Lupus Nephritis Subtype Classification with only Slide Level labels

Amit Sharma[*1] AMIT.S@RESEARCH.IIIT.AC.IN
[1] *Center for Visual Information Technology, International Institute of Information Technology, Hyderabad, India*
Ekansh Chauhan[*1] EKANSH.CHAUHAN@RESEARCH.IIIT.AC.IN
Megha S Uppin[2] MEGHA_HARKE@YAHOO.CO.IN
[2] *Department of Pathology, Nizam's Institute Of Medical Sciences, Hyderabad, India*
Liza Rajasekhar[3] LIZARAJASEKHAR@GMAIL.COM
[3] *Department of Clinical Immunology and Rheumatology, Nizam's Institute Of Medical Sciences, Hyderabad, India*
C V Jawahar[1] JAWAHAR@IIIT.AC.IN
P K Vinod[4] VINOD.PK@IIIT.AC.IN
[4] *Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, 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 the 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.

Keywords: 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 (Dasari et al., 2019). 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.

[*] Contributed equally

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 (Rajkomar et al., 2018). Digital pathology has significantly advanced due to its capacity to extract intricate patterns and features from complex medical data (Wu and Moeckel, 2023; Ahmed et al., 2022). Improvements in image analysis have led to significant advancements in various aspects of renal pathology, including automated detection and classification of glomerular lesions (Sheehan and Korstanje, 2018; Ginley et al., 2019), and identification of interstitial fibrosis (Zheng et al., 2021a). Advanced imaging techniques and molecular analyses may assist, but standardization and consensus in interpretation remain ongoing challenges.

Traditional LN classification follows a two-step process: first identifying glomeruli types, then classifying LN based on these types, heavily dependent on detailed glomeruli annotations (Sheehan and Korstanje, 2018; Zheng et al., 2021b). Yet, annotating glomeruli on large-scale WSIs is impractical in clinical settings due to their massive size and memory limitations, leading to patching and streaming solutions (Campanella et al., 2019; Pinckaers et al., 2020). Previous studies mainly differentiated LN from non-LN, not addressing subtype classification (Wang et al., 2023), which is complicated by similar glomerular types across subtypes and the unequal contribution of glomeruli to classification. (Cicalese et al., 2020) proposed an end-to-end LN subtype classification method, but it required manual segmentation on mice biopsies, not directly applicable to human samples due to differences in physiology and pathology.

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. Multiple Instance Learning (MIL) has been extensively explored for other areas of digital histopathology (Campanella et al., 2019), but not much has been reported or explored in renal pathology.

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:

- We focus on creating a 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].

- We also introduce a novel architecture, LupusNet, an explainable MIL-based model that significantly improves LN subtype classification by integrating Gated and Multi-Head Attention, underscoring the critical requirement to learn the morphological differences between LN class 4 & 5.

- To the best of our knowledge, we present the first end-to-end pipeline for LN subtype classification, designed to achieve efficient diagnosis and classification by relying only on slide-level labels, easing clinical workload and facilitating practical integration.

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

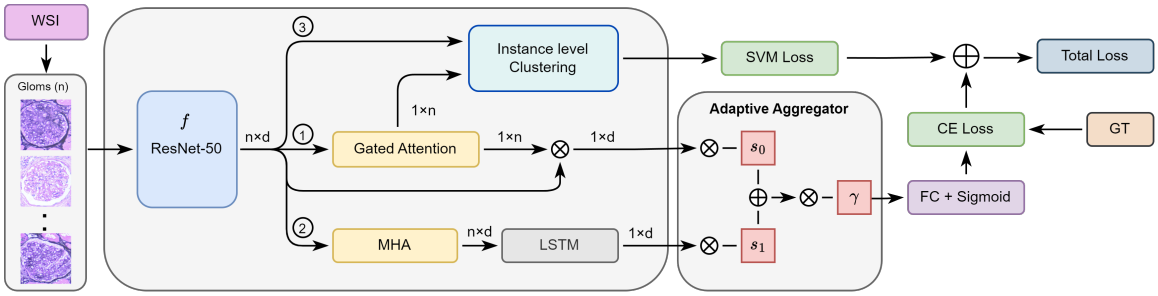


Figure 1: **LupusNet**: Proposed architecture for our lupus nephritis classifier. Gated attention identifies each glomerulus's importance, while multi-head attention (MHA) discerns their contextual relationships.

2. Materials 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, there are 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 (Weening et al., 2004). 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 belongs to the positive class, then the bag is considered positive. Conversely, if all the instances in a bag belong to the negative class, the bag is considered negative.

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

Our methodology extends this formulation to multiple positive classes for subtype LN classification. Unlike lung, brain, and breast datasets, renal pathology primarily focuses on a limited region of interest, particularly the glomerular area, 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, extracted using fine-tuned YOLOv4 model (Hemmatirad et al., 2023), 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 (Kang et al., 2023). We built on CLAM principles (Lu et al., 2021), 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).

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

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

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

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 $V = HW_i^V$, for the i^{th} head, are derived using trainable parameters W_i^Q, W_i^K, W_i^V, and W_o linearly transforms the multi-head outputs. d_k is used for scaling to prevent the dot product from becoming too large, and C^s is the bi-LSTM processed output context vector from Branch 2 on A^s.

Furthermore, we use softmax normalized learnable parameters s_0 and s_1 to adaptively aggregate contributions from each pipeline's output. A scaling learnable parameter γ fine-tunes 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 (Peters et al., 2018).

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

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 (Lu et al., 2021) 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. All methods were implemented in PyTorch and trained on a single NVIDIA RTX 3080ti GPU. The patch size for YOLOv4-based glom detector was set to 6000×6000, and the MIL 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 $lr = 1e4$. Batch size is set to 1 for all models.

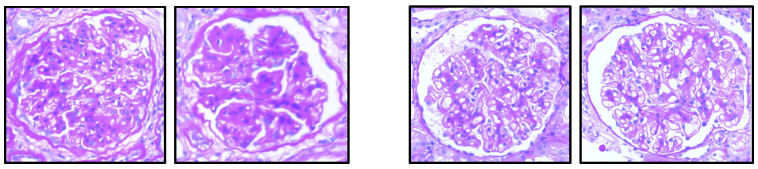
Table 1: Comparing our proposed model (LupusNet) with baselines, averaging results (in %) over 10-fold cross-validation on test cohort. Input types include GP (Only Glomeruli Patches) and AP (All Patches).

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**

3.2. Quantitative analysis

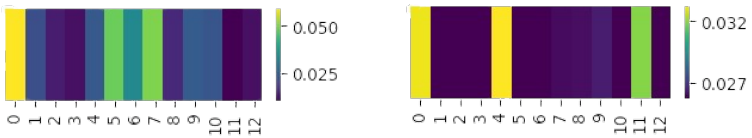
We established baselines using a pseudo-labeling approach for lack of detailed glomerulus-level labels by assigning whole slide labels to all glomeruli and tested models like AlexNet, ResNet, and DenseNet, with ResNet-101 performing best (1). These experiments underscored the challenge of label inconsistency among similar glomeruli in lupus classes 4 and 5, affecting model accuracy and emphasizing the need for alternative methods in the absence of precisely labeled datasets.

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. Additional observation showed LupusNet outperforming CLAM-SB (GP), by a significant F1-score improvement for class 5 LN (65.17% to 77.03%), highlighting its efficacy in distinguishing the two classes and reducing false positives and enhancing precision.



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

Figure 2: Comparison of visual features between subtype samples. (a) involves proliferative changes in the glomeruli, whereas (b) shows thickening of the glomerular basement membrane



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

Figure 3: Attention weights of both branches for a class 5 sample (a) Gated Attention and (b) Multi-head Attention

Table 2: Ablation study with module variations.
L=LSTM; **G**=Gated Attention; **C**=Clustering (Instance level)

Model	Test AUC	Test F1	Test ACC
LSTM	64.00	56.27	60.00
L+G	81.65	67.00	71.11
L+G+C	85.00	74.91	77.78
LupusNet (Ours)	**91.00**	**77.30**	**81.11**

3.3. Qualitative analysis

Figure 3 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 (Figure: 2) 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.

4. Ablation Study

In our ablation study, we methodically introduced various architectural components to evaluate their individual and combined effects on the model's performance. Beginning with a basic LSTM model as our starting point, we then integrated Gated Attention and Instance-level clustering. Each addition led to noticeable improvements in performance, as shown in Table 2, with our final model, LupusNet, outperforming all other configurations. This step-by-step process helped us identify the specific contributions of each component to the model's overall effectiveness in classifying two LN classes. We further optimized LupusNet by adjusting the learning rates and the number of Multi-Head Attention (MHA) blocks (Figure 4).

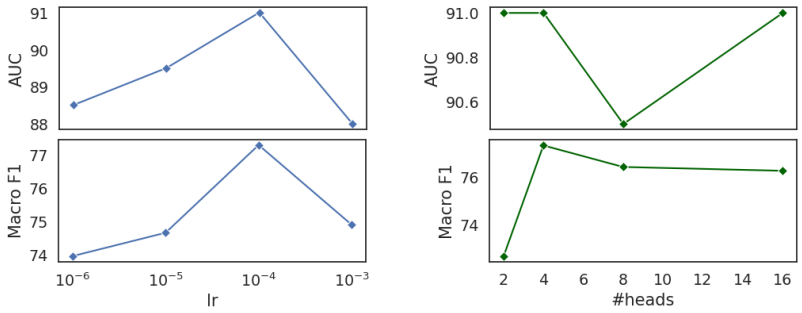


Figure 4: Hyperparameter tuning of LupusNet based on the optimized value of learning rate (left) and number of attention heads (right)

5. 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. Due to the limited data size, other MIL-based models incorporating transformers (Shao et al., 2021) were deemed sub-optimal for our case. 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 studying 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.

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.

Compliance with Ethical Standards

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

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.

References

Alhassan Ali Ahmed, Mohamed Abouzid, and Elżbieta Kaczmarek. Deep learning approaches in histopathology. *Cancers*, 14(21):5264, 2022. ISSN 2072-6694. doi: 10.3390/cancers14215264.

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.

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*, 25(2):315–324, 2020.

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.

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.

Kimia Hemmatirad, Morteza Babaie, Jeffrey Hodgin, Liron Pantanowitz, and HR Tizhoosh. An investigation into glomeruli detection in kidney h&e and pas images using yolo. *arXiv preprint arXiv:2307.13199*, 2023.

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*, pages 3344–3354, 2023.

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*, 5(6):555–570, June 2021. doi: 10.1038/s41551-020-00682-w. URL https://doi.org/10.1038/s41551-020-00682-w.

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

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*, 44(3):1581–1590, 2020.

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*, 1(1), 2018.

ISSN 2398-6352. doi: 10.1038/s41746-018-0029-1. URL https://dx.doi.org/10.1038/s41746-018-0029-1.

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. URL https://api.semanticscholar.org/CorpusID:235294100.

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*, 315(6):F1644–F1651, 2018. ISSN 1931-857X. doi: 10.1152/ajprenal.00629.2017.

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*, 72(6):1315–1324, 2023. ISSN 1023-3830. doi: 10.1007/s00011-023-01755-7.

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*, 65(2):521–530, 2004.

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

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*, 2021a.

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*, 2021b.