

Clinical-Injection Transformer with Domain-Adapted MAE for Lupus Nephritis Prognosis Prediction

Yuewen Huang¹[0009-0005-1255-3993], Zhitao Ye²[0009-0004-7103-8214],
Guangnan Feng¹[0000-0002-1382-280X], Fudan Zheng^{1*}[0000-0002-9664-012X],
Xia Gao^{2*}[0000-0003-4412-5166], and Yutong Lu^{1*}[0000-0001-5315-3375]

¹ Sun Yat-sen University, Guangzhou, China

² Department of Nephrology, Guangzhou Women and Children's Medical Center,
Guangzhou Medical University, Guangzhou, China
huangyw66@mail2.sysu.edu.cn, yezht@foxmail.com, fenggn7@mail.sysu.edu.cn,
zhengfd5@mail.sysu.edu.cn, gaoksiagz@vip.163.com,
luyutong@mail.sysu.edu.cn

Abstract. Lupus nephritis (LN) is a severe complication of systemic lupus erythematosus that affects pediatric patients with significantly greater severity and worse renal outcomes compared to adults. Despite the urgent clinical need, predicting pediatric LN prognosis remains unexplored in computational pathology. Furthermore, the only existing histopathology-based approach for LN relies on multiple costly staining protocols and fails to integrate complementary clinical data. To address these gaps, we propose the first multimodal computational pathology framework for three-class treatment response prediction (complete remission, partial response, and no response) in pediatric LN, utilizing only routine PAS-stained biopsies and structured clinical data. Our framework introduces two key methodological innovations. First, a Clinical-Injection Transformer (CIT) embeds clinical features as condition tokens into patch-level self-attention, facilitating implicit and bidirectional cross-modal interactions within a unified attention space. Second, we design a decoupled representation-knowledge adaptation strategy using a domain-adapted Masked Autoencoder (MAE). This strategy explicitly separates self-supervised morphological feature learning from pathological knowledge extraction. Additionally, we introduce a multi-granularity morphological type injection mechanism to bridge distilled classification knowledge with downstream prognostic predictions at both the instance and patient levels. Evaluated on a cohort of 71 pediatric LN patients with KDIGO-standardized labels, our method achieves a three-class accuracy of 90.1% and an AUC of 89.4%, demonstrating its potential as a highly accurate and cost-effective prognostic tool.

Keywords: Lupus Nephritis · Multimodal Fusion · Clinical-Injection Transformer · Representation-Knowledge Decoupling · Computational Pathology.

1 Introduction

Lupus nephritis (LN) affects 50–82% of pediatric systemic lupus erythematosus (SLE) patients [1]—substantially more than adults (20–40%) [2]—with greater disease severity and worse long-term renal outcomes [3]. Yet childhood-onset SLE itself is exceptionally rare, with an incidence of only 0.3–0.9 per 100,000 children-years—approximately one-sixth of the adult rate [4, 5]. When further restricted to biopsy-confirmed cases with longitudinal follow-up and digitized pathology, even the largest multicenter pediatric LN cohorts comprise only about 300 patients [3]. Besides, its treatment response varies significantly—complete remission (CR), partial response (PR), or no response (NR)—and early identification is critical.

With the growing adoption of deep learning in medical image analysis, existing approaches for LN prognosis fall into two disjoint tracks: machine learning models based on clinical biomarkers have shown promise for treatment outcome and relapse prediction [6], but discard the rich morphological information in renal biopsies; on the other hand, the only histopathology-based approach [7] requires four costly staining protocols without clinical data integration. No method combines histopathology with clinical data for LN prognosis, and pediatric LN [1] lacks any biopsy-image-based prediction approach.

Meanwhile, multimodal fusion of histopathology and clinical or genomic data has emerged as a key direction. However, methods such as MCAT [8], Surv-Path [9], CMTA [10], and HEALNet [11] typically require complex dual-stream architectures prone to overfitting on small cohorts. This motivates our Clinical-Injection Transformer, which injects clinical features as condition tokens into unified self-attention for parameter-efficient cross-modal interaction.

In addition, self-supervised pretraining, such as Masked Autoencoder (MAE)-based approaches [12] have shown effectiveness for medical image classification [13] and pathology representation learning [14]. These motivate us to explore MAE-based self-supervised pretraining on glomerulus patches to learn domain-specific representations.

In this work, we propose a multimodal framework for pediatric LN three-class prognosis prediction. Our contributions are:

1. The **first multimodal computational pathology framework** for LN treatment response prediction from routine single-stain histopathology with clinical data, under KDIGO-standardized three-class criteria (CR/PR/NR).
2. **Clinical-Injection Transformer (CIT)**, which injects clinical features as condition tokens into patch-level self-attention, enabling bidirectional cross-modal interaction—simpler and more parameter-efficient than cross-attention [8] or separate fusion modules [9].
3. **Decoupled representation-knowledge adaptation** that separates self-supervised feature learning from morphological classification, preserving prognostically relevant features (+7.1% over DINOv2 [15]).
4. **Multi-granularity morphological type injection** bridging distilled classification knowledge with prognosis at patch and patient levels (+2.3% Acc, +5.3% M-F1).

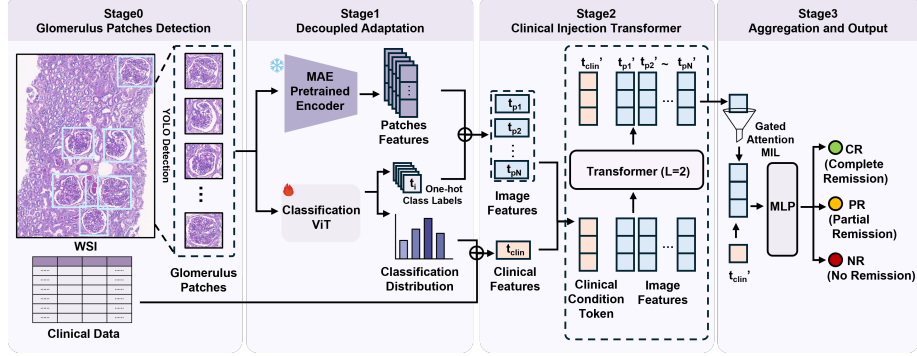


Fig. 1. Overview of the proposed framework.

2 Method

2.1 Framework Overview

Our framework (Fig. 1) consists of four stages: (0) automated glomerulus detection from PAS-stained WSIs, producing N cropped glomerulus patches per patient; (1) decoupled representation-knowledge adaptation—a ViT-B/16 encoder is pretrained via MAE [12] on 4,826 glomerulus patches, then serves two paths: the *frozen* pretrained encoder extracts patch representations $\{\mathbf{x}_i \in \mathbb{R}^{768}\}_{i=1}^N$ (representation path), while a *finetuned* copy with a classification head produces discrete morphological type labels $\{t_i\}_{i=1}^N$ (knowledge path), which are combined with clinical features $\mathbf{c} \in \mathbb{R}^{d_c}$; (2) Clinical-Injection Transformer fuses patch features (with multi-granularity type injection) and clinical condition token via unified self-attention for bidirectional cross-modal interaction; (3) gated attention-based multiple instance learning (MIL) aggregation pools enriched patch tokens into a patient-level representation, which is concatenated with the enriched clinical token and classified into CR/PR/NR via a classification head.

2.2 Decoupled Representation-Knowledge Adaptation

Finetuning for morphological classification narrows the representation space toward category-level discrimination, at the cost of subtle morphological cues—such as texture variations and structural irregularities—that are not rewarded by the classification objective yet may prove essential for predicting treatment response. We address this by *decoupling* domain adaptation into two complementary paths:

Representation path: self-supervised pretraining. A ViT-B/16 encoder is pretrained using masked autoencoding [12] on approximately 5,000 glomerulus patches, including publicly available annotated glomerular data [16] with an in-house collection from 137 patients at a collaborating center. This expanded pretraining corpus ensures exposure to diverse glomerular morphologies across institutions and staining variations. Following standard MAE, 75%

of patches are masked and the encoder learns to reconstruct the masked regions, capturing broad tissue textures and structural patterns. The pretrained encoder is *frozen* for downstream feature extraction, preserving its rich, task-agnostic representations.

Knowledge path: supervised morphological classification. A copy of the pretrained encoder is finetuned for 5-class glomerular morphological classification (achieving 92% accuracy): mesangial proliferative, normal, endocapillary proliferative, crescentic, and sclerotic. Rather than using the finetuned encoder for feature extraction, we *distill* its acquired pathological knowledge into discrete type labels, which are subsequently injected into the downstream model via multi-granularity type injection (Sec. 2.4).

Our ablation (Sec. 3.3) confirms this effect: frozen self-supervised features outperform finetuned features by +6.0% Acc, indicating that classification-oriented adaptation indeed discards prognostically relevant information. By separating representation learning from knowledge extraction, we preserve the representational richness of self-supervised features while still leveraging pathological semantics through structured knowledge injection.

2.3 Clinical-Injection Transformer (CIT)

The core of our fusion module integrates clinical features with patch-level image features through a unified self-attention mechanism.

Clinical condition token. Given clinical features $\mathbf{c} \in \mathbb{R}^{d_c}$, we project them into the patch feature space:

$$\mathbf{z}_{\text{cond}} = \text{LayerNorm}(\text{ReLU}(\mathbf{W}_c \mathbf{c} + \mathbf{b}_c)) \in \mathbb{R}^{d_h} \quad (1)$$

where $d_h=256$ is the hidden dimension. This condition token represents the patient’s clinical context.

Unified self-attention. Patch features $\{\mathbf{x}_i\}_{i=1}^N$ are projected to $\{\mathbf{z}_i \in \mathbb{R}^{d_h}\}$ and concatenated with the condition token:

$$\mathbf{Z}^{(0)} = [\mathbf{z}_{\text{cond}}; \mathbf{z}_1; \mathbf{z}_2; \dots; \mathbf{z}_N] \in \mathbb{R}^{(N+1) \times d_h} \quad (2)$$

This sequence is processed by $L=2$ Transformer encoder layers (4 attention heads per layer):

$$\mathbf{Z}^{(l)} = \text{TransformerLayer}(\mathbf{Z}^{(l-1)}), \quad l = 1, \dots, L \quad (3)$$

By placing clinical and image tokens in the *same* self-attention sequence, each clinical feature naturally attends to all patch features and vice versa, enabling **implicit bidirectional interaction** without explicit cross-attention modules. The entire CIT module contains approximately 0.56M parameters, substantially fewer than dual-stream cross-attention architectures, mitigating overfitting on small datasets.

Attention-based MIL aggregation. After the Transformer, patch representations are aggregated into a patient-level image representation via gated

attention-based MIL pooling [17]:

$$a_i = \frac{\exp(\mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{z}_i^{(L)}) \odot \sigma(\mathbf{U}\mathbf{z}_i^{(L)})))}{\sum_j \exp(\mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{z}_j^{(L)}) \odot \sigma(\mathbf{U}\mathbf{z}_j^{(L)})))}, \quad \mathbf{h}_{\text{img}} = \sum_{i=1}^N a_i \mathbf{z}_i^{(L)} \quad (4)$$

The final patient representation $\mathbf{h} = [\mathbf{h}_{\text{img}}; \mathbf{z}_{\text{cond}}^{(L)}]$ concatenates the aggregated image representation with the enriched clinical token, followed by a classification head.

2.4 Multi-Granularity Morphological Type Injection

We bridge the knowledge path and downstream prognosis through multi-granularity feature injection, providing the model with explicit morphological semantics at complementary hierarchical levels:

Patch-level injection. Each patch’s morphological type label (from the knowledge path) is encoded as a one-hot vector $\mathbf{t}_i \in \{0, 1\}^K$ ($K=5$ classes) and concatenated with the image feature: $\tilde{\mathbf{x}}_i = [\mathbf{x}_i; \mathbf{t}_i] \in \mathbb{R}^{d+K}$. This informs the Transformer of each patch’s morphological identity.

Patient-level injection. The type distribution $\mathbf{d} = \frac{1}{N} \sum_i \mathbf{t}_i \in [0, 1]^K$ is concatenated with clinical features: $\tilde{\mathbf{c}} = [\mathbf{c}; \mathbf{d}] \in \mathbb{R}^{d_c+K}$, encoding overall lesion composition (e.g., proportion of sclerotic or crescentic glomeruli).

Unlike multi-granularity approaches operating within a single modality [18], our injection spans *across* modalities, transferring distilled classification knowledge to prognosis at corresponding hierarchical levels. Patch-level morphological identity constrains instance-level attention, while patient-level composition informs global clinical interaction; these two granularities are designed to act jointly and provide complementary signals. To regularize training under severe class imbalance, we apply manifold Mixup [19] at the patient-level representation space, where interpolation in the learned embedding captures semantically meaningful morphological summaries.

3 Experiments

3.1 Dataset and Setup

Ethics approval and data collection. This study was approved by the Institutional Review Board of Guangzhou Women and Children’s Medical Center. We collected PAS-stained renal biopsy WSIs and clinical records from 180 pediatric LN patients at a tertiary children’s hospital. A YOLO-based detection model [16] extracts glomeruli (99% sensitivity). After inclusion criteria (complete biopsy digitization, sufficient clinical measurements, evaluable 12-month KDIGO outcome), 71 patients (CR=49, PR=10, NR=12) with 2,925 glomerulus patches (448×448, mean 31.5/patient) formed the final cohort.

Clinical features. The 59 clinical features span: (i) 3 demographics (age, sex, disease duration); (ii) 22 baseline (0-month) features across urinalysis (urine

protein, protein-to-creatinine ratio, urine RBC/WBC), blood routine (hemoglobin, platelets, WBC), renal biochemistry (serum creatinine, eGFR, BUN, albumin), immunological markers (C3, C4, ANA, anti-dsDNA), anthropometrics (BMI, blood pressure), and SLEDAI-2K; (iii) 34 three-month features (17 repeated measures + 17 deltas). The 3-month features and deltas are temporally distinct from the 12-month outcome. Missing values are handled by MICE [20], applied exclusively to pre-outcome features and strictly excluding outcome labels.

Cross-validation and training. All experiments use 5-fold cross-validation with 3 random seeds. To prevent data leakage, MICE imputation is performed independently within each fold—the imputer is fitted solely on the training split and applied to transform the unseen validation/test splits. The morphological classification model (knowledge path, Sec. 2.2) is retrained per fold using only training and validation splits; its discrete predictions on the held-out test patients serve as type labels $\{t_i\}_{i=1}^N$ for downstream prognosis. Outcome labels follow KDIGO [21]: CR requires proteinuria normalization and stable creatinine; PR requires $\geq 50\%$ proteinuria reduction; NR otherwise. We evaluate baseline-only (0m, 25 dims) and baseline+3-month (0m+3m, 59 dims). Models use AdamW (lr=1e-3, wd=5e-4), cosine annealing, early stopping on weighted F1 (patience 50), weighted cross-entropy (class weights inverse to frequency), manifold Mixup [19] ($\alpha=0.4$), and label smoothing ($\epsilon=0.05$). Metrics: accuracy (Acc), weighted F1 (W-F1), macro-averaged F1 (M-F1), AUC (mean $\pm$ std).

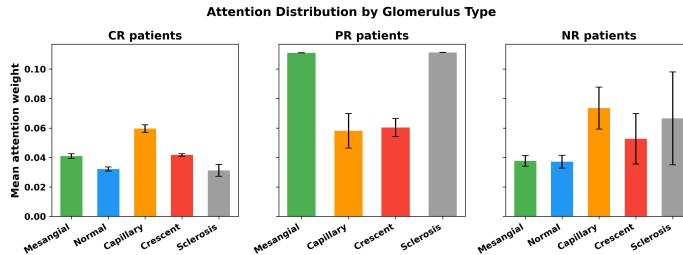
3.2 Main Results

Table 1 presents three-class prognosis prediction results. We evaluate the progression from single-modality prediction to multimodal fusion with increasing clinical context. ABMIL [17], TransMIL [22], and CLAM [23] serve as image-only MIL baselines, Clinical MLP provides the single-modality clinical baseline, and Late Fusion concatenates image and clinical features without cross-modal interaction. Our CIT is evaluated with two clinical input configurations: baseline-only (0m, 25 dimensions) and baseline plus 3-month follow-up (0m+3m, 59 dimensions including treatment response features).

Key observations: (1) **Single modalities are insufficient:** image-only MIL baselines (ABMIL 68.6%, TransMIL 67.6%, CLAM 65.8% Acc) plateau with low M-F1 ($\leq 51.3\%$), indicating poor minority class discrimination regardless of architecture. Clinical MLP (80.7% Acc) improves with temporal features but remains insufficient for reliable three-class discrimination. (2) **Cross-modal fusion provides substantial improvement:** Late Fusion reaches 83.1% Acc, while CIT achieves 90.1% Acc (+7.0%) through conditional attention. (3) **Temporal follow-up enhances prediction:** incorporating 3-month features improves CIT from 86.4% to 90.1% Acc (M-F1: 77.0% $\rightarrow$ 83.9%, AUC: 88.0% $\rightarrow$ 89.4%). This 3-month assessment is a KDIGO-established milestone [21], providing prognostic predictions 9 months prior to the final outcome and a clinically actionable window for timely therapeutic adjustments. (4) **Attention maps reveal clinically meaningful patterns:** PR and NR patients show elevated MIL attention to mesangial proliferative and sclerotic glomeruli—pathological types associated

Table 1. Three-class prognosis prediction (CR/PR/NR) on $n=71$ patients. Results are 5-fold cross-validation mean $\pm$ std. Best in **bold**.

Method	Acc(%)	W-F1(%)	M-F1(%)	AUC(%)
ABMIL [17] (image-only)	68.6 $\pm$ 1.8	62.1 $\pm$ 1.0	38.9 $\pm$ 2.0	45.2 $\pm$ 3.3
TransMIL [22] (image-only)	67.6 $\pm$ 2.1	63.5 $\pm$ 2.1	43.5 $\pm$ 2.1	60.6 $\pm$ 0.6
CLAM [23] (image-only)	65.8 $\pm$ 0.8	65.4 $\pm$ 1.4	51.3 $\pm$ 2.5	61.0 $\pm$ 1.5
Clinical MLP (0m+3m)	80.7 $\pm$ 4.3	80.4 $\pm$ 3.3	70.2 $\pm$ 4.9	81.9 $\pm$ 3.2
Late Fusion (0m+3m)	83.1 $\pm$ 2.3	82.5 $\pm$ 2.7	74.6 $\pm$ 3.4	85.4 $\pm$ 2.1
CIT (Ours, 0m)	86.4 $\pm$ 2.6	85.0 $\pm$ 2.6	77.0 $\pm$ 4.3	88.0 $\pm$ 1.9
CIT (Ours, 0m+3m)	90.1$\pm$1.1	89.2$\pm$1.0	83.9$\pm$1.3	89.4$\pm$2.9

**Fig. 2.** Mean MIL attention weights by glomerulus morphological type for each outcome class (3-seed aggregate, 0m+3m $\rightarrow$ 12m). PR and NR patients show elevated attention to pathological types (mesangial proliferative, sclerotic), while CR patients exhibit more uniform attention distribution.

with disease activity and chronicity—while CR patients exhibit uniform attention across types (Fig. 2), suggesting the model learns to focus on morphologically abnormal glomeruli for adverse outcome prediction. All CIT improvements over baselines are statistically significant ($p<0.05$, Wilcoxon signed-rank test on 15 paired fold-level observations).

3.3 Ablation Studies

We conduct systematic ablation experiments to validate each component’s contribution. All ablations are performed on the main 0m+3m $\rightarrow$ 12m setting ($n=71$) for consistency.

Fusion architecture (Table 2a). CIT outperforms cross-attention [8] and late fusion by +7.0% Acc (90.1% vs. 83.1%), demonstrating that condition-token injection is more effective than dual-stream architectures on small datasets.

Feature extractor (Table 2b). Domain-adapted MAE pretrained features outperform DINOv2 by +7.1% Acc and ResNet50 by +8.1%. MAE finetuned features (84.1%) show only marginal gains over these baselines, confirming that finetuning narrows representations while frozen self-supervised features preserve richer morphological diversity.

Morphological type injection (Table 2c). Multi-granularity injection leverages complementary morphological cues across scales—patch-level identity con-

Table 2. Ablation studies. (a) Fusion architecture, (b) feature extractor, and (c) morphological type injection, all on 0m+3m→12m ($n=71$). 5-fold cross-validation, mean±std over 3 seeds.

(a) Fusion Architecture (0m+3m→12m, $n=71$)				
Method	Acc(%)	W-F1(%)	M-F1(%)	AUC(%)
Cross-Attention [8]	83.1±3.9	83.6±3.7	76.7±5.8	87.1±5.5
Late Fusion	83.1±2.3	82.5±2.7	74.6±3.4	85.4±2.1
CIT (Ours)	90.1±1.1	89.2±1.0	83.9±1.3	89.4±2.9
(b) Feature Extractor (0m+3m→12m, $n=71$)				
Feature Source	Acc(%)	W-F1(%)	M-F1(%)	AUC(%)
DINOv2 ViT-B/14 [15]	83.0±4.0	82.3±4.1	75.0±5.3	83.1±5.3
ResNet50 (ImageNet)	82.0±1.1	81.5±1.2	72.7±1.6	81.4±2.6
MAE ViT-B/16 (finetuned)	84.1±2.4	84.1±2.2	76.9±3.7	83.0±3.0
MAE ViT-B/16 (pretrained, Ours)	90.1±1.1	89.2±1.0	83.9±1.3	89.4±2.9
(c) Morphological Type Injection (0m+3m→12m, $n=71$)				
Configuration	Acc(%)	W-F1(%)	M-F1(%)	AUC(%)
No type features	87.8±2.4	86.5±2.6	78.4±3.2	85.4±5.2
Multi-granularity injection (Ours)	90.1±1.1	89.2±1.0	83.9±1.3	89.4±2.9

strains local attention, patient-level composition informs global context—achieving 90.1% Acc (+2.3%), 83.9% M-F1 (+5.5%), and 89.4% AUC (+4.0%) over the no-type baseline.

4 Conclusion

We presented a multimodal computational pathology framework for pediatric LN prognosis prediction, integrating automated glomerulus detection, domain-adapted MAE feature learning, and a Clinical-Injection Transformer. CIT enables efficient bidirectional cross-modal interaction through condition token injection, while decoupled representation-knowledge adaptation preserves prognostically relevant features and leverages morphological knowledge via multi-granularity type injection. On 71 patients, our method achieves 90.1% accuracy and 89.4% AUC, demonstrating AI-assisted prognosis potential in rare pediatric kidney diseases.

Limitations include the single-center design and modest cohort size. However, this scale is consistent with the field—even the largest multicenter biopsy-confirmed childhood LN studies report approximately 300 patients [3], reflecting the low incidence of pediatric SLE (0.3–0.9 per 100,000 children-years) and the challenges of acquiring invasive biopsies, digitizing histopathology, and maintaining longitudinal follow-up. We are pursuing multi-center validation in ongoing work.

Acknowledgments. We sincerely thank all the anonymous reviewers for their valuable feedback. This work is supported by the National Key R&D Program of China

under Grant NO. 2025YFB3003501, Guangxi Natural Science Foundation - General Program Project No2024JJA140184, Guangzhou Science and Technology Plan Projects (Grant No.2025A03J4374, No.2025A04J4616 and 2025A03J4463) and the National Natural Science Foundation of China (NSFC): No.62502552

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Pinheiro, S.V.B., Dias, R.F., Fabiano, R.C.G., Araujo, S.D.A.: Pediatric lupus nephritis. *Brazilian Journal of Nephrology* **41**, 252–265 (2018)
2. Almaani, S., Meara, A., Rovin, B.H.: Update on lupus nephritis. *Clinical Journal of the American Society of Nephrology* **12**(5), 825–835 (2017)
3. Chan, E.Y.H., Yap, D.Y.H., Wong, W.T., Wong, W.H.S., Wong, S.W., Lin, K.Y.K., et al.: Long-term outcomes of children and adolescents with biopsy-proven childhood-onset lupus nephritis. *Kidney International Reports* **8**(1), 141–150 (2023)
4. Kamphuis, S., Silverman, E.D.: Prevalence and burden of pediatric-onset systemic lupus erythematosus. *Nature Reviews Rheumatology* **6**(9), 538–546 (2010)
5. Tian, J., Zhang, D., Yao, X., Huang, Y., Lu, Q.: Global epidemiology of systemic lupus erythematosus: a comprehensive systematic analysis and modelling study. *Annals of the Rheumatic Diseases* **82**(3), 351–356 (2023)
6. Huang, S., Chen, Y., Song, Y., Wu, K., Chen, T., Zhang, Y., et al.: Deep learning model to predict lupus nephritis renal flare based on dynamic multivariable time-series data. *BMJ Open* **14**(3), e071821 (2024)
7. Cheng, C., Li, B., Li, J., Wang, Y., Xiao, H., Lian, X., et al.: Multi-stain deep learning prediction model of treatment response in lupus nephritis based on renal histopathology. *Kidney International* **107**(4), 714–727 (2025)
8. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., et al.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4025 (2021)
9. Jaume, G., Vaidya, A., Chen, R.J., Williamson, D.F., Liang, P.P., Mahmood, F.: Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11579–11590 (2024)
10. Zhou, F., Chen, H.: Cross-modal translation and alignment for survival analysis. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21485–21494 (2023)
11. Hemker, K., Simidjievski, N., Jamnik, M.: HEALNet: multimodal fusion for heterogeneous biomedical data. In: *Advances in Neural Information Processing Systems* **37**, 64479–64498 (2024)
12. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009 (2022)
13. Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., Prasanna, P.: Self pre-training with masked autoencoders for medical image classification and segmentation. In: *Proceedings of the IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–6. IEEE (2023)

14. Wu, K., Zheng, Y., Shi, J., Xie, F., Jiang, Z.: Position-aware masked autoencoder for histopathology WSI representation learning. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 714–724. Springer, Cham (2023)
15. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., et al.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
16. Tang, Y., He, Y., Nath, V., Guo, P., Deng, R., Yao, T., et al.: Holohisto: End-to-end gigapixel WSI segmentation with 4k resolution sequential tokenization. arXiv preprint arXiv:2407.03307 (2024)
17. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of the International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)
18. Deng, R., Cui, C., Remedios, L.W., Bao, S., Womick, R.M., Chiron, S., et al.: Cross-scale multi-instance learning for pathological image diagnosis. *Medical Image Analysis* **94**, 103124 (2024)
19. Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., Bengio, Y.: Manifold mixup: Better representations by interpolating hidden states. In: Proceedings of the International Conference on Machine Learning, pp. 6438–6447. PMLR (2019)
20. Van Buuren, S., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* **45**, 1–67 (2011)
21. Rovin, B.H., Ayoub, I.M., Chan, T.M., Liu, Z.H., Mejía-Vilet, J.M., Floege, J.: KDIGO 2024 Clinical Practice Guideline for the management of lupus nephritis. *Kidney International* **105**(1), S1–S69 (2024)
22. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X.: TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In: Advances in Neural Information Processing Systems **34**, 2136–2147 (2021)
23. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering* **5**(6), 555–570 (2021)