npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02160-6

Domain-adaptive semi-supervised learning for efficient rare pathological lesion detection with minimal annotation

Isao Matsui[1,2,3,44] ✉, Ayumi Matsumoto[1,44], Atsuhiro Imai[1], Hiroki Okushima[1], Hirohiko Niioka[4], Masatoshi Abe[1], Natsune Tamai[1], Hajime Nagasu[3,5], Eiichiro Kanda[3,6], Eiichiro Uchino[3,7,8], Tadashi Sofue[3,9], Toshiyuki Imasawa[3,10], Yuichiro Yano[3,11], Hiroshi Kinashi[12], Ken-ichi Miyoshi[13], Tamaki Harada[14], Yasuyuki Nagasawa[15], Keiji Fujimoto[16], Yuka Kurokawa[17], Sawako Kato[18], Ryohei Kaseda[19], Masahiro Koizumi[20], Yasuo Kusunoki[21], Masaki Ohya[22], Yoshimasa Kawazoe[23], Hiroyuki Abe[24], Yuta Matsukuma[25], Takaaki Kosugi[26], Yoshiyasu Ueda[27], Naohiko Fujii[28], Masanobu Takeji[21], Akira Suzuki[29], Katsuyuki Nagatoya[30], Kazumasa Oka[31], Yutaka Ando[32], Masaaki Izumi[33], Toshiyuki Komiya[34], Tatsuo Tsukamoto[35], Imari Mimura[36], Takahiro Kuragano[37], Toshiaki Nakano[25], Kazuhiko Tsuruya[26], Yasuhiko Ito[12], Tetsuo Minamino[9], Osamu Yamaguchi[13], Suguru Yamamoto[19], Hirotaka Komaba[20], Kengo Furuichi[16], Kei Fukami[17], Shin-ichi Araki[38], Takao Masaki[39], Naotake Tsuboi[40], Hitoshi Yokoyama[16], Akira Shimizu[41], Tetsuo Ushiku[24], Shoichi Maruyama[18], Motoko Yanagita[7,42], Masaomi Nangaku[36], Ryohei Yamamoto[43], Kazunori Inoue[1] & Yoshitaka Isaka[1,3]

Artificial intelligence for rare pathological lesion detection faces dual challenges: expert annotation scarcity and domain shifts across institutions. Using multi-institutional kidney biopsies from 22 hospitals with 3 scanner types (NDPI, VSI, SVS), we demonstrate that model performance decreases dramatically across domains, with up to 70.3% reduction in detection precision for rare lesions such as crescents and segmental sclerosis (comprising only 2-3% of annotations). We present an approach integrating semi-supervised learning with residual CycleGAN-based domain adaptation, reducing mean Fréchet inception distance between institutions from 55.9 to 20.2 while preserving diagnostic morphology. We identified context-dependent optimal strategies: semi-supervised learning with 50% confidence threshold excelled in same-hospital scenarios (15.2-17.7% improvement for rare lesions), while our combined GAN-Semi-Supervised approach demonstrated superior performance in cross-scanner scenarios between NDPI and VSI formats (up to 63.4% improvement for crescents). This methodology enables robust performance across diverse healthcare settings with minimal expert annotation.

Histopathological examination remains the gold standard for diagnosing numerous diseases. However, this critical process faces significant challenges, including interobserver variability and time-intensive analysis[1,2]. Although artificial intelligence (AI) has transformed medical imaging, pathology lags behind radiology in terms of clinical AI implementation, with Food and Drug Administration (FDA)-approved pathology systems predominantly limited to specific applications[3]. This disparity stems from two critical challenges: the extensive annotation burden in medical imaging, which is particularly pronounced in pathology requiring specialized expertise, and the substantial domain shifts unique to pathology caused by inter-institutional variations in staining protocols and scanning equipment[4].

Semi-supervised learning offers a promising solution to annotation constraints by leveraging both labeled and unlabeled data through consistency regularization and pseudo-labeling[5–9]. However, these methods have primarily been validated on datasets with balanced class distributions. Rare pathological findings typically comprise only 2–3% of data, yet they often indicate severe diseases that require immediate intervention. This extreme class imbalance significantly inhibits reliable pseudo-labeling[5,6,10]. Concurrently, domain shift across institutions significantly affects model

A full list of affiliations appears at the end of the paper. ✉e-mail: matsui@kid.med.osaka-u.ac.jp

performance in multicenter pathology studies[4,11,12]. Cycle-consistent generative adversarial networks (CycleGANs) have emerged as a promising approach by which to address these variations through unsupervised image-to-image translation[13–15]. However, recent studies have demonstrated that conventional CycleGANs may inadvertently alter critical morphological features during transformation, particularly in fine-grained structures relevant to diagnosis, and may thereby compromise diagnostic accuracy[16–18].

Recent advances in foundation models and self-supervised learning have shown potential in medical imaging[19–21]. Our previous work on self-supervised learning for cropped glomerular image classification demonstrated improved performance with minimal annotations[22]. While these approaches offer powerful representations, they encounter challenges in pathological lesion detection: precise spatial localization requirements favor specialized object detection architectures over global representation models[23], computational demands of large foundation models limit clinical practicality[24], and extreme class imbalance requires specific handling mechanisms that are not inherently addressed by general-purpose models.

In this study, we investigated the effectiveness of combining semi-supervised learning with modified CycleGAN-based data augmentation for the detection of rare pathological lesions, using glomerular lesions in kidney biopsy images as test cases. We selected You Only Look Once version 8 (YOLOv8) as our detection framework owing to three key advantages: its lightweight architecture enables deployment on standard clinical hardware, its object detection-specific design is optimized for precise localization, and its distributed focal loss directly addresses class imbalance challenges. These characteristics make YOLOv8 particularly suitable for rare lesion detection while allowing the seamless integration of our semi-supervised and domain adaptation techniques.

Our investigation aimed to (1) evaluate semi-supervised learning with YOLOv8 for rare lesion detection, (2) explore the benefits of combining this approach with residual CycleGAN-based domain adaptation, and (3) identify optimal training strategies that minimize annotation burden while maintaining robust performance across diverse clinical settings. Though demonstrated on renal pathology, our approach addresses universal challenges in pathological image analysis and offers broadly applicable methodological innovations that complement recent advances in medical imaging AI.

Results

Study population and dataset characteristics

Both the development and test cohorts predominantly included patients with chronic kidney disease (CKD) stages G2-3b A3, encompassing a broad spectrum of renal pathologies (Supplementary Tables 1 and 2).

The image dataset, acquired from 22 distinct medical institutions, comprised 3 file formats (NDPI, VSI, and SVS) corresponding to different slide scanners, with whole slide image (WSI) counts per institution ranging from 50 to 249 in the development cohort and from 42 to 221 in the test cohort (Table 1). Annotations in the PAS-stained images classify the glomerular structures into four categories: glomeruli (excluding global sclerosis, crescents, and segmental sclerosis), global sclerosis, crescents, and segmental sclerosis (Supplementary Fig. 1). Global sclerosis represented approximately 16% of all annotations, whereas crescent and segmental sclerosis were relatively rare, constituting only 2–3% of the total annotations (Table 1).

A feature analysis of the cropped glomerular images revealed that scanner type was the primary factor contributing to image variability, with distinct clustering patterns observed across hospitals (Fig. 1a and Supplementary Fig. 2). In the feature space, globally sclerotic glomeruli demonstrated relatively cohesive distribution patterns, whereas crescent and segmental scleroses exhibited more dispersed distributions with partial clustering (Fig. 1a). The quantification of domain shifts using the Fréchet inception distance (FID) confirmed these visual observations, with a high mean FID value of 55.9 across all hospital pairs (Fig. 1b). A single hospital using the SVS format exhibited substantial domain shifts, particularly when compared with VSI-format hospitals (Fig. 1b). These findings highlight that the detection of crescents and segmental sclerosis presents particularly challenging tasks across different scanner types.

Development of baseline supervised models and their performance

For model training and evaluation, the WSIs were divided into non-overlapping patches of three sizes (Supplementary Fig. 3 and Supplementary Table 3). We strategically selected three representative hospitals to develop baseline supervised models and efficiently optimize the hyperparameters. Hospitals 02 (NDPI format), 06 (VSI format), and 22 (SVS format, sole provider) were selected as representatives of each scanner file format. While Hospitals 01 and 02 had similar numbers of WSIs, Hospital 02 was selected for its higher number of annotated lesions, and Hospital 06 for its substantial data volume and because it is the lead authors' affiliated institution.

Initial supervised learning was performed independently on data from each selected hospital. The hospital data was divided at the WSI level into development and test sets in approximately equal proportions (Table 1 and Supplementary Fig. 4). The development set was further split at the WSI level by performing five iterations of random training-validation splits with a 2:1 ratio. Higher proportions were allocated to the validation and test sets to ensure adequate representation of rare glomerular lesions. We assessed the model robustness across three distinct test categories: Category 1 (Cat. 1), comprising images from the same hospital but not used during model development; Category 2 (Cat. 2) comprising images from different hospitals using identical scanner types, and Category 3 (Cat. 3), comprising images obtained using different scanner types (Supplementary Fig. 4a). Category 3 was further subdivided as follows. For the NDPI/VSI-based models, Cat. 3-1 represents a cross-evaluation between these formats, and Cat. 3-2 describes the evaluation of the SVS data. For the SVS-based models, Cat. 3-1 and Cat. 3-2 represent the evaluations of the NDPI and VSI data, respectively.

The baseline models demonstrated lower detection performance for rare lesions compared to common ones, with AP_{50} values for crescents and segmental sclerosis class consistently lower than those for glomerulus class across all test categories (Table 2). Additionally, these models exhibited significant performance degradation (indicated by bold values in Table 2) when tested on images from different hospitals, even those using identical scanner types. This cross-institutional performance decline was particularly pronounced for rare lesions when evaluating across different scanner types, with crescent detection AP_{50} showing a dramatic 70.3% reduction (from 0.64 to 0.19, calculated as 1–(0.19/0.64)) when the VSI-derived Hospital 06 model was evaluated on NDPI format images. Similar substantial degradation was observed for segmental sclerosis detection, highlighting the pronounced impact of domain shift on rare lesion identification.

Optimization of baseline supervised models before semi-supervised learning

To optimize the model performance before implementing semi-supervised learning, we first evaluated the detection performance for different patch sizes. Using large patch predictions as a reference, medium patches showed comparable performance, whereas small patches demonstrated significantly decreased performance (Supplementary Table 4). We then tested whether restricting model training to only large and medium patches (LM models) would improve efficiency, but we found no significant improvement over models trained with all patch sizes and occasionally observed decreased performance (Supplementary Table 5). Consequently, we retained all patch sizes for model training while restricting the evaluation to large and medium patches.

We also evaluated three augmentation strategies that target patches containing rare lesions (crescent or segmental sclerosis). Contrary to expectations, none of these strategies improved the detection performance compared with non-augmented models, and some even decreased performance (Supplementary Table 6). Additionally, benchmark comparison with a pathology foundation model revealed that our standard YOLOv8

Table 1 | Distribution of file formats, WSI counts, and annotations across hospitals

Hospital	WSI format	WSI count	Development cohort				WSI count	Test cohort			
			Annotation count (fraction of annotated data)					Annotation count (fraction of annotated data)			
			Glomerulus	Global sclerosis	Crescent	Segmental sclerosis		Glomerulus	Global sclerosis	Crescent	Segmental Sclerosis
Hosp01	NDPI	249	5378 (0.778)	1191 (0.172)	169(0.024)	179 (0.026)	221	4312 (0.760)	1176 (0.207)	78 (0.014)	106 (0.019)
Hosp02	NDPI	244	6011(0.766)	1344 (0.171)	229 (0.029)	262 (0.033)	205	5478 (0.801)	916 (0.134)	252 (0.037)	189 (0.028)
Hosp03	NDPI	195	6940 (0.774)	1522 (0.170)	223 (0.025)	279 (0.031)	168	6455 (0.795)	1248 (0.154)	191 (0.024)	222 (0.027)
Hosp04	NDPI	155	4278 (0.747)	1143 (0.200)	145 (0.025)	162 (0.028)	134	4330 (0.792)	948 (0.174)	52 (0.010)	134 (0.025)
Hosp05	VSI	148	3833 (0.778)	727 (0.148)	169 (0.034)	199 (0.040)	127	3262 (0.808)	595 (0.147)	84 (0.021)	98 (0.024)
Hosp06	VSI	145	3600 (0.807)	619 (0.139)	140 (0.031)	100 (0.022)	125	2952 (0.818)	554 (0.154)	77 (0.021)	26 (0.007)
Hosp07	VSI	125	2948 (0.791)	636 (0.171)	61 (0.016)	83 (0.022)	109	2519 (0.760)	687 (0.207)	60 (0.018)	49 (0.015)
Hosp08	VSI	122	2848 (0.806)	525 (0.149)	111 (0.031)	51 (0.014)	102	2098 (0.742)	616 (0.218)	82 (0.029)	30 (0.011)
Hosp09	VSI	117	1415 (0.696)	421 (0.207)	103 (0.051)	93 (0.046)	110	1538 (0.755)	382 (0.188)	55 (0.027)	62 (0.030)
Hosp10	VSI	107	4275 (0.870)	497 (0.101)	103 (0.021)	38 (0.008)	93	3344 (0.849)	478 (0.121)	84 (0.021)	35 (0.009)
Hosp11	VSI	96	1389 (0.683)	437 (0.215)	83 (0.041)	125 (0.061)	84	1328 (0.726)	361 (0.197)	77 (0.042)	63 (0.034)
Hosp12	VSI	92	2325 (0.853)	321 (0.118)	36 (0.013)	43 (0.016)	80	2184 (0.846)	338 (0.131)	47 (0.018)	13 (0.005)
Hosp13	VSI	86	1482 (0.713)	448 (0.216)	103 (0.050)	45 (0.022)	75	1396 (0.751)	372 (0.200)	60 (0.032)	31 (0.017)
Hosp14	VSI	85	1815 (0.798)	337 (0.148)	91 (0.040)	31 (0.014)	74	1631 (0.768)	412 (0.194)	59 (0.028)	22 (0.010)
Hosp15	VSI	73	1444 (0.770)	331 (0.177)	68 (0.036)	32 (0.017)	66	1461 (0.745)	331 (0.169)	140 (0.071)	28 (0.014)
Hosp16	VSI	72	957 (0.786)	184 (0.151)	26 (0.021)	50 (0.041)	65	996 (0.761)	238 (0.182)	39 (0.030)	36 (0.028)
Hosp17	VSI	72	1214 (0.780)	237 (0.152)	51 (0.033)	54 (0.035)	63	1037 (0.754)	228 (0.166)	66 (0.048)	45 (0.033)
Hosp18	VSI	69	1390 (0.859)	183 (0.113)	14 (0.009)	31 (0.019)	60	1155 (0.808)	198 (0.138)	36 (0.025)	41 (0.029)
Hosp19	VSI	66	1479 (0.833)	208 (0.117)	59 (0.033)	30 (0.017)	61	1361 (0.780)	274 (0.157)	77 (0.044)	32 (0.018)
Hosp20	VSI	66	1069 (0.798)	205 (0.153)	41 (0.031)	24 (0.018)	61	966 (0.820)	180 (0.153)	21 (0.018)	11 (0.009)
Hosp21	VSI	50	664 (0.778)	162 (0.190)	13 (0.015)	14 (0.016)	42	546 (0.763)	131 (0.183)	12 (0.017)	27 (0.038)
Hosp22	SVS	64	2093 (0.725)	670 (0.232)	62 (0.021)	60 (0.021)	53	1492 (0.732)	458 (0.225)	52 (0.026)	37 (0.018)
Sum		2498	58847 (0.782)	12348 (0.164)	2100 (0.028)	1985 (0.026)	2178	51841 (0.785)	11121 (0.169)	1701 (0.026)	1337 (0.020)

This table presents an overview of the dataset characteristics across participating hospitals. Each kidney biopsy procedure results in the creation of a single WSI, with the number of WSIs corresponding directly to the number of biopsies performed. Annotations are categorized into four types: glomerulus (excluding global sclerosis, crescents, and segmental sclerosis), global sclerosis, crescents, and segmental sclerosis. For each annotation type, both the absolute count and the percentage relative to the total annotations for that hospital within its respective cohort (development or test) are provided. *WSI* whole slide image, *Hosp* hospital, *NDPI* Hamamatsu Photonics file format, *VSI* Evident file format, *SVS* Leica file format.

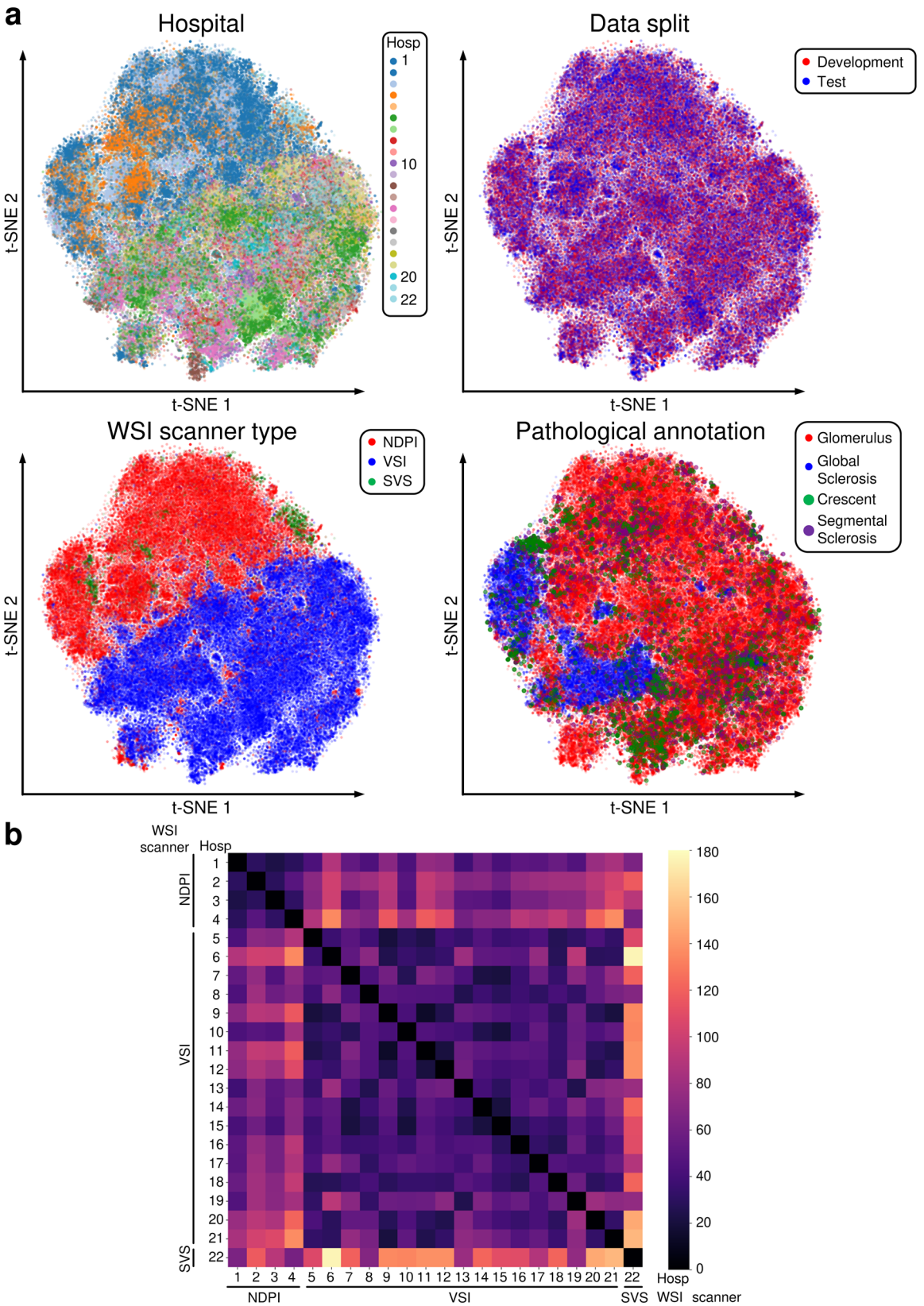


Fig. 1 | Visualization of glomerular features from PAS-stained kidney biopsy images. a *t*-distributed stochastic neighbor embedding (t-SNE) plots showing the distribution of features extracted from glomerular images using ImageNet-pretrained ResNet50. Top row: distribution by hospital source (left) and development/test split (right). Bottom row: distribution by WSI scanner type (left) and pathological annotation (right). Points representing crescent and segmental sclerosis lesions are enlarged for better visibility owing to their rarity. **b** Heatmap visualization of the Fréchet inception distance (FID) among hospitals, with a high mean FID value of 55.9 across all hospital pairs. WSI whole slide image, NDPI Hamamatsu Photonics file format, VSI Evident file format, SVS Leica file format, Hosp hospital.

Table 2 | Performance of baseline supervised models and impact of domain shift

Train data	Test data	mAP_{50}	AP_{50}			
			Glomerulus	Global sclerosis	Crescent	Segmental sclerosis
Hosp02	(Cat.1) Hosp02	0.73 [0.72–0.74] Ref.	0.94 [0.94–0.94] Ref.	0.83 [0.83–0.83] Ref.	0.69 [0.68–0.71] Ref.	0.46 [0.44–0.48] Ref.
	(Cat. 2)NDPI w/o Hosp02	**0.65 [0.64–0.66] $P < 0.01$**	**0.93 [0.93–0.93] $P < 0.01$**	**0.79 [0.79–0.79] $P < 0.01$**	**0.45 [0.43–0.48] $P < 0.01$**	**0.42 [0.41–0.44] $P < 0.01$**
	(Cat. 3-1) VSI	**0.64 [0.63–0.64] $P < 0.01$**	**0.91 [0.90–0.91] $P < 0.01$**	**0.75 [0.75–0.76] $P < 0.01$**	**0.56 [0.53–0.59] $P < 0.01$**	**0.33 [0.31–0.35]$P < 0.01$**
	(Cat. 3-2) SVS	**0.71 [0.71–0.72] $P < 0.01$**	**0.92 [0.92–0.92] $P < 0.01$**	0.83 [0.82–0.83] $P = 0.55$	**0.64 [0.61–0.67] $P < 0.01$**	0.47 [0.43–0.50] $P = 0.83$
Hosp06	(Cat. 1) Hosp06	0.69 [0.66–0.72] Ref.	0.94 [0.93–0.94] Ref.	0.78 [0.77–0.80] Ref.	0.64 [0.59–0.69] Ref.	0.39 [0.30–0.49] Ref.
	(Cat. 2) VSI w/oHosp06	**0.58 [0.55–0.61]$P < 0.01$**	**0.91 [0.90–0.91] $P < 0.01$**	**0.72 [0.70–0.75]$P = 0.03$**	**0.47 [0.43–0.52] $P < 0.01$**	**0.21 [0.15–0.27] $P < 0.01$**
	(Cat. 3-1) NDPI	**0.48 [0.42–0.54] $P < 0.01$**	**0.88 [0.87–0.89] $P < 0.01$**	**0.63 [0.59–0.68] $P < 0.01$**	**0.19 [0.08–0.30] $P < 0.01$**	**0.23 [0.15–0.31] $P < 0.01$**
	(Cat. 3-2) SVS	**0.49 [0.43–0.55] $P < 0.01$**	**0.86 [0.85–0.88] $P < 0.01$**	**0.67 [0.61–0.74] $P < 0.01$**	**0.24 [0.12–0.36] $P < 0.01$**	**0.18 [0.10–0.26] $P < 0.01$**
Hosp22	(Cat. 1) Hosp22	0.68 [0.66–0.69] Ref.	0.91 [0.91–0.92] Ref.	0.81 [0.80–0.82] Ref.	0.64 [0.59–0.68] Ref.	0.34 [0.32–0.36] Ref.
	(Cat. 3-1) NDPI	**0.58 [0.56–0.60] $P < 0.01$**	0.91 [0.91–0.91] $P = 0.21$	**0.74 [0.72–0.76] $P < 0.01$**	**0.40 [0.36–0.44] $P < 0.01$**	**0.26 [0.23–0.29] $P < 0.01$**
	(Cat. 3-2) VSI	**0.50 [0.48–0.53] $P < 0.01$**	**0.88 [0.87–0.88] $P < 0.01$**	**0.70 [0.68–0.72] $P < 0.01$**	**0.32 [0.25–0.39] $P < 0.01$**	**0.12 [0.10–0.14] $P < 0.01$**

This table presents the performance of baseline supervised learning models in detecting glomerular lesions on PAS-stained renal biopsy images. Models were independently developed using data from Hospitals 02, 06, and 22. Each hospital's WSIs were divided into large (L), medium (M), and small (S) patches for training, validation, and testing (Supplementary Fig. 3). For each hospital, five models were developed through random splits at the WSI level (2/3 training, 1/3 validation) (Supplementary Fig. 4a). Performance metrics include mean average precision at 50% Intersection over Union (mAP_{50}) and AP_{50} for each class, presented as means with 95% confidence intervals.

Test data were categorized as: Category 1 (Cat. 1)—images from the same hospital not used for model development; Category 2 (Cat. 2)—images from other hospitals with identical scanner types; and Category 3—images from hospitals with different scanner types. Category 3 was further subdivided: for NDPI/VSI-based models, Cat. 3-1 represented cross-evaluation between these formats and Cat. 3-2 denotes evaluation on SVS data; for SVS-based models, Cat. 3-1 and Cat. 3-2 represented evaluation on NDPI and VSI data, respectively. Statistical comparisons were performed using Dunnett's test comparing Category 2 and 3 performance against Category 1. Bold indicates significant differences ($P < 0.05$). *WSI* whole slide image, *w/o* without, *Ref* reference, *Hosp* hospital, *NDPI* Hamamatsu Photonics file format, *VSI* Evident file format, AP_{50} average precision at 50% Intersection over Union, mAP_{50} mean AP_{50}. *Ref* reference category used as baseline for statistical comparison.

approach consistently outperformed UNI-based models across evaluation scenarios (Supplementary Table 7)[25]. Based on these comprehensive findings, models trained with all patch sizes and without specific augmentation (Baseline YOLO) were selected as the foundation for subsequent semi-supervised learning experiments.

Semi-supervised learning parameter optimization and performance evaluation

Our semi-supervised approach involved generating pseudo-labels for images from other hospitals in the development cohort using Baseline YOLO and then combining high-confidence pseudo-labeled patches with the original labeled data for model training (Fig. 2). To determine optimal parameters, we evaluated various confidence thresholds (50%, 70%, and 90%) and iteration strategies (Fig. 3 and Supplementary Fig. 5). A pooled model using a completely labeled dataset served as an upper performance benchmark. The models using a 90% threshold (Semi-90-1) showed minimal improvement over the baseline (Fig. 3). For the 70% and 50% thresholds, we tested both single-iteration models (Semi-70-1, Semi-50-1) and second-iteration models (Semi-70-2, Semi-50-2) trained using refined pseudo-labels from their respective first-iteration models. Notably, the second-iteration models frequently underperformed compared with the first-iteration models, indicating that a single iteration was optimal (Fig. 3).

Between Semi-70-1 and Semi-50-1, the 50% threshold model demonstrated superior performance, improving nine AP_{50} metrics in Category 1, versus seven metrics with a 70% threshold (both improved three metrics for rare lesions). In Category 2, the Semi-50-1 improved three AP_{50} metrics, all for rare lesions, whereas the Semi-70-1 improved three metrics, but only two for rare lesions. Category 3 evaluations showed comparable results for these thresholds. Based on these trends, we selected the Semi-50-1 (Semi-Supervised YOLO) as the optimal configuration.

Semi-supervised learning using external datasets from the Kidney Precision Medicine Project (KPMP) demonstrated performance improvements in several Category 1 and 2 evaluations (Supplementary Fig. 6). The category 3 evaluations showed mixed results, with some parameters improving and others deteriorating, similar to the results of the primary analysis.

Modified CycleGAN-based augmentation strategy to mitigate domain shift

Despite improvements through semi-supervised learning, domain shift remains a significant challenge. While Semi-Supervised YOLO models achieved multiple AP_{50} improvements in the same-hospital evaluations, the benefits were minimal when assessed using data from other institutions (Fig. 3). To address these domain shifts, we implemented a residual CycleGAN (Supplementary Fig. 7)[18]. In addition to the conventional CycleGAN loss functions, residual loss was introduced to prevent excessive transformation, thus enabling the transformation of pathological images while preserving their morphological integrity, which is essential for retaining diagnostically important fine structures.

We developed hospital-specific residual CycleGAN models to generate style-transformed images and subsequently validated their morphological fidelity (Fig. 4, Tables 3, 4, and Supplementary Figs. 8–11). Feature analysis confirmed a significant reduction in domain gaps among institutions, with the mean FID value decreasing from 55.9 to 20.2 across all hospital pairs (Supplementary Fig. 10). Visual assessment by expert nephrologists demonstrated high preservation of diagnostic morphology across residual CycleGAN transformations (Tables 3, 4). Overall, 86.3 ± 3.3% of transformed images received Score 0 (no artifacts), with an additional 12.7 ± 2.8% classified as clinically acceptable (Score 1). Only 0.9 ± 0.4% images showed diagnostic impairment (Score 2). Class-specific analysis revealed that normal glomeruli demonstrated the highest fidelity, while global sclerosis and segmental sclerosis showed slightly more transformation artifacts but remained predominantly clinically acceptable (Table 3). Transformation artifacts in global sclerosis and segmental sclerosis predominantly occurred in areas with sclerotic tissue, where relatively uniform

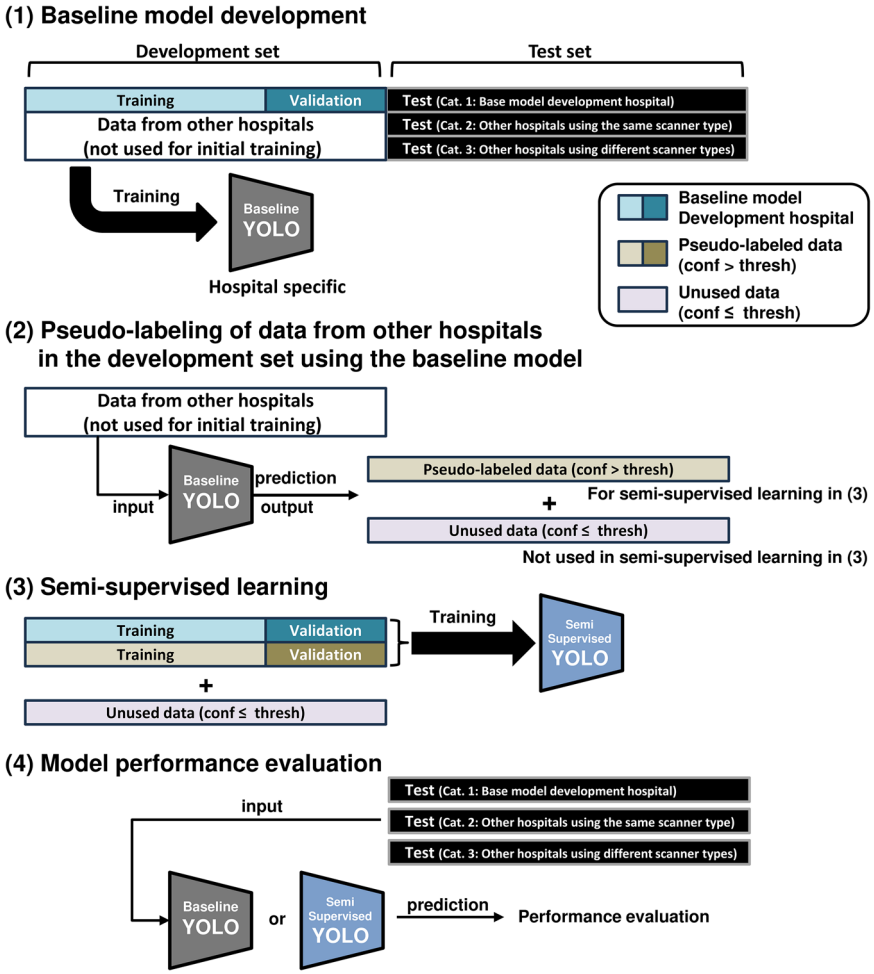


Fig. 2 | Semi-supervised learning framework for glomerular lesion detection. Schematic overview of our four-step semi-supervised learning approach: (1) Baseline supervised model (Baseline YOLO) development using data from a single hospital; (2) generation of pseudo-labels for data from other hospitals using the Baseline YOLO, selecting images with prediction confidence above a predetermined threshold; (3) semi-supervised learning by combining original labeled data with high-confidence pseudo-labeled data to create the Semi-Supervised YOLO model; (4) performance evaluation comparing Baseline YOLO and Semi-Supervised YOLO across three test categories: Category 1 (same hospital), Category 2 (other hospitals using same scanner type), and Category 3 (other hospitals using different scanner types).

intensity regions appeared more susceptible to residual CycleGAN-induced texture variations (Supplementary Fig. 11). Crescent lesions maintained high morphological preservation (Table 3). Scanner-specific transformation analyses showed variable performance, with within-scanner transformations achieving better fidelity scores compared to cross-scanner transformations (Table 4). However, even in cross-scanner transformations, only 0.3 to 2.1% images showed diagnostic impairment.

Our implementation then followed three key steps. First, we developed hospital-specific style transformations, training 21 separate residual CycleGANs for each baseline hospital (Fig. 4a). Second, we trained the GAN-Augmented YOLO models using both the original images from the baseline hospital and their residual CycleGAN-transformed versions. Third, we combined this approach with semi-supervised learning using GAN-Augmented models to generate pseudo-labels for images from other hospitals and selected patches with confidence scores above 50% to train the GAN-Semi-Supervised YOLO models (Fig. 4c). We systematically compared four model configurations: Baseline YOLO, Semi-Supervised YOLO, GAN-Augmented YOLO, and GAN-Semi-Supervised YOLO (Fig. 4d).

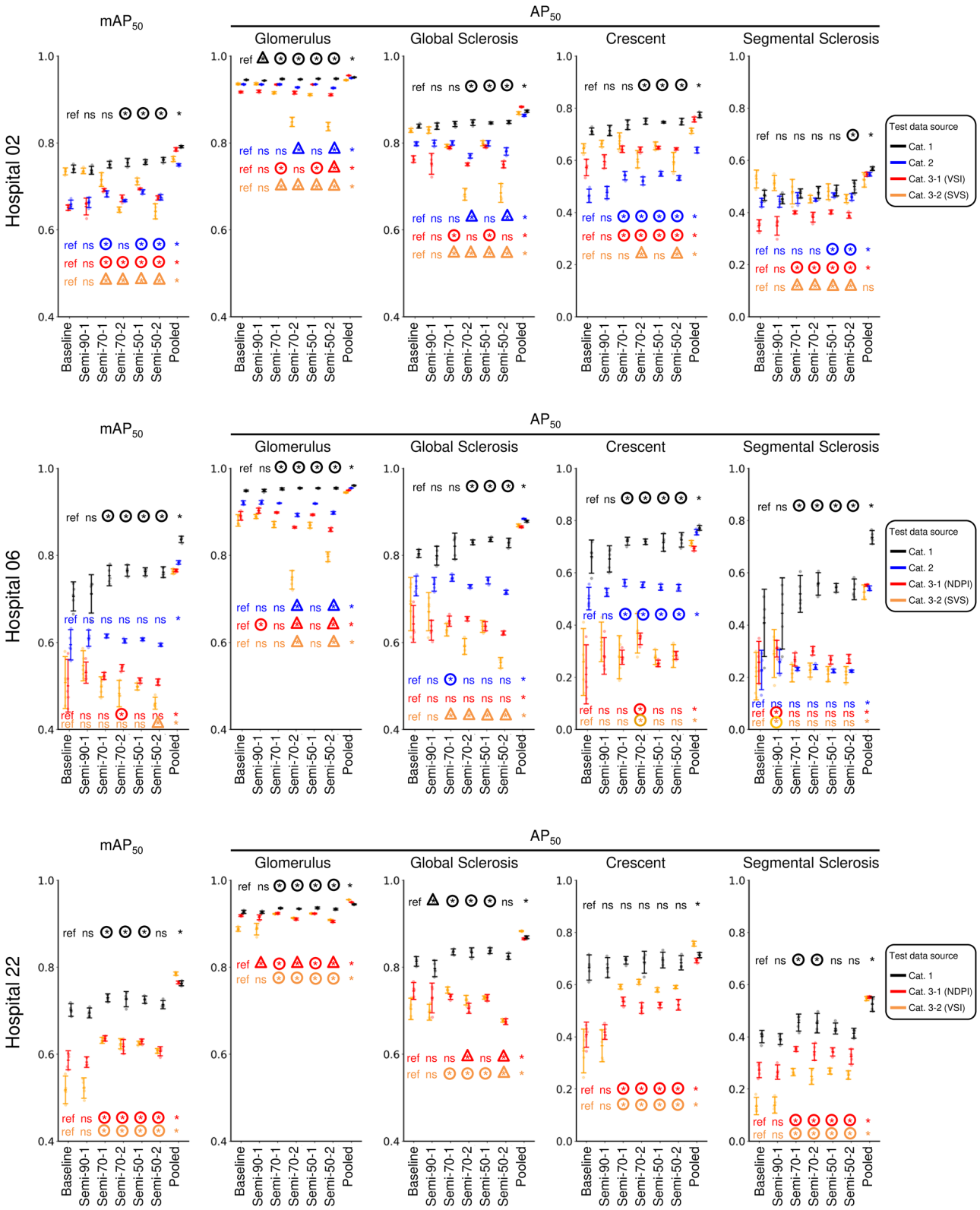
Performance comparison across different model training strategies and testing scenarios

We evaluated our four modeling approaches using radar charts to visualize the AP_{50} values for each glomerular class across the different testing categories. For the three hospitals, the performance patterns varied according to lesion type and testing scenario (Fig. 5). For the glomerulus class, all models demonstrated consistently high detection performances, with the AP_{50} values typically exceeding 0.9. Global sclerosis detection showed similar patterns, with slightly more pronounced variations among the models. For rare lesions, the differences between the training strategies became more apparent. In crescent detection, Category 1 testing showed that Semi-Supervised YOLO (blue lines in Fig. 5) and GAN-Semi-Supervised YOLO (red) improved performance, whereas GAN-Augmentation alone (yellow) often decreased performance. However, in Category 3 testing, the Semi-Supervised YOLO (blue) showed limited improvement, whereas the combined GAN-Semi-Supervised approach (red) demonstrated superior performance. For segmental sclerosis detection, which is the most challenging task, GAN-Semi-Supervised YOLO (red) showed advantages in cross-scanner evaluation.

To validate the generalizability, we extended our analysis to eight additional hospitals: Hospitals 04, 08, 10, 12, 14, 16, 18, and 20 (Supplementary Fig. 12). The results showed a lower performance in hospitals with smaller training datasets (hospitals with higher ID numbers typically had fewer samples, as shown in Table 1). For crescent detection, Semi-Supervised YOLO (blue) improved the performance in Category 1 testing, whereas GAN-Semi-Supervised YOLO (red) excelled in Category 3 evaluations. For segmental sclerosis, Semi-Supervised YOLO (blue) generally outperformed other models in Category 1, whereas GAN-Semi-Supervised YOLO (red) demonstrated advantages in Category 3-1 testing. However, in Category 3-2 testing, no consistent improvement was observed with any approach, highlighting the difficulty in adapting to the SVS scanner format.

Investigation of SVS format limitations through single-hospital simulation

To investigate whether the limited performance improvements on SVS data resulted from single-hospital origin effects, we conducted simulation experiments using Hospital 02 (NDPI) and Hospital 06 (VSI) as source



hospitals. We compared GAN-Semi-Supervised models with limited cross-format diversity against our main models, incorporating all available hospitals. The simulation revealed variable effects of single-hospital limitations (Supplementary Table 8). Hospital 02-based models showed decreased performance when cross-format training was limited to a single VSI hospital, with a notable reduction in segmental sclerosis detection. Conversely, Hospital 06-based models maintained overall performance when cross-format training was limited to a single NDPI hospital, though with decreased global sclerosis performance and improved crescent detection. These findings indicate that single-hospital limitations can affect cross-scanner performance, but the impact varies depending on scanner combination and source hospital characteristics. The limited performance improvements observed for SVS data likely result from multiple contributing factors beyond single-hospital origin alone.

Fig. 3 | Performance comparison of semi-supervised learning models with different confidence thresholds and iterations. The performances of the baseline and semi-supervised models for the three hospitals (Hospitals 02, 06, and 22) were evaluated using mAP_{50} and class-specific AP_{50}. The models included a baseline model (trained with single-hospital data); semi-supervised models with confidence thresholds of 90% (Semi-90-1), 70% (Semi-70-1 for the first iteration and Semi-70-2 for the second iteration), and 50% (Semi-50-1 for the first iteration, Semi-50-2 for the second iteration); and a pooled model representing the maximum achievable performance using all labeled data. Statistical significance was assessed using Dunnett's test with the baseline model as a reference. Asterisks indicate statistically significant differences ($P < 0.05$), with circles and triangles indicating improvement and deterioration, respectively. For the pooled models, significance testing was performed without using improvement/deterioration indicators, as these represent different training paradigms. Colors represent the different test categories: Category 1 (black), Category 2 (blue), and Category 3 (red/orange). Individual data points are represented as lighter-colored dots, along with their mean values and 95% confidence intervals. AP_{50} average precision at 50% Intersection over Union, mAP_{50}, mean AP_{50}.

Cross-institutional generalization and optimal strategy selection

To identify the optimal approach for each scenario, we performed systematic comparisons, using the mAP_{50} as the metric (Fig. 6a). The color-coded matrix revealed that Semi-Supervised YOLO (blue panel in Fig. 6a) most frequently excelled in Category 1. For Category 2, Semi-Supervised YOLO (blue) generally performed the best, although GAN-based approaches (yellow and orange) showed advantages for specific hospitals. Notably, in Category 3-1 testing, the GAN-Augmented (yellow) and GAN-Semi-Supervised YOLOs (orange) dominated, with the latter showing the highest prevalence of statistically significant improvements and the greatest magnitude of improvement. In Category 3-2 evaluations, several GAN-based approaches (yellow and orange) demonstrated improvements over baseline models.

An analysis of the improvement rates by glomerular class, aggregated across hospitals within the same test category, revealed that the greatest performance gains were achieved for rare lesions, particularly crescents (Fig. 6b). For both crescent and segmental sclerosis detection, the Semi-Supervised YOLO showed statistically significant improvements in same-hospital evaluations (Category 1). Semi-Supervised YOLO also demonstrated significant performance improvement for global sclerosis detection in the same-hospital evaluations (Category 1). However, for pathological analysis models, achieving robust performance not only on the same-institution data but also across different institutions is critically important. From this perspective, while Semi-Supervised YOLO showed significant improvement for crescent detection in the challenging cross-scanner scenarios between NDPI and VSI formats (Category 3-1), it failed to improve performance for global sclerosis and segmental sclerosis in these demanding conditions.

In contrast, GAN-Semi-Supervised YOLO demonstrated superior cross-institutional generalization capabilities, particularly for global sclerosis, crescent, and segmental sclerosis detection in Category 3-1 evaluations. For crescent detection, GAN-Semi-Supervised YOLO achieved performance improvements even in the most difficult Category 3-2 evaluations. Importantly, although residual-CycleGAN effectively addressed domain shifts between different hospitals and between NDPI and VSI scanner types, residual-CycleGAN-based adaptation alone tended to reduce performance on the source-hospital test data. This tradeoff highlights the value of our combined GAN-Semi-Supervised approach, which maintains strong performance across all testing scenarios.

Model interpretability analysis reveals enhanced feature learning

To understand the underlying mechanisms of performance improvement, we conducted explainability analysis using Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize model attention patterns. Comparative analysis between Baseline YOLO and GAN-Semi-Supervised YOLO revealed distinct differences in attention mechanisms across various pathological scenarios. For false positive reduction, our analysis demonstrated that GAN-Semi-Supervised YOLO exhibited more refined attention patterns. In cases where Baseline YOLO incorrectly classified atrophic tubules as global sclerosis, the improved model showed reduced attention to tubular structures while maintaining focus on actual glomerular regions (Supplementary Fig. 13a). Similarly, for segmental sclerosis misclassified as normal glomeruli by Baseline YOLO, GAN-Semi-Supervised YOLO demonstrated enhanced attention to sclerotic segments (Supplementary Fig. 13b). For false negative reduction, GAN-Semi-Supervised YOLO showed improved sensitivity in detecting subtle pathological features. In cases where Baseline YOLO failed to detect atypical crescents with minimal capillary loops, GAN-Semi-Supervised YOLO successfully focused attention on crescent formations, enabling accurate detection of these challenging lesions (Supplementary Fig. 14). These findings suggest that the combined domain adaptation and semi-supervised learning approach enables more robust feature learning, leading to improved discrimination of pathological structures.

Discussion

This study demonstrated that combining semi-supervised learning with residual CycleGAN-based domain adaptation enhances rare pathological lesion detection across diverse clinical settings while minimizing the annotation burden. Our approach addresses two fundamental challenges in AI-assisted pathological image analysis: limited expert annotations and domain shifts inherent in multicenter studies.

Our semi-supervised learning approach effectively leveraged unlabeled data, particularly for rare lesions. A 50% confidence threshold for pseudo-labeling yielded optimal results, suggesting that for rare entities, including more diverse samples with moderate confidence provides a greater benefit than does restricting to only the highest-confidence predictions. Notably, a single iteration of pseudo-labeling proved optimal, with second-iteration models frequently underperforming, likely owing to the amplification of initial prediction errors leading to confirmation bias[10,26–28].

Although semi-supervised learning improved same-hospital performance, its benefits diminished across different scanner types. Previous studies on semi-supervised learning have not quantitatively evaluated such domain shifts, making this finding significant[6–9]. Our systematic comparison revealed context-dependent optimal strategies. For same-hospital testing, the Semi-Supervised YOLO generally excelled, suggesting that when domain shifts were minimal, pseudo-labeling alone was sufficient. Conversely, models that use only GAN augmentation often exhibit decreased performance in same-hospital testing. This observation highlights the value of our combined GAN-Semi-Supervised approach, which balances domain adaptation with semi-supervised learning benefits and maintains robust performance across diverse testing scenarios. Notably, the greatest performance gains by GAN-Semi-Supervised YOLO were achieved for rare lesions, addressing the critical class imbalance challenges in pathology.

We benchmarked our approach against a state-of-the-art pathology foundation model, UNI[25], which was used as a frozen feature extractor for a detection head. Our end-to-end fine-tuned YOLOv8 model demonstrated superior performance on our specific task. This result does not refute the power of foundation models, but rather highlights several key considerations for their application. Firstly, foundation models like UNI are typically pre-trained on vast archives of Hematoxylin-eosin (HE)-stained images, whereas our dataset utilizes PAS staining; this significant difference in staining protocol likely limits the direct transferability of features. Secondly, our results suggest that for a specific and fine-grained detection task, end-to-end fine-tuning of a task-oriented model can be more effective than relying on features from a general-purpose foundation model. This underscores the importance of considering the trade-offs between the vast but general knowledge of foundation models and the focused efficiency of specialized models, particularly for deployment in specific clinical workflows.

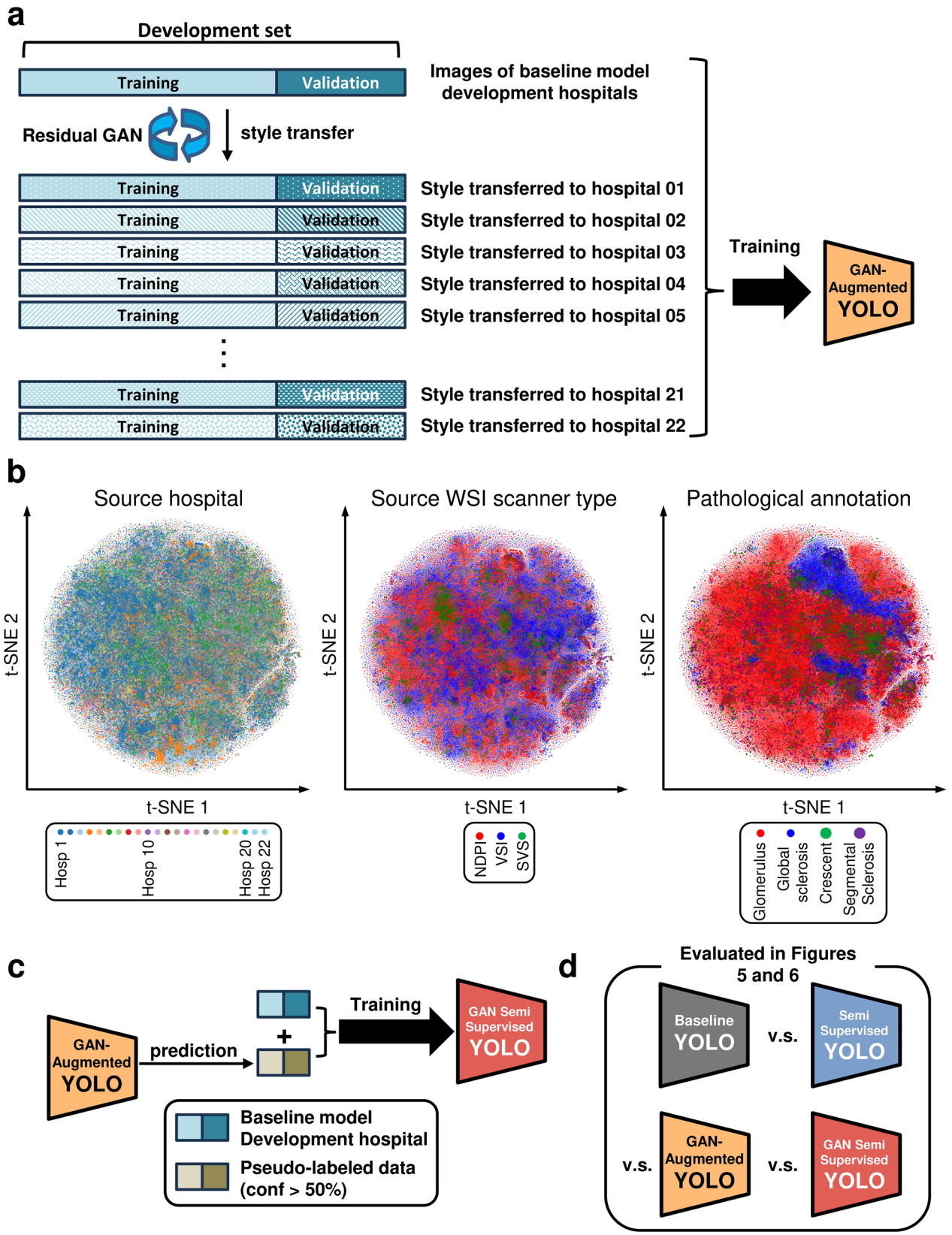


Fig. 4 | Integration of residual CycleGAN-based data adaptation with a semi-supervised learning strategy. a Residual CycleGAN-based data adaptation strategy for GAN-Augmented YOLO development. Images from a baseline hospital are transformed using residual CycleGANs to match the characteristics of all other hospitals in the development set. For each baseline hospital, 21 separate residual CycleGANs were developed for style transfer to each target hospital. The original and style-transferred images were combined to train a GAN-Augmented YOLO. **b** *t*-SNE visualization of features from all glomerular regions in GAN-augmented images, colored by source hospital (left), source scanner type (middle), and annotation type (right). **c** Semi-supervised learning process in which GAN-Augmented YOLOs generate pseudo-labels for original images from other hospitals. Images with confidence values above 50% were combined with original training data to create the GAN-Semi-Supervised YOLO model. **d** Evaluation strategy comparing different model configurations in Figs. 5 and 6.

Our findings have several practical implications: (1) effective AI systems can be developed with minimal expert annotation through a strategic combination of semi-supervised learning and domain adaptation, (2) addressing variations between scanner types is crucial when deploying pathology AI systems across institutions, (3) the computational efficiency of YOLOv8 enables deployment in standard clinical hardware, and (4) the efficacy of our approach for rare lesion detection addresses a critical gap in current pathology AI systems.

While our study presents promising results, it is important to acknowledge several limitations that also highlight important avenues for future research. Our approach demonstrated efficacy on PAS-stained glomerular lesions in kidney biopsies, but its broader generalizability requires further validation across different organs, pathological conditions, and other staining protocols. From a technical standpoint, we observed that our residual CycleGAN occasionally introduced minor visual artifacts on global and segmental sclerosis lesions. Although a blind review confirmed these were overwhelmingly acceptable for diagnosis, reducing these artifacts represents an opportunity for further performance enhancement. Furthermore, the practical deployment of this AI faces hurdles such as integration into clinical workflows and regulatory approval, and we have not yet established a definitive minimum required sample size for new institutions, as this likely depends on several factors. These limitations naturally guide future directions. Future work should focus on systematically validating the model on a wider variety of tissues and investigating minimum data requirements. Additionally, promising technical advancements could include exploring hybrid architectures that combine YOLOv8 with Transformers to better capture lesion context, linking the model's detections with clinical data to build prognostic models, and developing more lightweight model versions to enhance accessibility in resource-constrained settings[29–32].

In conclusion, our domain-adaptive semi-supervised learning approach effectively addresses the dual challenges of annotation scarcity and domain shifts in pathological image analysis, particularly regarding rare lesions. As demonstrated in renal pathology, our methodology may offer broadly applicable principles for developing efficient and generalizable AI systems across diverse clinical settings.

Table 3 | Visual assessment of residual CycleGAN-transformed images by nephrologists stratified by pathological class

	No artifacts (score 0 (%))	Clinically acceptable (score 1 (%))	Diagnostic impairment (score 2 (%))
Overall assessment			
	86.3 ± 3.3	12.7 ± 2.8	0.9 ± 0.4
Class-specific assessment			
Glomerulus	93.1 ± 2.1 Ref.	6.6 ± 1.9 Ref.	0.3 ± 0.1 Ref.
Global Sclerosis	**81.9 ± 4.4 *P* < 0.01**	**16.5 ± 3.5 *P* < 0.01**	1.6 ± 0.8 *P* = 0.32
Crescent	89.8 ± 2.5 *P* = 0.44	9.9 ± 2.3 *P* = 0.35	0.3 ± 0.2 *P* = 0.99
Segmental Sclerosis	**80.6 ± 4.3 *P* < 0.01**	**17.9 ± 4.1 *P* < 0.01**	1.5 ± 0.8 *P* = 0.32

Expert evaluation of morphological fidelity across 9240 transformed images by four independent nephrologists. Values represent mean ± standard deviation. Score 0: no artifacts; Score 1: minor artifacts, clinically acceptable; Score 2: significant artifacts with diagnostic impairment. Statistical comparisons were performed using Dunnett's test with the glomerulus as the reference group. Bold values indicate statistical significance. *Ref* reference category used as baseline for statistical comparison.

Methods

Study design and data acquisition

We conducted a retrospective analysis of native kidney biopsy specimens from patients aged ≥16 years who underwent renal biopsy between January 2014 and December 2018 in one of 22 Japanese hospitals (Supplementary Table 9). This study received central ethics approval from the Review Board of The University of Osaka Hospital (approval number: 17008-13), with subsequent institutional approval from each participating facility. An opt-out consent process was used. The study was conducted in accordance with the Declaration of Helsinki.

PAS-stained kidney biopsy sections were prepared according to the standard protocol at each participating hospital. The stained slides were digitized using one of three slide scanners: an Aperio ScanScope (Leica Biosystems, Wetzlar, Germany), a Hamamatsu NanoZoomer (Hamamatsu Photonics, Shizuoka, Japan), or a VS120 virtual slide microscope (Evident, Tokyo, Japan), producing images in SVS, NDPI, and VSI formats, respectively. Tissue samples from each biopsy procedure were embedded in a single paraffin block to generate a single WSI. The digitized WSIs were compressed to 1/8 of their original dimensions and converted to PNG format for lossless compression.

Ground-truth labeling and dataset preparation

Ground-truth labels were created using labelImg software by four expert nephrologists who classified all glomeruli into four categories:

1. Glomerulus (glomeruli without global or segmental sclerosis or crescent formation)
2. Global sclerosis (including both solidified and disappearing types)
3. Crescent (cellular and fibrocellular crescents)
4. Segmental sclerosis (sclerosis involving only a portion of the glomerular tuft)

Prior to the annotation process, these experts held a consensus meeting to establish detailed annotation guidelines, ensuring consistency across all

Table 4 | Visual assessment of residual CycleGAN-transformed images by nephrologists stratified by scanner type transformation

Pathological class	No artifacts (score 0 (%))	Clinically acceptable (score 1 (%))	Diagnostic impairment (score 2 (%))
NDPI to NDPI	86.1 ± 2.5 Ref.	12.4 ± 2.1 Ref.	1.5 ± 1.6 Ref.
NDPI to VSI	81.1 ± 4.0 *P* = 0.18	18.1 ± 3.7 *P* = 0.09	0.8 ± 0.9 *P* = 0.59
NDPI to SVS	79.7 ± 5.0 *P* = 0.08	**20.0 ± 4.6 *P* = 0.03**	0.3 ± 0.6 *P* = 0.28
VSI to VSI	88.7 ± 3.4 Ref.	10.6 ± 3.0 Ref.	0.7 ± 0.7 Ref.
VSI to NDPI	**79.7 ± 3.2 *P* < 0.01**	**18.2 ± 2.5 *P* < 0.01**	2.1 ± 1.7 *P* = 0.21
VSI to SVS	87.2 ± 2.2 *P* = 0.71	12.1 ± 1.9 *P* = 0.62	0.7 ± 0.7 *P* = 0.99
SVS to NDPI	95.9 ± 1.2	3.8 ± 1.0	0.3 ± 0.6
SVS to VSI	91.7 ± 2.6	7.8 ± 1.6	0.5 ± 1.1

Expert evaluation of morphological fidelity across scanner type transformations. Values represent mean ± standard deviation. Score 0: no artifacts; Score 1: minor artifacts, clinically acceptable; Score 2: significant artifacts with diagnostic impairment. Statistical comparisons performed using Dunnett's test with same-scanner transformations (NDPI to NDPI, VSI to VSI) as reference groups for each score category. SVS transformations were not statistically compared due to the absence of a within-SVS reference. Bold values indicate statistical significance. *Ref* reference category used as baseline for statistical comparison.

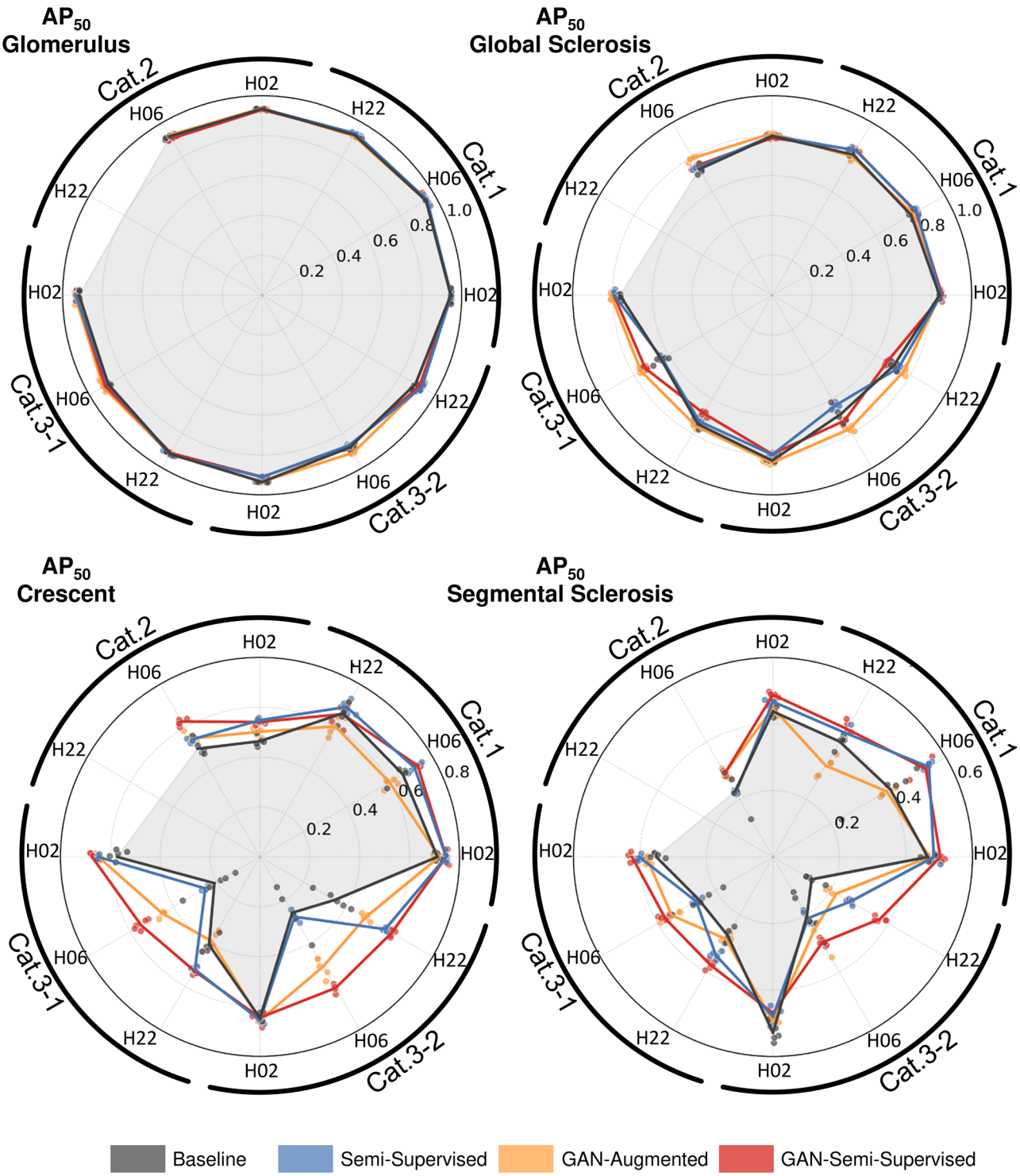
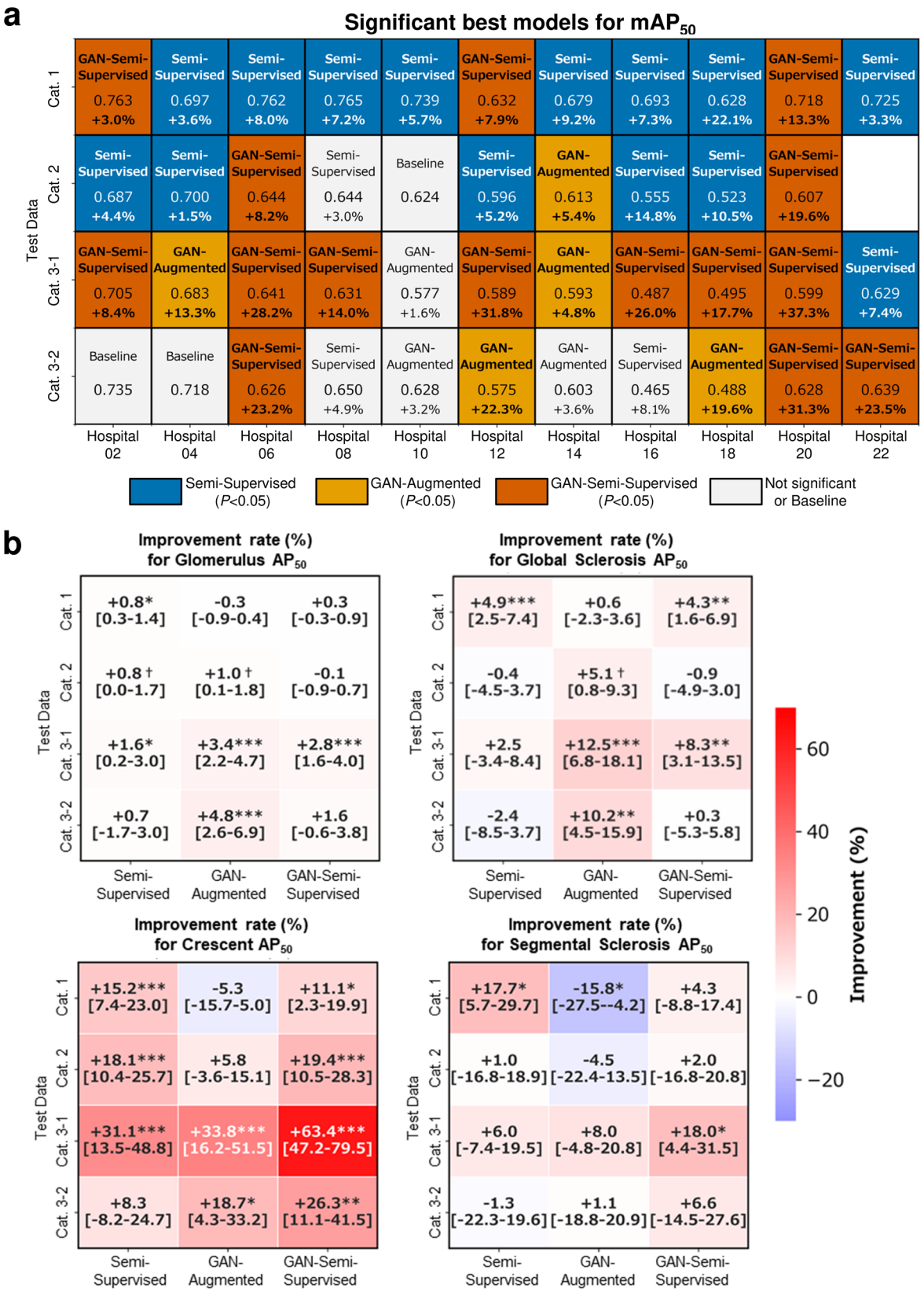


Fig. 5 | Performance comparison across different training strategies. Radar charts displaying AP_{50} values for each glomerular class (glomerulus, global sclerosis, crescent, and segmental sclerosis) for models based on Hospitals 02, 06, and 22. Each chart is divided into quadrants representing test categories: first quadrant (Cat. 1), second quadrant (Cat. 2), third quadrant (Cat. 3-1), and fourth quadrant (Cat. 3-2). Within each category, data points are arranged in the order specified by the hospital. The gray-filled area represents the Baseline YOLO performance, whereas the blue, yellow, and red lines indicate the Semi-Supervised, GAN-Augmented, and GAN-Semi-Supervised YOLO models, respectively. Mean values are connected by lines, with individual data points shown as dots. Note that the Category 2 evaluations for Hospital 22 are absent, as it was the sole hospital providing SVS format data.

annotators. To maintain annotation quality, a subset of images was cross-checked by the lead authors to assess inter-rater agreement.

The WSI dataset was divided into development and test sets at the WSI level. For feature analysis and domain shift assessment, annotated glomerular regions were cropped from PAS-stained WSIs, and 2048-dimensional feature vectors were extracted using an ImageNet pre-trained ResNet50 model. FID was computed between patches from different institutions using a pre-trained Inception v3 model with bootstrap sampling (1000 iterations, 100 images per iteration).



Model development and training

For the baseline supervised model development, WSIs from a single hospital were selected. We used three representative hospitals (one for each scanner type) to conduct hyperparameter tuning and establish optimized model configurations. For each hospital, five iterations of random training-validation splits with a 2:1 ratio were performed at the WSI level to train and validate the models.

We employed the standard YOLOv8l architecture without modifications, utilizing its anchor-free design that eliminates the need for custom anchor configurations. The YOLOv8l model was selected over YOLOv8x

Fig. 6 | Performance comparison of training strategies across hospitals and glomerular classes. a Color-coded matrix displaying the best-performing training strategy for each hospital-testing category combination. Each cell shows the strategy type (top), its mAP_{50} value (middle), and its percentage improvement over Baseline YOLO (bottom). Cell colors indicate statistically significant improvements ($P < 0.05$, Dunnett's test): blue for Semi-Supervised, yellow for GAN-Augmented, and orange for GAN-Semi-Supervised. Gray cells represent non-significant improvements or cases in which the baseline performed best. The testing categories are arranged vertically from Cat. 1 (self-facility) to Cat. 3-2 (different scanner types). **b** Heatmap visualization showing relative improvement rates (%) compared to baseline models across different test data categories (vertical axis) and training strategies (horizontal axis). Each cell displays the mean improvement rate with significance markers (upper value) and 95% confidence intervals (lower bracket). Statistical significance was assessed using Welch's *t* test with false discovery rate (FDR) correction within each test environment: $^{**}P < 0.001$, $^{*}P < 0.01$, $P < 0.05$ (FDR-corrected), $^{\dagger}P < 0.05$ (uncorrected only). Color intensity represents improvement magnitude using a blue-white-red scale, where blue indicates deterioration, white represents no change, and red indicates improvement.

due to memory constraints while maintaining comparable detection performance. Each model was trained using a single GPU with a batch size of 64 to ensure sufficient representation of rare lesions. Models were initialized with pre-trained weights from Common Objects in Context (COCO) datasets and trained for a maximum of 500 epochs with early stopping (patience = 50). Data augmentation included modified settings for hsv_h (0.2), degrees (90.0), and flipud (0.5), while maintaining default RandAugment settings to enhance model robustness across varied imaging conditions. YOLOv8 utilizes a composite loss function with three weighted components: CIoU loss (weight=7.5), Varifocal loss (weight=0.5), and Distribution focal loss (weight=1.5). The total training loss is computed as the weighted sum of these components.

For model evaluation, inference was performed using a confidence threshold of 0.001 and Non-Maximum Suppression with IoU threshold of 0.5. Performance metrics were calculated at AP_{50} following standard object detection evaluation protocols. All implementations utilized the Ultralytics YOLOv8 framework without custom modifications to the detection pipeline.

Semi-supervised learning

For each baseline model, predictions were generated for images from other institutions. Predictions exceeding predefined confidence thresholds were converted into pseudo-labels. Semi-supervised models were then developed by combining the original labeled data with these pseudo-labeled images and initialized with weights from their corresponding baseline models.

External dataset for semi-supervised learning

PAS-stained kidney biopsy WSIs were downloaded from the KPMP (https://www.kpmp.org/), accessed on July 7, 2024. A total of 487 SVS files were obtained from this external source. These images were processed using the same protocol applied to the multi-institutional hospital dataset. The processed KPMP images were then incorporated into the semi-supervised learning framework.

Benchmark comparison with a pathology foundation model

We conducted benchmark comparisons using UNI, a general-purpose foundation model trained on HE-stained pathology images[25]. We compared our standard YOLOv8l approach (COCO pretraining with end-to-end fine-tuning) against UNI as a frozen feature extractor backbone with a trainable YOLOv8l detection head. The UNI implementation utilized pre-trained weights from the official repository (https://github.com/mahmoodlab/UNI), with only detection components trained on our PAS-stained kidney biopsy dataset. For UNI integration with YOLOv8, we adapted the Yolo-DinoV2 framework (https://github.com/itsprakhar/Yolo-DinoV2.git), replacing the original Yolo-DinoV2/ultralytics/nn/modules/pre-trained_vit.py with our modified version available in our GitHub repository to enable UNI backbone integration. Training parameters remained identical between approaches to ensure fair comparison.

Domain adaptation with residual CycleGAN

To address scanner-specific variations, we implemented a residual CycleGAN to transform images between hospitals. To prevent excessive morphological alterations, we incorporated a residual loss term that constrains the magnitude of pixel-wise transformations:

$$L\_residual = \lambda\_residual \times (||G\_AB(A) - A||\_1 + ||G\_BA(B) - B||\_1)$$

where G_AB(A) represents the generated image when transforming real image A to domain B, G_BA(B) represents the generated image when transforming real image B to domain A, and $||\cdot||\_1$ denotes the pixel-wise L1 norm.

For each source hospital, 21 residual CycleGANs were trained to generate hospital-matched versions of the dataset. These GAN-transformed images retained their original annotation labels and were combined with the original images from the source hospital to train the GAN-Augmented YOLO model.

Visual assessment of residual CycleGAN-transformed images

To validate the morphological fidelity of residual CycleGAN-transformed images, we conducted a systematic visual assessment by four independent nephrologists experienced in renal biopsy interpretation. We randomly selected 5 representative glomerular images from each pathological class from each of the 22 hospitals, resulting in 20 original images per hospital and 440 total original images across all institutions. These glomerular images were cropped from patch images based on YOLO annotation boundaries. For domain adaptation validation, patch images containing these glomeruli were first transformed using the 21 residual CycleGAN models, and then the corresponding glomerular regions were cropped from the transformed patches, yielding 9240 CycleGAN-transformed glomerular images. The nephrologists, blinded to transformation details, evaluated each transformed image using a 3-point scoring system: Score 0 (no artifacts—diagnostically identical to original), Score 1 (minor artifacts present but clinically acceptable without diagnostic impairment), and Score 2 (significant artifacts causing diagnostic impairment).

Single-hospital simulation experiment

To investigate the impact of single-hospital limitations on cross-scanner generalization, we conducted simulation experiments using Hospital 02 (NDPI) and Hospital 06 (VSI) as source hospitals. We created single-hospital cross-format GAN-Semi-Supervised YOLO models and compared them with our main multi-hospital models.

(1) Multi-hospital cross-format models: used our main GAN-Semi-Supervised YOLO models from the primary analysis that incorporated all available hospitals.

(2) Single-hospital cross-format models: models created using only one hospital from the target scanner format (Hospital 02 source with all NDPI hospitals, Hospital 22 (SVS), and single VSI hospital (Hospital 06); Hospital 06 source with all VSI hospitals, Hospital 22 (SVS), and single NDPI hospital (Hospital 02)), mimicking the SVS single-hospital situation. The single-hospital models followed identical GAN-Semi-Supervised training procedures. Performance evaluation was conducted on test data from the single cross-format hospital used in the limited models (Hospital 06 for Hospital 02-based models, Hospital 02 for Hospital 06-based models).

Color-coded matrix visualization

A color-coded matrix visualization (Fig. 6a) was applied to identify the statistically superior model for each hospital-category combination. In this matrix, each cell represented a specific hospital-testing category pairing, displaying three key pieces of information: (1) the name of the best-performing model type (Baseline, Semi-Supervised, GAN-Augmented, or GAN-Semi-Supervised), (2) the absolute mAP_{50} value achieved by this model (average of five cross-validation seeds), and (3) the relative performance improvement percentage calculated as [(advanced model mAP_{50})–baseline mAP_{50}]/(baseline mAP_{50}) × 100%. Cells were color-coded only when the best model showed statistically significant improvement over the baseline ($P < 0.05$ by Dunnett's test).

We created heatmap visualizations to summarize percentage improvements by testing category for each glomerular class (Fig. 6b). Unlike the color-coded matrix visualization, which shows hospital-specific evaluations, this heatmap presents aggregate performance across all hospitals within each test category. Performance comparisons in Fig. 6b were assessed using Welch's t test, with relative improvement rates calculated as [(comparison model performance—baseline performance)/baseline performance] × 100%. Statistical significance was evaluated at $\alpha = 0.05$, with Benjamini–Hochberg false discovery rate (FDR) correction applied within each test data category.

Explainability analysis using gradient-weighted class activation mapping (Grad-CAM)

To investigate the mechanisms underlying performance improvements, we implemented Grad-CAM analysis using the YOLOv8 Explainer framework[33]. Grad-CAM visualizations were generated for representative cases demonstrating typical failure modes of Baseline YOLO and corresponding improvements achieved by GAN-Semi-Supervised YOLO. For each analyzed case, we extracted attention heatmaps overlaid on original PAS-stained images, with activation intensity represented by color gradients ranging from blue (low activation) to red (high activation). Ground truth annotations (GT) and model predictions (Pred) were displayed with bounding boxes, using abbreviations: Glo (glomerulus), GloSc (global sclerosis), Cres (crescent), and SegSc (segmental sclerosis).

Computing environment

A custom-built computer with a CPU (Ryzen Threadripper PRO 5975WX, Advanced Micro Devices, Santa Clara, CA, USA) and GPUs (RTX™ 6000 Ada or RTX™ A6000, 48 GB, NVIDIA Corporation, Santa Clara, CA, USA) was used for all calculations. Ubuntu 22.04 LTS was installed as the operating system. All methods were implemented using Python 3.8 and PyTorch 2.2.0.

Performance evaluation and statistical analysis

Statistical analyses were performed using the SciPy Statistics Library in Python, with a P value < 0.05 considered statistically significant. The specific statistical methods applied for each comparison are detailed in the respective figure and table legends.

Data availability

Clinical datasets are private and subject to restrictions to safeguard privacy. All the trained models are available at https://drive.google.com/drive/folders/15tVY1dzanf9ExYxcaNHeZ-n2UE2eQ8fp?usp=sharing and https://drive.google.com/drive/folders/1kVlOhKmIYmkOTSo1SgipJ2LJTIe_JswG?usp=sharing.

Code availability

All source code for the model and data preprocessing is available at https://github.com/NephrologyOsakaUniv/2025-Domain-adaptive-semi-supervised-learning-for-efficient-rare-pathological-lesion-detection.

Received: 11 May 2025; Accepted: 7 November 2025;
Published online: 23 November 2025

References

1. Tizhoosh, H. R. et al. Searching images for consensus: can AI remove observer variability in pathology? *Am. J. Pathol.* **191**, 1702–1708 (2021).
2. Burt, A. D., Lackner, C. & Tiniakos, D. G. Diagnosis and assessment of NAFLD: definitions and histopathological classification. *Semin. Liver Dis.* **35**, 207–220 (2015).
3. Satturwar, S. & Parwani, A. V. Artificial intelligence-enabled prostate cancer diagnosis and prognosis: current state and future implications. *Adv. Anat. Pathol.* **31**, 136–144 (2024).
4. Stacke, K., Eilertsen, G., Unger, J. & Lundstrom, C. Measuring domain shift for deep learning in histopathology. *IEEE J. Biomed. Health Inf.* **25**, 325–336 (2021).
5. Chen, X. et al. Recent advances and clinical applications of deep learning in medical image analysis. *Med. Image Anal.* **79**, 102444 (2022).
6. Yurt, M. et al. Semi-supervised learning of MRI synthesis without fully-sampled ground truths. *IEEE Trans. Med. Imaging* **41**, 3895–3906 (2022).
7. Cheplygina, V., de Bruijne, M. & Pluim, J. P. W. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* **54**, 280–296 (2019).
8. Berthelot, D. et al. MixMatch: a holistic approach to semi-supervised learning. *Adv. Neural Inf. Process Syst.* **32**, https://arxiv.org/abs/1905.02249 (2019).
9. Sohn, K. et al. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Adv. Neural Inf. Process Syst.* **33**, 596–608 (2020).
10. Arazo, E. et al. Pseudo-labeling and confirmation bias in deep semi-supervised learning. *Proceedings of the International Joint Conference on Neural Networks* https://doi.org/10.1109/IJCNN48605.2020.9207304 (2020).
11. Wilm, F. et al. Mind the gap: scanner-induced domain shifts pose challenges for representation learning in histopathology. *Proceedings - International Symposium on Biomedical Imaging* **2023-April** (2023).
12. Mehrtens, H. A., Kurz, A., Bucher, T. C. & Brinker, T. J. Benchmarking common uncertainty estimation methods with histopathological images under domain shift and label noise. *Med. Image Anal.* **89**, 102914 (2023).
13. Alajaji, S. A. et al. Generative adversarial networks in digital histopathology: current applications, limitations, ethical considerations, and future directions. *Mod. Pathol.* **37**, 100369 (2024).
14. Runz, M. et al. Normalization of HE-stained histological images using cycle consistent generative adversarial networks. *Diagn. Pathol.* **16**, 1–10 (2021).
15. Hetz, M. J., Bucher, T. C. & Brinker, T. J. Multi-domain stain normalization for digital pathology: A cycle-consistent adversarial network for whole slide images. *Med. Image Anal.* **94**, 103149 (2024).
16. Dong, J. et al. Constrained CycleGAN for effective generation of ultrasound sector images of improved spatial resolution. *Phys. Med. Biol.* **68**, 125007 (2023).
17. Yang, H. et al. Unpaired brain mr-to-ct synthesis using a structure-constrained cyclegan. *Lect. Notes Comput. Sci.* **11045 LNCS**, 174–182 (2018).
18. de Bel, T., Bokhorst, J. M., van der Laak, J. & Litjens, G. Residual cyclegan for robust domain transformation of histopathological tissue slides. *Med. Image Anal.* **70**, 102004 (2021).
19. Yang, Z. et al. A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. *Nat. Commun.* **16**, 2366 (2025).
20. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
21. Wang, X. et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 970–978 (2024).

22. Abe, M. et al. Self-supervised learning for feature extraction from glomerular images and disease classification with minimal annotations. *J. Am. Soc. Nephrol.* **36**, 471–486 (2025).
23. Redmon, J. S. D. R. G. A. F. (YOLO) you only look once. *Cvpr* 2016-December, 779–788 (2016).
24. Pyzer-Knapp, E. O. et al. Foundation models for materials discovery – current state and future directions. *npj Comput. Mater.* **11**, 1–10 (2025).
25. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
26. Fang, P., Feng, R., Liu, C. & Wen, R. Boundary sample-based class-weighted semi-supervised learning for malignant tumor classification of medical imaging. *Med. Biol. Eng. Comput.* **62**, 2987–2997 (2024).
27. Long, J., Ren, Y., Yang, C., Ren, P. & Zeng, Z. MDT: semi-supervised medical image segmentation with mixup-decoupling training. *Phys. Med. Biol.* **69**, (2024).
28. Yang, Q., Chen, Z. & Yuan, Y. Hierarchical bias mitigation for semi-supervised medical image classification. *IEEE Trans. Med. Imaging* **42**, 2200–2210 (2023).
29. Wang, S., Xia, C., Lv, F. & Shi, Y. RT-DETRv3: real-time end-to-end object detection with hierarchical dense positive supervision. 11–13 (2024).
30. Liu, S. et al. DAB-DETR: dynamic anchor boxes are better queries for DETR. *ICLR 2022 - 10th International Conference on Learning Representations* (2022).
31. Yan, Y. & Niu, K. Improved DN-DETR for safety helmet wearing detection. *2023 5th International Conference on Frontiers Technology of Information and Computer, ICFTIC* 874–877 https://doi.org/10.1109/ICFTIC59930.2023.10456362 (2023).
32. Carion, N. et al. End-to-end object detection with transformers. *Lect. Notes Comput. Sci.* **12346 LNCS**, 213–229 (2020).
33. Borah, P. P. S., Kashyap, D., Laskar, R. A. & Sarmah, A. J. A comprehensive study on explainable AI using YOLO and post hoc method on medical diagnosis. *J. Phys. Conf. Ser.* **2919**, 012045 (2024).

Acknowledgements

The authors thank Ms. Naoko Horimoto for technical assistance. This research was supported by grants from JSPS KAKENHI (19K0872, 20K17280, 21H02935), Health Labor Sciences Research Grant (K23FC1048a), Japan Agency for Medical Research and Development (J240705006), Japanese Society of Women Nephrologists, Osaka Kidney Foundation, The Takano Science Foundation, Manpei Suzuki Diabetes Foundation, Shimazu Science Foundation, and Nishikawa Medical Foundation.

Author contributions

I.M.: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing–original draft, visualization, funding acquisition. A.M.: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing–original draft, funding acquisition. A.I.: resources, data curation, writing–review & editing, visualization. H.O.: resources, data curation, writing–review & editing, visualization. H.N.: conceptualization, methodology, writing–review & editing. M.A.: conceptualization, methodology, data curation, validation, writing–review & editing. N.T.: resources, data curation, writing–review & editing, visualization. H.N., E.K., E.U., T.S., T.I., Y.Y., H.K., K.M., T.H., Y.N., K.F., Y.K., S.K., R.K., M.K., Y.K., M.O., Y.K., H.A., Y.M., T.K., Y.U., N.F., M.T., A.S., K.N., K.O., Y.A., M.I., T.K., T.T., I.M., T.K., T.N., K.T., Y.I., T.M., O.Y., S.Y., H.K., K.F., K.F., S.A., T.M., N.T., H.Y., A.S., T.U., S.M., M.Y., M.N., and R.Y.: resources, project administration, writing–review & editing. K.I.: conceptualization, methodology, investigation, resources, data curation, writing–review & editing. Y.I.: conceptualization, writing–review & editing, supervision, project administration, funding acquisition.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-02160-6.

Correspondence and requests for materials should be addressed to Isao Matsui.

[1]Department of Nephrology, Graduate School of Medicine, The University of Osaka, Suita, Osaka, Japan. [2]Transdimensional Life Imaging Division, Institute for Open and Transdisciplinary Research Initiatives, The University of Osaka, Suita, Osaka, Japan. [3]Subcommittee of AI/ICT Infrastructure Construction, Japanese Society of Nephrology, Bunkyo-ku, Tokyo, Japan. [4]Data-Driven Innovation Initiative, Kyushu University, Higashi-ku, Fukuoka, Japan. [5]Department of Nephrology and Hypertension, Kawasaki Medical School, Kurashiki, Okayama, Japan. [6]Health Data Science, Kawasaki Medical School, Kurashiki, Okayama, Japan. [7]Department of Nephrology, Kyoto University Graduate School of Medicine, Kyoto, Kyoto, Japan. [8]Department of Biomedical Data Intelligence, Kyoto University Graduate School of Medicine, Kyoto, Kyoto, Japan. [9]Department of Cardiorenal and Cerebrovascular Medicine, Kagawa University, Kita-gun, Kagawa, Japan. [10]Department of Nephrology, National Hospital Organization Chiba Medical Center Chibahigashi National Hospital, Chiba, Chiba, Japan. [11]Department of General Medicine, Juntendo University Faculty of Medicine, Bunkyo-ku, Tokyo, Japan. [12]Department of Nephrology and Rheumatology, Aichi Medical University, Nagakute, Aichi, Japan. [13]Department of Cardiology, Pulmonology, Hypertension and Nephrology, Ehime University Graduate School of Medicine, Toon, Ehime, Japan. [14]Department of Nephrology, Higashiosaka City Medical Center, Higashiosaka, Osaka, Japan. [15]Department of General Internal Medicine, Hyogo medical University, Nishinomiya, Hyogo, Japan. [16]Department of Nephrology, Kanazawa Medical University School of Medicine, Kahoku, Ishikawa, Japan. [17]Division of Nephrology, Department of Medicine, Kurume

University School of Medicine, Kurume, Fukuoka, Japan. [18]Department of Nephrology, Nagoya University Graduate School of Medicine, Nagoya, Aichi, Japan. [19]Division of Clinical Nephrology and Rheumatology, Niigata University Graduate School of Medical and Dental Sciences, Niigata, Niigata, Japan. [20]Division of Nephrology, Endocrinology and Metabolism, Tokai University School of Medicine, Isehara, Kanagawa, Japan. [21]Department of Nephrology, Toyonaka Municipal Hospital, Toyonaka, Osaka, Japan. [22]Department of Nephrology, Kindai University Nara Hospital, Ikoma, Nara, Japan. [23]Artificial Intelligence and Digital Twin in Healthcare, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku, Tokyo, Japan. [24]Department of Pathology, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku, Tokyo, Japan. [25]Department of Medicine and Clinical Science, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan. [26]Department of Nephrology, Nara Medical University, Kashihara, Nara, Japan. [27]Department of Kidney Disease and Hypertension, Osaka General Medical Center, Osaka, Japan. [28]Department of Nephrology, Hyogo Prefectural Nishinomiya Hospital, Nishinomiya, Hyogo, Japan. [29]Department of Internal Medicine, Japan Community Health Care Organization Osaka Hospital, Osaka, Osaka, Japan. [30]Department of Nephrology, Osaka Rosai Hospital, Sakai, Osaka, Japan. [31]Department of Pathology, Hyogo Prefectural Nishinomiya Hospital, Nishinomiya, Hyogo, Japan. [32]Department of Nephrology, National Hospital Organization Osaka Minami Medical Center, Kawachinagano, Osaka, Japan. [33]Division of Kidney and Dialysis, Department of Internal Medicine, Kansai Rosai Hospital, Amagasaki, Hyogo, Japan. [34]Department of Nephrology, Japanese Red Cross Otsu Hospital, Otsu, Shiga, Japan. [35]Department of Nephrology and Dialysis, Medical Research Institute Kitano Hospital, PIIF Tazuke-Kofukai, Osaka, Osaka, Japan. [36]Division of Nephrology and Endocrinology, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku, Tokyo, Japan. [37]Division of Kidney, Dialysis and Cardiology, Department of Internal Medicine, Hyogo medical University, Nishinomiya, Hyogo, Japan. [38]Department of Nephrology, School of Medicine, Wakayama Medical University, Wakayama, Wakayama, Japan. [39]Department of Nephrology, Hiroshima University Hospital, Hiroshima, Hiroshima, Japan. [40]Department of Nephrology, Fujita Health University School of Medicine, Toyoake, Aichi, Japan. [41]Department of Analytic Human Pathology, Graduate School of Medicine, Nippon Medical School, Tokyo, Japan. [42]Institute for the Advanced Study of Human Biology (ASHBi), Kyoto University, Kyoto, Japan. [43]Health and Counseling Center, The University of Osaka, Suita, Osaka, Japan. [44]These authors contributed equally: Isao Matsui, Ayumi Matsumoto. ✉e-mail: matsui@kid.med.osaka-u.ac.jp