LUPUS NEPHRITIS SUBTYPE CLASSIFICATION WITH ONLY SLIDE LEVEL LABELS

Amit Sharma[1], *Ekansh Chauhan*[1], *Megha S Uppin*[2], *Liza Rajasekhar*[2], *C V Jawahar*[1], *P K Vinod*[1]

[1]International Institute of Information Technology, Hyderabad, India
[2]Nizam's Institute Of Medical Sciences, Hyderabad, India

ABSTRACT

Lupus Nephritis classification has historically relied on labor-intensive and meticulous glomerular-level labeling of renal structures in whole slide images (WSIs). However, this approach presents a formidable challenge due to its tedious and resource-intensive nature, limiting its scalability and practicality in clinical settings. In response to this challenge, our work introduces a novel methodology that utilizes only slide-level labels, eliminating the need for granular glomerular-level labeling. A comprehensive multi-stained lupus nephritis digital histopathology WSI dataset was created from the Indian population, which is the largest of its kind. *LupusNet*, a deep learning MIL-based model, was developed for subtype classification of LN. The results underscore its effectiveness, achieving an AUC score of 91.0%, an F1-score of 77.3%, and an accuracy of 81.1% on our dataset in distinguishing membranous and diffused classes of LN.

***Index Terms*—** Lupus Nephritis, Weakly Supervised Learning, Whole Slide Image, Binary Classification

1. INTRODUCTION

Lupus Nephritis (LN) is one of the most severe manifestations of systemic lupus erythematosus (SLE), an autoimmune disease, due to its potential for severe renal damage and the intricate diagnostic and classification process. The complex nature of this disease is worsened by the substantial-high inter and intra-observer variability in histopathological renal biopsies [1]. As some classes of LN exhibit varying levels of aggressiveness, a precise classification of these classes becomes crucial in assessing fatality risks, predicting long-term prognosis, and determining an effective therapeutic approach. The microscopic nuances in these classes require experienced pathologists for accurate identification, and overlap cases or mixed patterns can further complicate it.

Deep learning has recently emerged as a powerful tool in medical AI and healthcare, revolutionizing various aspects of medicine, from diagnosis and treatment to drug discovery and patient monitoring [2]. Digital pathology has significantly advanced due to its capacity to extract intricate patterns and features from complex medical data [3, 4]. Improvements in image analysis have led to significant advancements in various aspects of renal pathology, including automated detection and classification of glomerular lesions [5, 6], and identification of interstitial fibrosis [7]. Advanced imaging techniques and molecular analyses may assist, but standardization and consensus in interpretation remain ongoing challenges.

Prior methods for LN classification typically adopt a two-stage process: first, classifying glomeruli types, and second, determining the LN class based on identified glomeruli types. This method heavily relies on glomeruli-level annotations and classification performance, constituting a distinct research domain [5, 8]. However, annotating numerous glomeruli is impractical for clinical settings, particularly for large WSIs up to 100k x 100k pixels. Additionally, classifying LN using WSIs faces challenges due to memory constraints, leading to proposed workarounds like patching and streaming [9, 10]. While previous efforts focused only on distinguishing between LN and non-LN without classifying subtypes [11], classifying LN subtypes proves challenging due to the presence of similar glomerular types across different subtypes. Notably, not all glomeruli contribute equally to the classification process, adding complexity to subtype differentiation. Although [12] aimed for an end-to-end LN subtype pipeline, their approach required manual glomeruli segmentation from WSIs and was evaluated on mice kidney biopsies, which differ substantially from human kidney biopsies in physiological and pathological features.

In contrast, our work simplifies this process by creating an end-to-end pipeline that does not necessitate reliance on glomeruli class labels at any intermediate stage. This method is called Multiple Instance Learning (MIL). MIL has been extensively explored for other areas of digital histopathology [9], but not much has been reported or explored in renal pathology. Our pipeline uses automatic glomeruli detection within WSIs, relying solely on slide-level labels.

While digital pathology has made strides, the LN classification research faces challenges such as access to the datasets and lack of consensus among medical professionals regarding its classification. In light of these considerations, the principal contributions of our work are as follows:

- **Dataset:** We focus on creating valuable dataset of LN to drive research (computational and medical) in kidney diseases. This dataset, featuring multi-stained whole slide images, stands as one of the largest collections for lupus

nephritis, which is a part of the consortium India Pathology Dataset (IPD) [1].

- **LN Classifier**: We introduce LupusNet, an interpretable MIL-based classification model for predicting LN subtype classes using only glomeruli patches in a WSI.

2. MATERIAL AND METHOD

2.1. Data Acquisition & Description

In this study, biopsy specimens of 166 patients (retrospective and prospective cases) in different subclasses (ranging from 1 to 6) of LN from the Nizam Institute of Medical Sciences (NIMS) in Hyderabad, India, were digitalized. A total of 540 WSIs were digitalized using the Morphle Optimus 6X Scanner, with each WSI captured at a maximum magnification of 40x and stored in the widely used TIFF format.

Within this repository of 540 WSIs, we identified four distinct categories of stained images, specifically Hematoxylin and Eosin (H&E), Periodic Acid-Schiff (PAS), methenamine silver Periodic Acid-Schiff (mt-PAS), and silver methenamine Periodic Acid-Schiff (sm-PAS). In this dataset, LN classes 4 (diffused proliferated) and 5 (membranous) exhibited the highest representation, with 62 and 53 cases, respectively.

Class 4 LN displays a varied glomerular appearance characterized by widespread inflammation, cellular proliferation, and diverse lesions, whereas class 5 LN demonstrates a uniform appearance due to immune complex deposition, resulting in a membranous pattern [13]. Consequently, our study focused primarily on observations and results for these two prominent LN class classifications using PAS-stained slides, highlighting carbohydrates, glycogen, and glycoproteins, aiding the identification of renal structures.

This India region-specific dataset is created to support global collaboration in lupus nephritis research. It helps add diversity to the other existing cohort, offering insights into potential regional and ethnic variations in the disease.

2.2. Methodology

We aim to learn a function that can predict the presence or absence of a condition within a WSI based on its constituent patches. Mathematically, this problem can be defined as follows: We are provided with a dataset containing pairs of bag-labels $\{(X_i, Y_i)\}_{i=1}^{D}$. Each X_i represents a collection of instances (patches) within a bag, and Y_i is the label assigned to that bag. Each bag X_i contains a variable number of instances $\{x_1, x_2, \ldots, x_N\} \in X_i$. These instances have labels $\{y_1, y_2, \ldots, y_N\}$ with $y_n \in \{0, 1\}$. However, the labels for individual instances are unknown during the training phase. If any instance in a bag is positive (i.e., belongs to the positive class), then the bag is considered positive. Conversely, if all the instances in a bag are negative, the bag is considered negative (i.e., belongs to the negative class).

$$Y_i = \begin{cases} 1, & \text{if } \exists x_n \in X_i \text{ such that } y_n = 1 \\ 0, & \text{otherwise} \end{cases}$$

Our methodology addresses the unique aspects of renal pathology, distinct from datasets like lung, brain, and breast tissues. Unlike these datasets, renal pathology primarily centers on a limited region of interest, mainly focusing on the glomerular area and allowing us to use recurrent networks. Glomeruli play a pivotal role in various renal diseases, including LN. Instead of providing MIL with all WSI patches, we exclusively use glomerular patches, enhancing precision by avoiding potential noise. Recognizing the laborious labeling at the glomerular area, we aimed to eliminate the need for intermediate glomerular-level labels; thus, opting for weakly supervised approaches is an appropriate option.

Our novel end-to-end MIL architecture for LN classification, LupusNet, works on raw glomerular patches in WSIs, with two key components: (a) Feature Extractor (f) and (b) Feature Aggregator (g), jointly trained. f transforms inputs into an information-rich feature space using a ResNet-50 network pre-trained on histopathology images [14]. We built on CLAM principles [15], which utilizes gated attention pooling and instance-level clustering to distinguish positive from negative samples. Gated attention, however, cannot fully exploit the uniformity of class 5 lupus nephritis glomeruli, hindering its ability to achieve optimal efficacy in capturing its consistent patterns. We hypothesize that adding contextual information among all glomeruli patches will improve the performance. To address this, we integrate self-attention and Bi-LSTM into the MIL framework, enhancing contextual understanding among instances (patches) in a WSI.

Suppose, in a WSI bag X, we have N glomerular patches, and the Feature Extractor f transforms each image $x_n \in \mathbb{R}^{224\times224\times3}$ into a h vector of dimension $d \in \mathbb{R}^{1\times d}$. For N such images, we obtain a matrix $H \in \mathbb{R}^{N\times d}$ (eq: 1). Our feature aggregator can further be divided into three branches: (1) Gated Attention Pooling, (2) Self-Attention + LSTM and (3) Instance-level Clustering. In Branch 1, the gated attention block assigns attention scores $A^g = \{a_1^g, a_2^g, \ldots, a_N^g\} \in \mathbb{R}^{1\times N}$ to every instance (eq: 2), followed by instance-level clustering using A^g as pseudo labels for confident instances (Branch 3).

$$H = f(X; \Theta) \quad \text{where } H = \{h_1, h_2, \ldots, h_N\} \tag{1}$$

$$a_k^g = \frac{W_c^T(\tanh(W_a h_k^T) \odot \sigma(W_b h_k^T))}{\sum_{j=1}^{N} W_c^T(\tanh(W_a h_j^T) \odot \sigma(W_b h_j^T))} \tag{2}$$

$$C^g = \sum_{k=1}^{N} a_k^g h_k \tag{3}$$

[1] https://hai.iiit.ac.in/ipd/

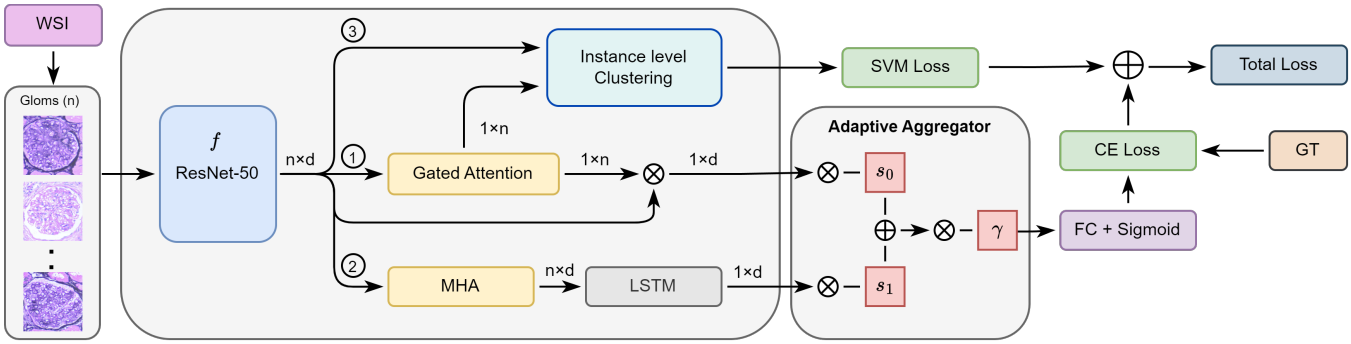


Fig. 1: **LupusNet**: Proposed architecture for our one-step lupus nephritis classification. The gated attention will capture the individual importance of each glomeruli, and multi-head attention (MHA) will capture the relative context between the glomeruli to provide a comprehensive view of the whole bag before a label for the bag is predicted

where W_a, W_b and W_c are trainable parameters, a_k^g can be supposed as positive probability of instances. σ represents sigmoid function and $\odot$ represents element-wise multiplication. C^g is the output context vector of Branch 1 (eq: 3).

In Branch 2, initially, H goes to MHA, yielding contextualized output among instances (A^s). Self-attention (eq: 4) enables context consideration between every instance pair, and the multi-head mechanism focuses on modeling various such contextual relationships and dependencies among instances. The attention scores obtained from different heads, n_h is a total number of heads, are concatenated, and a linear transformation is applied to ensure that the resulting shape matches the input, resulting in $\mathbb{R}^{n\times d}$ (eq: 5). To further process this contextualized information, we employ LSTM, which uses gating mechanisms and outputs the hidden layer of the last time step $\mathbb{R}^{1\times d}$.

$$a_i^{self} = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \tag{4}$$

$$A^s = (a_1^{self} \oplus a_2^{self} \oplus \ldots \oplus a_{n_h}^{self}) W_o \tag{5}$$

where $Q_i = HW_i^Q, K = HW_i^K$ and $V = HW_i^V$ with W_i^Q, W_i^K and W_i^V trainable parameters for i^th head and W_o is also trainable parameter and used to linear transform the concatenated output from multiple heads. d_k is used for scaling to prevent the dot product from becoming too large, and C^s is the output context vector of Branch 2 after applying bi-LSTM on A^s.

Furthermore, we use softmax normalized learnable parameters s_0 and s_1 to aggregate contributions from each pipeline's output adaptively. A scaling learnable parameter γ finetunes the overall merged output contribution, introducing an additional degree of freedom in the weighting process (eq: 6). Inspired by attention principles, this approach facilitates contextual understanding and dynamic weighting for effective information extraction from both branches. It draws parallels from multiple layer fusion of contextual embeddings in ELMO during downstream task [16].

$$logits = \gamma\left(s_0 C^g + s_1 C^s\right) \tag{6}$$

After applying the adaptive aggregation method, a binary classifier with a single neuron and a sigmoid activation function is used to estimate the probabilities, y, of a slide being positive. Subsequently, binary cross-entropy loss is computed at the slide level (Branch 1 and 2) while Smooth SVM loss [15] is applied for instance-level clustering (Branch 3). The Smooth SVM loss, a generalization of traditional cross-entropy classification loss, accommodates diverse margin values and temperature scaling strategies, providing flexibility to mitigate overfitting. The rationale for choosing Smooth SVM loss lies in addressing potential noise in pseudo-labels, offering robustness in the presence of uncertainties. The total loss, as per Equation 7, is calculated as the weighted sum of both losses, where H' and $A^{g\prime}$ are the subset of H and A^g respectively, $\hat{y}$ is ground truth and β is a hyper-parameter.

$$J = \beta\,\text{BCE}(y, \hat{y}) + (1-\beta)\,\text{Smooth-SVM}(H', A^{g\prime}) \tag{7}$$

3. RESULTS

3.1. Experimentation Details

For a robust evaluation of classification performance, we employed 10-fold cross-validation. Using a YOLO-based glom detector with 6000×6000 patches generated through the [15] pipeline, all methods were implemented in PyTorch and trained on a single NVIDIA RTX 3080ti GPU. The training involved 50-200 epochs with early stopping. $n_h = 4$, $\beta = 0.8$, a Bi-LSTM hidden dimension of 512, and Adam optimizer with $lr = 1e5$. Batch size is set to 1 for all models.

3.2. Quantitative analysis

We established baselines by initiating natural and intuitive experiments, employing a pseudo-labeling approach to address the lack of glomerulus-level labels, and assigning all

glomeruli the same label as their corresponding whole slide labels. Deep learning models, including AlexNet, ResNet-18, ResNet-101, ResNet-152, DenseNet-121, SqueezeNet, and ShuffleNet (pretrained or finetuned on ImageNet), were applied to evaluate this strategy, with the best-performing ResNet-101's result detailed in Table 1. However, the experiments revealed a significant challenge: The potential inconsistency in labeling similar glomeruli types between lupus class 4 and class 5 cases. This ambiguity has impacted model performance, highlighting the need for exploring alternative techniques, especially in the absence of fully supervised datasets with precise labels.

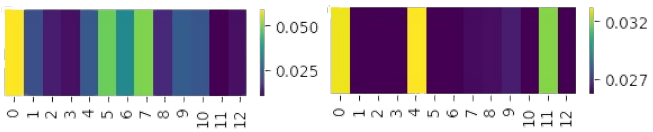
Model	Input	Test AUC	Test F1	Test ACC
ResNet-101	GP	52.88	44.12	53.23
CLAM-SB	AP	57.65	52.22	52.43
CLAM-SB	GP	86.00	72.80	75.55
LupusNet (Ours)	**GP**	**91.00**	**77.30**	**81.11**

Table 1: Comparing our proposed model (LupusNet) with baselines, averaging results (in %) over 10-fold cross-validation on test cohort. Input types include GP (Only Glomeruli Patches) and AP (All Patches).

Afterward, we employed a weakly supervised CLAM single-branched variant (CLAM-SB) and our proposed LupusNet on the in-house dataset. Results are presented for both scenarios, wherein we either input all the WSI patches or just the glomeruli patches. The conclusive findings, as shown in Table 1, demonstrate that LupusNet outperforms all baseline models. We can empirically observe a significant performance improvement when only glomeruli patches are provided, consequently reducing noise to the CLAM-SB model. LupusNet, outperforming CLAM-SB (GP), demonstrates a significant F1-score improvement for class 5 LN (65.17% to 77.03%), highlighting its efficacy in reducing false positives and enhancing precision.

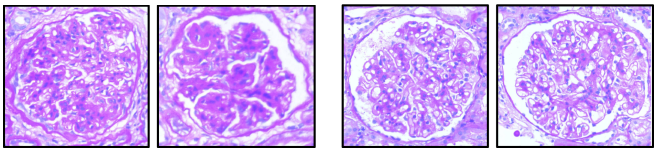
3.3. Qualitative analysis

Figure 2 represents the interpretability of a test sample, which contains multiple glomerulus images. It uses attention weight distributions using heatmaps from the two branches of our model. Here, MHA (Branch 2) focuses on glomeruli patterns, prioritizing context at the WSI level. This contextualization is crucial for capturing the uniform membranous patterns of class 5 LN (fig: 3) and thus highlights the importance of MHA for improved classification performance compared to relying only on gated attention (Branch 1), which exhibits a diverse focus necessary for class 4 LN, which shows diffuse proliferation pattern.



(a) Branch-1 (b) Branch-2

Fig. 2: Attention weights of both branches for a class 5 sample (a) Gated Attention and (b) Multi-head Attention



(a) Class 4 glomeruli (b) Class 5 glomeruli

Fig. 3: Comparison of visual features between subtype samples. (a) involves proliferative changes in the glomeruli, whereas (b) shows thickening of the glomerular basement membrane

4. DISCUSSION AND CONCLUSION

Our study introduces LupusNet, a MIL-based model for lupus nephritis classification that uses only slide-level labels, eliminating the necessity for glomeruli-level labels. Other MIL-based models incorporating transformers [17] were deemed sub-optimal for our case due to the limited data size. However, we recognized the need for self-attention among glomeruli for context inclusion. Therefore, our work includes this aspect without increasing network complexity while retaining interpretability for pathologists. This study is a valuable reference for pathologists to address inter/intra-variability. Additionally, it holds significance for researchers engaged in the study of other diverse renal diseases beyond the specific focus on LN. It also contributes to renal pathology research by creating a digital whole slide image dataset. While LupusNet exhibits promising results, there are areas for potential improvement. Our future work involves improving glomeruli detection models and feature aggregators, which could extract even better contextual information from glomeruli. This study also

Data Availability Statement: The dataset generated and/or analyzed during the current study is available from the authors on reasonable request within the terms of the data use agreement and compliance with ethical and legal requirements.

5. COMPLIANCE WITH ETHICAL STANDARDS

Procedures in studies with human participants adhered to ethical standards set by institutional (NIMS) and/or national research committees (ICMR).

6. ACKNOWLEDGMENTS

We acknowledge IHub-Data, IIIT Hyderabad (H1-002) for financial assistance. We also thank Dr. Manasa Kondamadugu for project coordination, Ms. Ramya Alugam, and Mr. Akula Rajesh Goud for data digitalization and organization.

7. REFERENCES

[1] Shobha Dasari, Ashish Chakraborty, Luan Truong, and Chandra Mohan, "A systematic review of interpathologist agreement in histologic classification of lupus nephritis," *Kidney International Reports*, 2019.

[2] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, and et al., "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 1, 2018.

[3] Benjamin Wu and Gilbert Moeckel, "Application of digital pathology and machine learning in the liver, kidney and lung diseases," *Journal of Pathology Informatics*, 2023.

[4] Alhassan Ali Ahmed, Mohamed Abouzid, and Elżbieta Kaczmarek, "Deep learning approaches in histopathology," *Cancers*, vol. 14, no. 21, pp. 5264, 2022.

[5] Susan M. Sheehan and Ron Korstanje, "Automatic glomerular identification and quantification of histological phenotypes using image analysis and machine learning," *American Journal of Physiology-Renal Physiology*, vol. 315, no. 6, pp. F1644–F1651, 2018.

[6] Brandon Ginley, Brendon Lutnick, Kuang-Yu Jen, Agnes B Fogo, Sanjay Jain, Avi Rosenberg, Vighnesh Walavalkar, Gregory Wilding, John E Tomaszewski, Rabi Yacoub, et al., "Computational segmentation and classification of diabetic glomerulosclerosis," *Journal of the American Society of Nephrology: JASN*, 2019.

[7] Yi Zheng, Clarissa A. Cassol, Saemi Jung, Divya Veerapaneni, Vipul C. Chitalia, Kevin Y.M. Ren, Shubha S. Bellur, Peter Boor, Laura M. Barisoni, Sushrut S. Waikar, and et al., "Deep-learning–driven quantification of interstitial fibrosis in digitized kidney biopsies," *The American Journal of Pathology*, 2021.

[8] Zhaohui Zheng, Xiangsen Zhang, Jin Ding, Dingwen Zhang, Jihong Cui, Xianghui Fu, Junwei Han, and Ping Zhu, "Deep learning-based artificial intelligence system for automatic assessment of glomerular pathological findings in lupus nephritis," *Diagnostics*, 2021.

[9] Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, 2019.

[10] Hans Pinckaers, Bram Van Ginneken, and Geert Litjens, "Streaming convolutional neural networks for end-to-end learning with multi-megapixel images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1581–1590, 2020.

[11] Da-Cheng Wang, Wang-Dong Xu, Shen-Nan Wang, Xiang Wang, Wei Leng, Lu Fu, Xiao-Yan Liu, Zhen Qin, and An-Fang Huang, "Lupus nephritis or not? a simple and clinically friendly machine learning pipeline to help diagnosis of lupus nephritis," *Inflammation Research*, vol. 72, no. 6, pp. 1315–1324, 2023.

[12] Pietro Antonio Cicalese, Aryan Mobiny, Zahed Shahmoradi, Xiongfeng Yi, Chandra Mohan, and Hien Van Nguyen, "Kidney level lupus nephritis classification using uncertainty guided bayesian convolutional neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 315–324, 2020.

[13] Jan J Weening, Vivette D D'agati, Melvin M Schwartz, Surya V Seshan, Charles E Alpers, Gerald B Appel, James E Balow, JANA Bruijn, Terence Cook, Franco Ferrario, et al., "The classification of glomerulonephritis in systemic lupus erythematosus revisited," *Kidney international*, vol. 65, no. 2, pp. 521–530, 2004.

[14] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira, "Benchmarking self-supervised learning on diverse pathology datasets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3344–3354.

[15] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, June 2021.

[16] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep contextualized word representations," in *Human Language Technologies*. 2018, Association for Computational Linguistics.

[17] Zhucheng Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang, "Transmil: Transformer based correlated multiple instance learning for whole slide image classication," in *Neural Information Processing Systems*, 2021.