Eliminating Staining and Scanner Variability by Fine-Tuning AI for Inflammatory Cell Detection in Kidney Biopsies

Lisanne Huisman, Luuk Neervens, Martina Baricocchhi, Femke Aminetzah

Radboud University Nijmegen, The Netherlands
Email: lisanne.huisman@ru.nl, luuk.neervens@ru.nl, martina.baricocchhi@ru.nl, femke.aminetzah@ru.nl

***Abstract*—Accurate and robust detection of inflammatory cells in kidney transplant biopsies is crucial for assessing the risk of rejection and guiding treatment. However, acute and chronic rejection mediated by cellular immune infiltrates poses a significant threat to longevity. Pathologists assess these inflammatory infiltrates using the internationally adopted Banff classification grades, which reflect the mononuclear inflammatory cell burden across different renal compartments. These scores directly drive medical decision-making. Whole-slide imaging (WSI) of PAS-stained biopsies enables CNN-based tools to achieve pathologist-level performance, but stain and scanner variability remain a challenge.**

In this work, we train different Faster R-CNN models on PAS-Original and PAS-Diagnostic data and evaluate its performance on PAS-CpG data, to quantify transfer learning capabilities for inflammatory cell detection.

Fine-tuning on a small subset of PAS-CpG data (Model B) substantially improved recall (from 68% to 79%) and F1-score (from 0.63 to 0.66) compared to the baseline (Model A), approaching the performance of a fully supervised PAS-CpG model (Model C, F1 = 0.69). These findings demonstrate that lightweight adaptation is an effective strategy for mitigating staining variability, supporting the development of clinically robust, generalizable AI tools for renal transplant pathology.

***Index Terms*—Kidney transplantation, Banff scoring, domain adaptation, Faster R-CNN, automated inflammatory cell detection, whole-slide imaging.**

I. Introduction

Kidney transplantation remains the most effective treatment for end-stage renal disease. However, graft longevity is often compromised by acute and chronic rejection, which is driven by the infiltration of immune cells.

The Banff classification system is used to semi-quantitatively assess the presence of mononuclear inflammatory cells in various renal compartments. Since clinical diagnosis and treatment decisions rely heavily on biopsy lesion scores (BLS), objectivity and consistency in scoring are critical. Yet, the Banff system suffers from limited reproducibility and is time-consuming in routine pathology workflows. Consequently, automated biopsy assessment methods hold significant promise for reducing the workload of pathologists while improving the reliability and standardization of evaluations.

Manual cell counting is both labor-intensive and subject to inter- and intra-observer variability [1], [2]. Automated tools can provide more consistent assessments and have the potential to improve diagnostic accuracy and patient outcomes.

Whole-slide imaging (WSI), combined with convolutional neural networks (CNNs), has shown the ability to match or exceed pathologist-level performance in large-scale studies of histopathology [3]. However, these models often struggle with inter-scanner and inter-stain variability, which can lead to significant performance degradation when applied to slides prepared using different protocols [4].

In this study, we investigate the impact of inter-staining and inter-scanner variability on the detection of inflammatory cells in renal biopsies. Specifically, we examine how these domain shifts affect model performance across PAS-stained slides and PAS CpG slides.

In addition to highlighting the challenges posed by domain variability, we demonstrate that lightweight fine-tuning significantly improves model robustness. This approach mitigates performance loss due to domain shifts and supports the development of generalizable, cross-institutional models.

Such tools could accelerate diagnosis, enable early treatment decisions, and facilitate scalable deployment in clinical practice, including in under-resourced settings.

A. Study Objectives

This study aims to enhance the safety and effectiveness of kidney transplantation by improving the robustness and generalizability of automated inflammatory cell detection in histopathological slides. Specifically, we aim to:

1) Assess the transfer learning performance of a detector trained on PAS-Original/Diagnostic slides when applied to PAS-CpG slides.

2) Quantify the performance gain achieved through lightweight fine-tuning on a subset of PAS-CpG data.

3) Establish an upper-bound reference by training and evaluating a model exclusively on PAS-CpG slides.

II. Background

In kidney transplant recipients, the Banff classification system is the clinical standard for assessing the severity of graft rejection, based on the density and distribution of mononuclear leukocytes (MNLs), including lymphocytes and monocytes, within renal compartments. Elevated MNL infiltration indicates immune-mediated injury and is directly associated with reduced graft survival. As such, accurate and reproducible quantification of inflammation plays a crucial role in guiding immunosuppressive treatment and clinical decision-making.

Pathologists traditionally process renal biopsies using histological staining, such as Periodic Acid-Schiff (PAS), and manually evaluate them. However, this manual Banff scoring approach is labor-intensive and prone to both inter- and intra-observer variability, limiting its scalability and reliability in routine practice [1], [2].

With the advent of whole-slide imaging (WSI), labs can now digitalize histological slides at high resolution, enabling computational analysis. Convolutional neural network (CNN)-based methods have achieved pathologist-level performance in tasks such as segmenting renal structures and detecting glomerulosclerosis and inflammation [1].

However, these models often exhibit performance degradation when applied to slides that differ in scanner type or staining protocol. For example, lymphocyte segmentation performance can decrease by 5–10% when models are evaluated on slides from a different scanner, unless domain adaptation or normalization techniques are employed [4].

Digital pathology and artificial intelligence (AI) offer promising solutions for automating and standardizing the detection of inflammatory cells. Earlier approaches relied on heuristic image processing methods such as thresholding or color segmentation, but these were limited in robustness to staining variation and cell clustering.

Modern deep learning methods, in contrast, can learn complex morphological features of MNLs directly from annotated data, enabling more accurate and efficient detection [5].

For instance, Hermsen et al. [1] applied CNNs to quantify inflammation and chronic lesions in renal transplant biopsies and reported strong correlations between AI-generated scores and pathologist-assigned Banff grades.

Similarly, Jacq et al. [2] demonstrated that deep learning models could accurately evaluate interstitial inflammation and capillaritis, highlighting the potential for AI to support standardized rejection diagnostics.

To enhance model generalization across domains, transfer learning has been widely adopted. This approach involves initializing a model with pretrained weights—often from large-scale datasets like ImageNet—and then fine-tuning it on task-specific medical data.

He et al. [6] found that ImageNet-pretrained ResNet architectures matched or exceeded the performance of models trained from scratch in both convergence speed and accuracy on medical imaging tasks.

Similarly, Sharma and Maji [7] demonstrated that transfer learning significantly improved classification performance across organ types and histology tasks.

In renal pathology, Li et al. [8] trained a U-Net–based segmentation model on frozen-section WSIs, using VGG16 encoder weights pretrained on ImageNet. Their transfer learning approach achieved higher Dice coefficients and more robust generalization on external test sets compared to randomly initialized models.

In these cases, early convolutional layers were typically frozen to preserve general visual features, while deeper layers were fine-tuned for domain-specific learning. This strategy not only reduced overfitting but also accelerated convergence [7], [8].

For the current study, we adopt the Faster R-CNN architecture—a two-stage object detection framework consisting of a region proposal network (RPN) followed by classification and bounding box regression. Faster R-CNN is particularly well-suited to histopathology applications due to its ability to detect small, densely packed objects such as nuclei and immune cells. Its multi-scale feature extraction via feature pyramid networks (FPNs) and robustness to class imbalance make it a strong candidate for MNL detection [9]–[12].

These characteristics, combined with its extensive validation in medical imaging, motivated our selection of Faster R-CNN for this study.

III. Methods

A. Implementation Details

The implementation used the following Python packages:

- Standard libraries: `re`, `json`, `random`, `pathlib`, `glob`, `gc`, `os`, `shutil`, and `pickle`
- Numerical computing: NumPy (v1.21+)
- Image processing: Pillow/PIL (v8.0+), `rasterio` (v1.2+), and `tifffile` (v2021+)
- Deep learning: PyTorch and TorchVision (v1.10+ with `FasterRCNN_ResNet50_FPN_Weights`)
- Visualization: Matplotlib (v3.4+)
- Utilities: `tqdm` (v4.60+)

The whole model pipeline can be seen in figure 8.

B. Data

The dataset comprises 153 whole slide images (WSIs) collected from six pathology departments across Europe. A total of 231 regions of interest (ROIs) have been annotated using dot markers to identify inflammatory cells as either monocytes or lymphocytes. Each ROI spans an area of approximately 0.32 ± 0.22 mm^2. On average, each WSI contains an estimated 350 lymphocyte and 180 monocyte annotations, providing a rich dataset for training and evaluation of immune cell detection models. To ensure balanced learning, the training set was constructed to include a similar number of lymphocyte and monocyte examples. This was done to prevent the model from becoming biased toward a particular cell type when detecting inflammatory cells.

As discussed before, each PAS slide has three staining versions: a CpG profile, a diagnostic profile, and an original profile (see figure 1). These staining variations are a critical component of this research, allowing the model to learn from and adapt to differences in staining protocols across institutions.

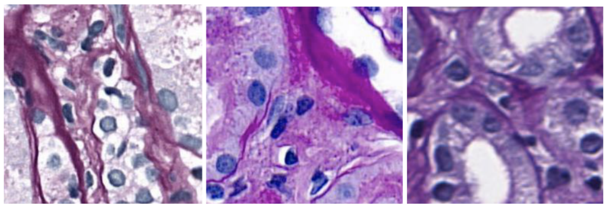


Fig. 1: The three PAS staining types used in this study. From left to right, CpG staining, diagnostic staining, and original staining.

C. Pre-processing

This study utilizes whole slide images (WSIs), which are high-resolution tissue scans commonly used in computational pathology. Due to their extremely large size, directly using WSIs for model training is computationally expensive and memory-demanding. To mitigate this, a pre-processing pipeline was developed to reduce the data size while preserving the structural and biological relevance of the tissue.

The pipeline begins by extracting regions of interest (ROIs) based on annotated coordinates of inflammatory cells, including both lymphocytes and monocytes. We use each ROI as the center of a patch measuring 256 × 256 pixels. We then export these patches as PNG images to streamline downstream processing and reduce memory requirements.

For testing, a similar strategy is applied. Only here, patches of size 256 by 256 are randomly selected from the WSI and are assigned a unique identifier, which can be used to trace back where the patch appears in the original image. After obtaining the detection results as pixel coordinates per testing patch, the patch coordinates can be translated into the WSI coordinates, yielding the correct output format for computing the performance metrics. We then compare the predicted inflammatory cell coordinates to the closest recorded ROI coordinate as can be found in the json pixel file for that respective WSI. We denote predictions that fall within 5 micrometers of the ground truth cell coordinates by the true positive rate. Incorrectly detected inflammatory cells are denoted by the false positive rate. The false negative rate refers to the missed detection of inflammatory cells.

This patch-based approach enables efficient model training, focusing specifically on regions of interest that contain inflammatory cells without compromising critical spatial information.

D. Model Architecture

To detect inflammatory cells in histology images, we employed the Faster R-CNN object detection architecture. Faster R-CNN is a two-stage deep learning model that first proposes candidate object regions and then classifies and refines those regions [13].

In the first stage, a Region Proposal Network (RPN) slides over the feature map and predicts bounding boxes and objectness scores. By sharing information with the detector, the RPN efficiently generates high-quality proposals after initial feature extraction.

In the second stage, the model refines these proposals using a classifier head that outputs the object class and adjusted bounding boxes.

Additionally, we employ a ResNet-50 backbone with a Feature Pyramid Network (FPN) to handle the scale diversity of cells. FPN enriches multi-scale feature maps by combining low-level detail with high-level semantic information, ultimately improving the detection of small objects like inflammatory nuclei (10–15 µm).

Given the sparsity and small size of these cells, Faster R-CNN with FPN is ideal, as it generates anchors at multiple scales and focuses on subtle features, allowing the RPN to filter out the background while the second stage verifies likely cell regions. This method performs even when most patches contain few or no cells.

Overall, the model's proposal refinement and non-maximal suppression strategies make it particularly effective for sparse object detection tasks.

E. Experimental Design: Models A, B, and C

Three variants of the Faster R-CNN model were trained and evaluated to systematically analyze generalization and domain adaptation effects across different PAS stain domains: Model A (baseline), Model B (fine-tuned), and Model C (fully trained on target domain).

Model A was trained from scratch using patches derived from the PAS-Original and PAS-Diagnostic cohorts. The training set consisted of 6322 patches (80% split), while validation was performed on 1581 patches (20% split). The model was optimized over 10 epochs using stochastic gradient descent (SGD) with an initial learning rate of 0.005, momentum of 0.9, and weight decay of 0.0005. After five epochs, the learning rate was reduced by a factor of 0.1 via a StepLR scheduler. A batch size of 4 was used.

Upon completion of training, Model A was evaluated on the PAS-CpG test set, consisting of 2025 randomly selected patches, without any further adaptation to the target domain.

Model B reused the pretrained weights of Model A and underwent a fine-tuning phase on PAS-CpG patches. Fine-tuning employed a small subset of PAS-CpG data with balanced sampling of monocyte and lymphocyte instances to mitigate class imbalance. The model was fine-tuned for an additional 5 epochs at a reduced learning rate of 1e-4, allowing adaptation to the target domain without erasing the general representations learned on the source domains.

The architecture, anchor strategy, and post-processing pipeline remained identical to those of Model A to ensure a controlled comparison of adaptation effects.

Following fine-tuning, Model B was evaluated on the same PAS-CpG test set used for Model A.

Model C was trained exclusively on PAS-CpG data. The training set included 6248 patches (80% split), and validation was performed on 1562 patches (20% split). The model was trained using the same hyperparameters and optimization settings as Models A and B.

Evaluation of Model C was likewise conducted on the PAS-CpG test set (2025 patches), establishing an upper-bound performance benchmark for fully supervised training on the target domain.

All models used the Faster R-CNN architecture with a ResNet-50 + FPN backbone. A consistent training pipeline was applied across experiments, including loss functions and evaluation metrics.

The comparative analysis of Models A, B, and C quantified the adaptation gain achievable through lightweight fine-tuning. It provided insights into the generalization capabilities of object detectors across stain domains without explicit domain alignment.

F. Coordinate Mapping and Visualization

We reconstructed the model's predictions back to tissue coordinates by retaining the original spatial location of each patch within the whole-slide image (WSI). This allowed us to accurately map detected cells onto their respective positions within the tissue context.

For visual feedback, the inflammatory cell predictions were displayed as red bounding boxes, 32 by 32 pixels in size, overlaid on each patch.

Additionally, ground truth coordinates were annotated using green 32 by 32 pixel boxes to visualize true positive, false positive, and false negative detections.

This double-overlay visualization enabled easy assessment and confirmation of the model's performance, facilitating the identification of systematic detection errors, such as over-detection in dense regions.

G. Evaluation Metrics

We evaluated model performance by comparing predicted and ground truth bounding boxes based on their spatial overlap.

We calculated precision and recall using Intersection over Union (IoU)–based matching, with a threshold of 0.5. This means a predicted bounding box was considered a true positive if it overlapped with a ground truth box by at least 50%. Given the pixel resolution, this corresponds approximately to a spatial tolerance of 5 µm, accounting for slight positional variations in cell annotations.

The F1-score, the harmonic mean of precision and recall, was also computed to provide a balanced measure of detection performance.

IV. Results

We evaluated model performance on the PAS-CpG test set, which consists of 2025 patches.

Table I summarizes the precision, recall, and F1-score achieved by the baseline Model A (baseline), Model B (fine-tuned on PAS-CpG) and Model C (fully trained on PAS-CpG), given a prediction score threshold of 0.55.

Figures 2–4 (2–4) depict the training-loss curves for Models A, B and C respectively, while Figures 5–7 (5–7) show the corresponding FROC plots.

TABLE I: Detection performance on the PAS-CpG test set

Model	**Precision**	**Recall**	**F1-score**
Model A (baseline)	0.5872	0.6817	0.6309
Model B (fine-tuned on PAS-CpG)	0.5659	0.7871	0.6584
Model C (trained on PAS-CpG)	0.6300	0.7502	0.6849

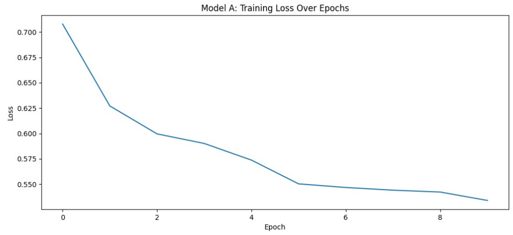


Fig. 2: Training loss curve for Model A (baseline). The curve shows the training loss (y-axis) versus the number of training epochs (x-axis).

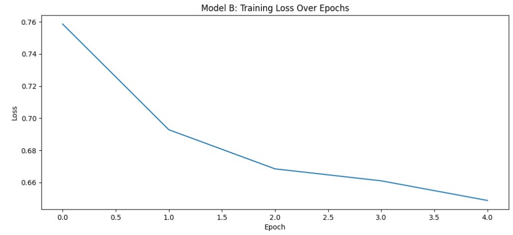


Fig. 3: Training loss curve for Model B (finetuned on PAS-CpG). The curve shows the training loss (y-axis) versus the number of training epochs (x-axis).

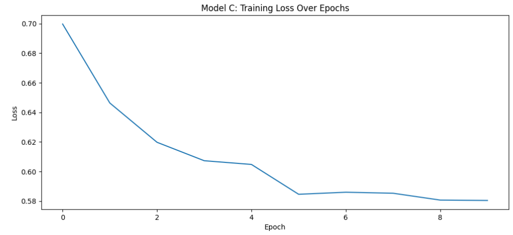


Fig. 4: Training loss curve for Model C (trained on PAS-CpG). The curve shows the training loss (y-axis) versus the number of training epochs (x-axis).

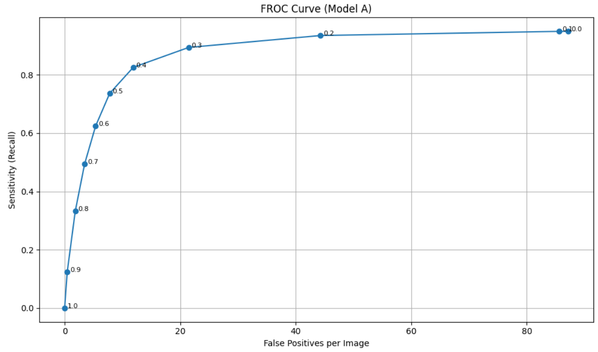


Fig. 5: FROC curve for Model A (baseline). The curve shows the sensitivity (y-axis) versus the false positives per image (x-axis) for different detection thresholds of 1.0 to 0.1.

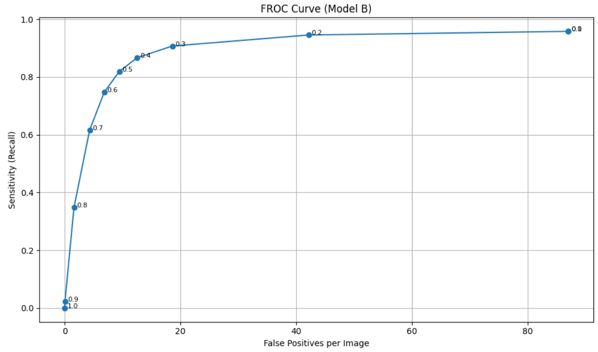


Fig. 6: FROC curve for Model B (finetuned on PAS-CpG). The curve shows the sensitivity (y-axis) versus the false positives per image (x-axis) for different detection thresholds of 1.0 to 0.1.

V. Discussion

A. Interpretation of Results

Model C achieved a higher precision, recall, and F1-score compared to the baseline Model A, confirming the benefit of domain-specific training on PAS-CpG data. In particular, recall improved from 68.17% to 75.02%, indicating fewer missed inflammatory cells. Precision also increased, demonstrating improved localization accuracy.

Fine-tuning the baseline model (Model B) on a subset of PAS-CpG data led to a notable improvement in recall, increasing from 68.17% in Model A to 78.71%, thus reducing the number of missed inflammatory cells. However, precision slightly decreased from 58.72% to 56.59%, resulting in a modest overall F1-score improvement (from 0.6309 to 0.6584).

The training loss curves (figures 2 and 4) show stable convergence across 10 epochs for all models.

This study set out to quantify how well AI models can generalize inflammatory cell detection across different PAS stain domains and whether lightweight fine-tuning can bridge the domain gap. The results clearly demonstrate the impact of domain shift on inflammatory cell detection performance. When tested on PAS-CpG slides, Model A exhibited moderate precision (59.3%) and recall (67.9%), consistent with prior reports that cross-stain variability reduces generalization.

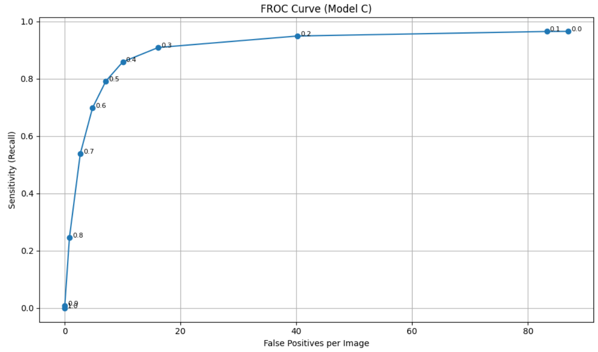


Fig. 7: FROC curve for Model C (trained on PAS-CpG). The curve shows the sensitivity (y-axis) versus the false positives per image (x-axis) for different detection thresholds of 1.0 to 0.1.

Model B underwent 5 additional fine-tuning epochs on PAS-CpG data, which led to its convergence as well (see figure 3).

Fine-tuning Model B on a small subset of PAS-CpG data substantially improved recall to 78.6%, with a modest increase in F1-score, confirming that lightweight domain adaptation can mitigate stain-related performance degradation.

Notably, fine-tuning resulted in a slight decrease in precision, which may indicate that the model is becoming more sensitive to subtle cell features at the expense of increased false positives.

Model C, trained fully on PAS-CpG, achieved the highest precision (63.6%) and recall (75.0%), establishing an upper performance bound. Together, these findings support the original hypothesis that fine-tuning enhances cross-domain robustness, but full target-domain training remains optimal when sufficient labeled data are available.

B. Limitations

We acknowledge the following limitations to provide a balanced interpretation of the findings:

Firstly, the current approach for extracting patches could be improved and standardized. As mentioned in section III-C, train patches were constructed around an inflammatory cell, meaning that the middle point of the patch always contained an inflammatory cell. Conversely, test patches were constructed in a more random fashion where patches always contain at least one inflammatory cell not necessarily in the center. In hindsight, the latter approach should have been taken for training patch construction as well. However, due to time constraints we were unable to re-run it on the training patches. Furthermore, we did not include any empty patches (i.e. patches without inflammatory cells) in the constructed testing set. This could have aided in a more robust model performance evaluation.

Moreover, considerable resource limitations restricted the number of epochs and training samples per epoch that could be utilized, which may have impacted the model's performance. This also meant that it was not feasible to perform hyperparameter optimization for the model, implying the chosen hyperparameters might not be optimal.

Lastly, it is worth noting that the original TIF files were converted to PNGs during the patching procedure. This conversion resulted in a loss of resolution and likely impaired the model's ability to distinguish between inflammatory cells and healthy ones accurately.

C. Future Work

Future work should address several limitations identified in this study.

First, standardizing the patch extraction process across training and testing is a priority, as differences in how patches were sampled likely introduced bias. Using randomized sampling strategies that reflect real-world data distributions will provide a more robust evaluation framework.

Second, expanding the training set and increasing the number of training epochs, combined with systematic hyperparameter optimization, could further improve model performance and stability.

Furthermore, exploring alternative model architectures and backbones, and comparing their performances across models A, B, and C, could provide valuable insights and potentially enhance the results.

Finally, training directly on original-resolution images could allow the model to capture finer morphological details, potentially improving discrimination between inflammatory and healthy cells.

VI. Conclusion

This study addressed the critical question of how to enhance the robustness and generalizability of deep learning models for detecting inflammatory cells across PAS-stained domains in renal biopsy slides.

The Faster R-CNN architecture demonstrated strong baseline performance. Still, we observed a clear domain shift when applying a model trained on PAS-Original and PAS-Diagnostic slides (Model A) to PAS-CpG data.

Introducing lightweight fine-tuning on a small subset of PAS-CpG slides (Model B) resulted in significant improvements in recall and overall F1-score, highlighting the value of domain adaptation for mitigating performance degradation.

Model C achieved the best performance through full training on PAS-CpG, confirming that while fine-tuning offers practical gains when target domain data is limited, full target-domain training remains the ideal approach when feasible.

References

[1] M. Hermsen, F. Ciompi, A. Adefidipe, A. Denic, A. Dendooven, B. H. Smith, D. van Midden, J. H. Bräsen, J. Kers, M. D. Stegall *et al.*, "Convolutional neural networks for the evaluation of chronic and inflammatory lesions in kidney transplant biopsies," *The American Journal of Pathology*, vol. 192, no. 10, pp. 1418–1432, 2022.

[2] A. Jacq, G. Tarris, A. Jaugey, M. Paindavoine, E. Maréchal, P. Bard, J.-M. Rebibou, M. Ansart, D. Calmo, J. Bamoulid *et al.*, "Automated evaluation with deep learning of total interstitial inflammation and peritubular capillaritis on kidney biopsies," *Nephrology Dialysis Transplantation*, vol. 38, no. 12, pp. 2786–2798, 2023.

[3] G. Campanella, M. G. Hanna, L. Geneslaw, and et al., "Clinical-grade computational pathology using weakly supervised deep learning on whole-slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[4] A. Khan, A. Janowczyk, F. Müller, A. Blank, H. G. Nguyen, C. Abbet, L. Studer, A. Lugli, H. Dawson, J.-P. Thiran *et al.*, "Impact of scanner variability on lymph node segmentation in computational pathology," *Journal of pathology informatics*, vol. 13, p. 100127, 2022.

[5] J. Hung, A. Goodman, D. Ravel, S. C. P. Lopes, G. W. Rangel, O. A. Nery, B. Malleret, F. Nosten, M. V. G. Lacerda, M. U. Ferreira, L. Rénia, M. T. Duraisingh, F. T. M. Costa, M. Marti, and A. E. Carpenter, "Keras R-CNN: library for cell detection in biological images using deep neural networks," *BMC Bioinformatics*, vol. 21, no. 1, 7 2020. [Online]. Available: https://doi.org/10.1186/s12859-020-03635-x

[6] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/eb1e78328c46506b46a4ac4a1e378b91-Paper.pdf

[7] Y. Sharmay, L. Ehsany, S. Syed, and D. E. Brown, "Histotransfer: understanding transfer learning for histopathology," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2021, pp. 1–4.

[8] X. Li, R. C. Davis, Y. Xu, Z. Wang, N. Souma, G. Sotolongo, J. Bell, M. Ellis, D. Howell, X. Shen *et al.*, "Deep learning segmentation of glomeruli on kidney donor frozen sections," *Journal of Medical Imaging*, vol. 8, no. 6, pp. 067 501–067 501, 2021.

[9] I. Ahmed and R. Das, "Comparative analysis of yolo and faster r-cnn models for detecting traffic object." *International Journal of Advanced Computer Science & Applications*, vol. 16, no. 3, 2025.

[10] M. Carranza-García, J. Torres-Mateo, P. Lara-Benítez, and J. García-Gutiérrez, "On the performance of one-stage and two-stage object detectors in autonomous vehicles using camera data," *Remote Sensing*, vol. 13, no. 1, p. 89, 2020.

[11] Y. Liu, S. Jin, Q. Shen, L. Chang, S. Fang, Y. Fan, H. Peng, and W. Yu, "A deep learning system to predict the histopathological results from urine cytopathological images," *Frontiers in Oncology*, vol. 12, p. 901586, 2022.

[12] Y. Kawazoe, K. Shimamoto, R. Yamaguchi, Y. Shintani-Domoto, H. Uozaki, M. Fukayama, and K. Ohe, "Faster r-cnn-based glomerular detection in multistained human whole slide images," *Journal of Imaging*, vol. 4, no. 7, p. 91, 2018.

[13] S. Ren, K. He, R. Girshick, and M. Research, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Tech. Rep. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

Appendix A
GitHub Repository

Code is available at https://github.com/LuukNeervens/AIMI_MONKEY.

Appendix B
Pipeline

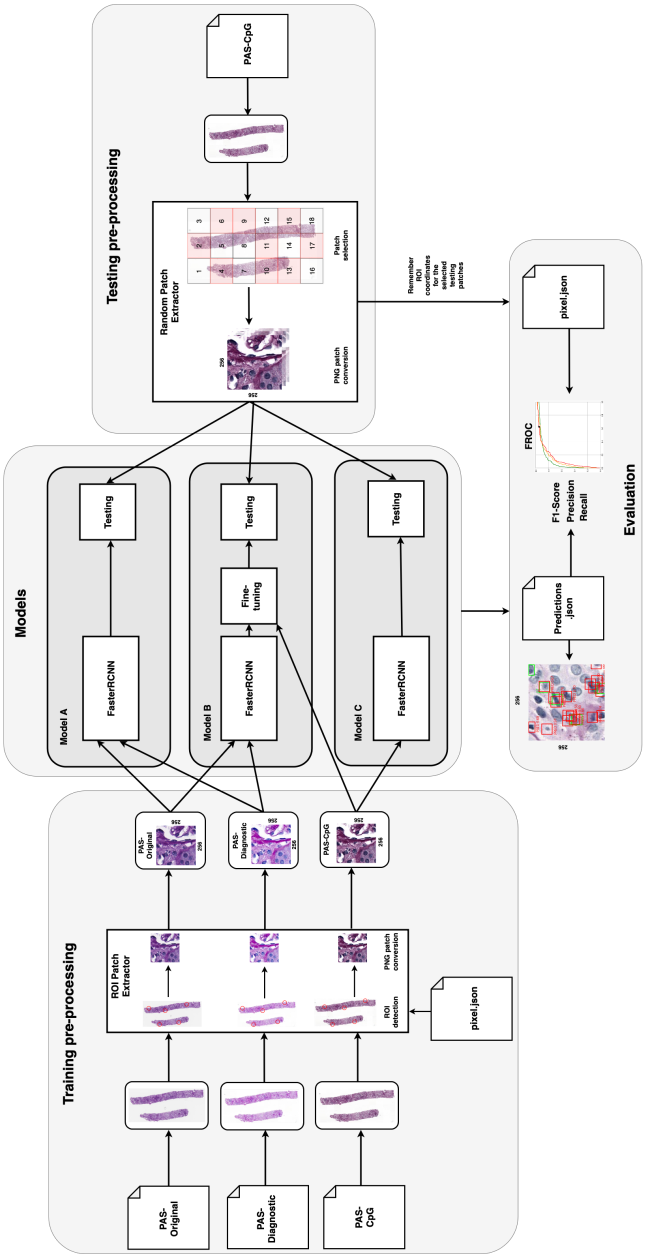
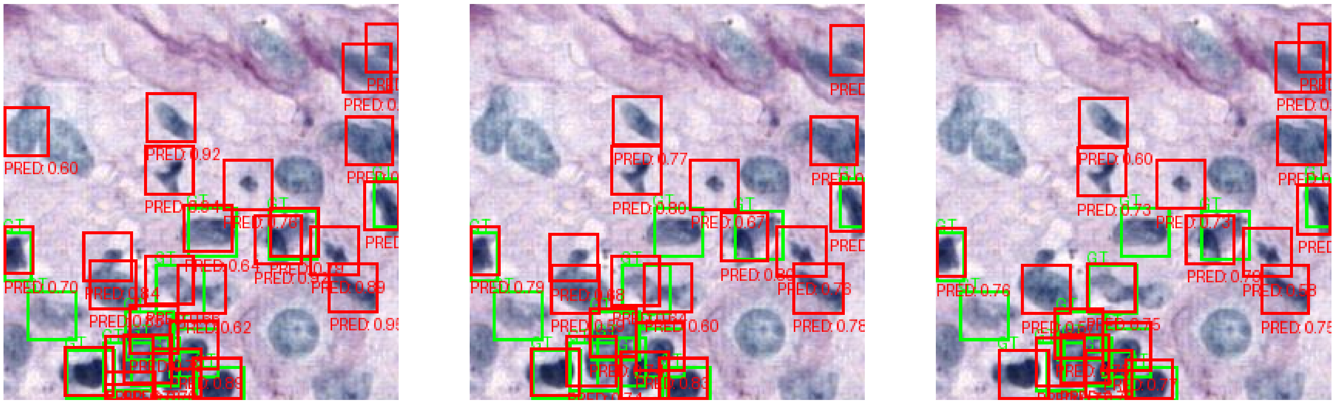


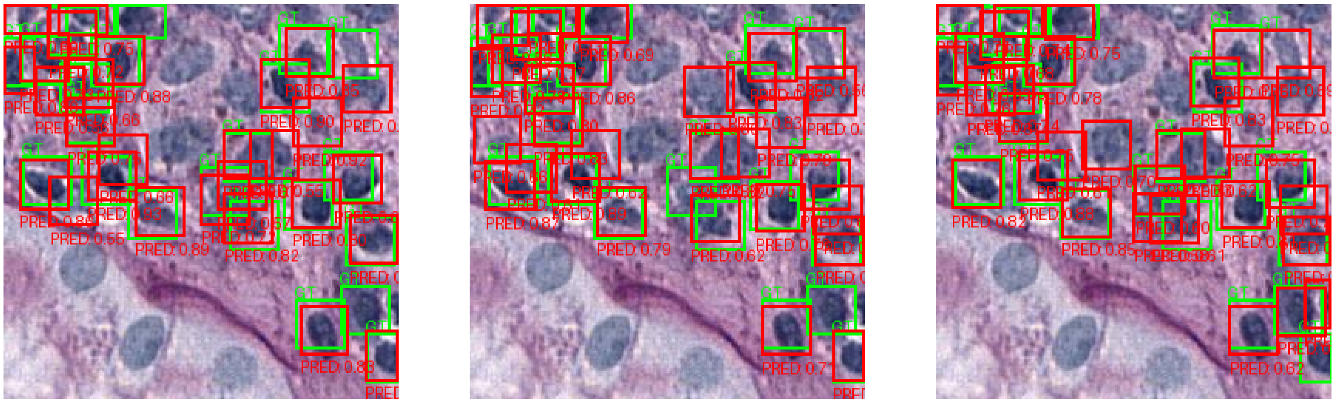
Fig. 8: Overview of the pre-processing, model training, and evaluation pipeline. The pipeline begins with pre-processing for training, where regions of interest (ROI) are extracted as PNG patches from PAS-CpG, PAS-Diagnostic, and PAS-Original whole-slide images (WSIs), along with corresponding ground truth cell coordinates. For testing, patches are randomly generated from PAS-CpG WSIs. Three Faster R-CNN models are evaluated: (A) trained on PAS-Diagnostic and PAS-Original slides and tested directly on PAS-CpG (baseline); (B) trained on PAS-Diagnostic and PAS-Original slides, fine-tuned on PAS-CpG, and tested on PAS-CpG; (C) trained and tested exclusively on PAS-CpG slides. Predicted inflammatory cell locations (in pixel coordinates) from all three models are compared against ground truth annotations to compute performance metrics including precision, recall, F1-score, and FROC curves. In addition, predictions and ground truths are visualized per patch for qualitative comparison.

Appendix C
Prediction comparison across models



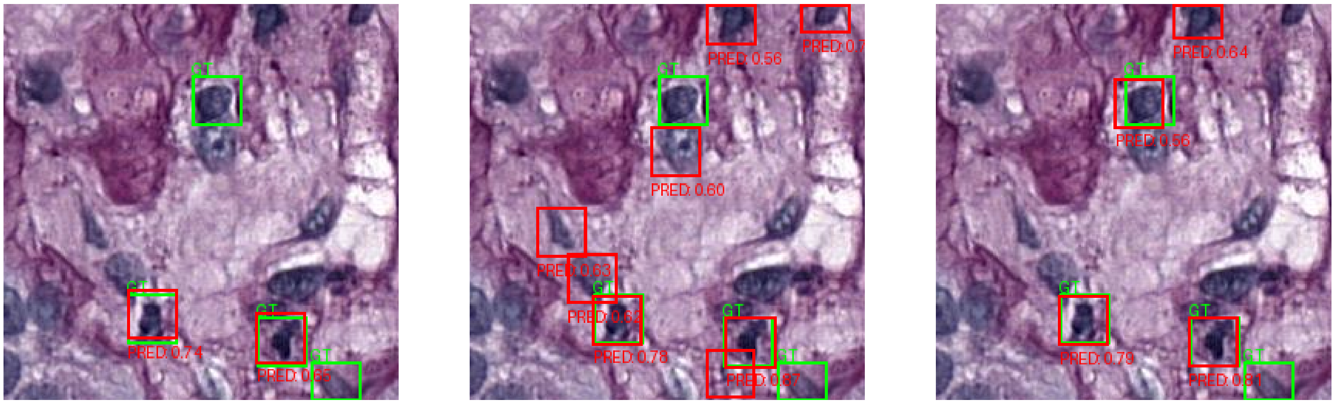
(a) Model A (b) Model B (c) Model C

Fig. 9: Model comparison for patch *A_P000001_PAS_CPG_x10560_y87276*



(a) Model A (b) Model B (c) Model C

Fig. 10: Model comparison for patch *C_P000038_PAS_CPG_x21904_y118224*



(a) Model A (b) Model B (c) Model C

Fig. 11: Model comparison for patch *D_P000015_PAS_CPG_x14134_y6099*