Optimizer: K Steps Forward, 1 Step Back*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[25] L. Jose, S. Liu *et al.*, "Artificial Intelligence–Assisted Classification of Gliomas Using Whole-Slide Images," *Archives of Pathology & Laboratory Medicine*, 11 2022. [Online]. Available: https://doi.org/10.5858/arpa.2021-0518-OA

[26] S. Liu, Z. Shah *et al.*, "Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning," *Scientific Reports*, vol. 10, 05 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-64588-y