Rethinking Multiple Instance Learning for Corneal Perforation Detection on Radial Anterior Segment Optical Coherence Tomography in Microbial Keratitis

Lucia H Rhode[1] LRHODE1@JH.EDU
[1] *Department of Electrical and Computer Engineering, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland, USA*

Kamini Reddy[2]
[2] *Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA*

Folahan Ibukun[2]

Subeesh Kuyyadiyil[3]
[3] *SNC Chitrakoot, Chitrakoot, Madhya Pradesh, India*

Elesh Jain[3]

Gautam Parmar[3]

Rama Chellappa[1]

Nakul S. Shekhawat[2] NSHEKHA1@JHMI.EDU

Editors: Under Review for MIDL 2026

Abstract

Multiple instance learning (MIL) is a dominant paradigm for weakly supervised pathology detection in medical imaging, yet its assumptions of large, unordered, and heterogeneous bags are violated in several clinical modalities. Radial anterior segment OCT (ASOCT) produces a fixed, small, and anatomically structured bag of six scans at predefined angles, raising the question of whether expressive attention-based pooling provides meaningful benefit under these constraints. We present the first systematic study of MIL pooling for microbial keratitis perforation detection in ASOCT, evaluating mean, max, attention-based MIL (ABMIL), and gated attention pooling across ResNet-50 and ViT-B/16 encoders on an infected-only cohort of 150 eyes (24 perforated) with patient-grouped stratified 5-fold cross-validation. Encoder choice consistently dominated performance, with pretrained ViT feature extractors consistently outperforming convolutional features across all pooling methods. Critically, learned attention pooling yields no meaningful advantage over mean pooling within this small, fixed-bag setting. These findings challenge the prevailing assumption that increasingly expressive MIL pooling is universally beneficial, demonstrating instead that for structured, low-cardinality bags common in certain biomedical imaging modalities, representation quality is the primary driver of performance while complex pooling provides limited gains.

Keywords: Anterior Segment Optical Coherence Tomography, Microbial Keratitis, Corneal Perforation, Corneal Imaging, Multiple Instance Learning, Attention Pooling, Weakly Supervised Learning.

1. Introduction

Microbial keratitis is the leading cause of corneal blindness globally (Flaxman et al., 2017; Whitcher et al., 2001), with the greatest disease burden in low- and middle-income countries (Upadhyay et al., 1991). Among its most devastating complications is corneal perforation, a full-thickness stromal defect caused by inflammatory collagen degradation (Jhanji et al., 2011). ASOCT provides high-resolution cross-sectional visualization of stromal architecture, infiltrate depth, and deeper anatomic damage that may not be apparent on slit-lamp examination alone (Chong et al., 2024; Konstantopoulos et al., 2011; Abdelghany et al., 2024). Prior work has reported near-perfect inter-grader agreement for perforation grading on ASOCT (Ibukun et al., 2026). However, ASOCT interpretation requires expertise that may be unavailable in remote or resource-limited settings, motivating automated detection. ASOCT-based deep learning for infectious keratitis is nascent for pathogen classification (Koyama et al., 2021), but perforation detection remains unexplored.

Each patient-eye-visit in the Heidelberg Anterion *Metrics* protocol yields six radial cross-sectional scans at evenly spaced angles. A perforation may be obvious in only a subset of angles, and clinicians review all six jointly before assigning a label. This naturally maps to multiple instance learning (MIL) (Ilse et al., 2018; Lu et al., 2021), where a bag carries a single label and the model learns to aggregate instance-level evidence. Although MIL has been widely successful in computational pathology, where slide-level labels are aggregated over thousands of patches, its performance in a small, fixed-bag regime (N=6) is much less studied, and it remains unclear whether learned attention meaningfully outperforms simple averaging (Zaheer et al., 2017) at this scale.

Contributions. (i) We formulate corneal perforation detection on radial ASOCT as a small, fixed-bag MIL task (N=6). (ii) We benchmark four mask-aware pooling heads on Convolutional Neural Network (CNN) and Vision Transformer (ViT) encoders under patient-grouped stratified 5-fold cross-validation. (iii) We demonstrate that head choice is secondary to encoder quality. Mean pooling matches or exceeds learned attention on a strong ViT, while pooling effects are volatile on a weaker ResNet-50.

2. Methods

Dataset and Labels. We enrolled consecutive patients with microbiologically confirmed bacterial or fungal keratitis who were treated at SNC Chitrakoot, a tertiary eye care center in Madhya Pradesh, India, between May 2024 and December 2024. ASOCT imaging used the Heidelberg Anterion Metrics App, acquiring six radial cross-sectional scans at $\{0°, 30°, 60°, 90°, 120°, 150°\}$. Two masked corneal specialists independently graded each eye for frank corneal perforation, defined as a full-thickness stromal defect with Descemet membrane discontinuity and/or iris plugging. Descemetoceles without frank perforation were assigned to the negative class (inter-grader Cohen's $\kappa \approx$ `0.98` (Ibukun et al., 2026)). We report on the infected-only cohort (24 perforated, 126 non-perforated eyes), which isolates the clinically harder discrimination. Full-cohort results including contralateral healthy controls appear in the appendix.

MIL Formulation. Each patient-eye-visit is a bag $\mathcal{B}_i = \{x_{i,1}, \ldots, x_{i,N}\}$ of N=6 single-channel ASOCT scans resized to 224×224 with binary label $y_i \in \{0, 1\}$. The bag model is $f(\mathcal{B}) = g\big(\rho\big(\{h_\theta(x_1), \ldots, h_\theta(x_N)\big)$, where h_θ is a shared instance encoder, $\mathbf{m} \in \{0, 1\}^N$ is

a boolean instance mask with m_k=0 denoting dropped scans, ρ is a permutation-invariant mask-aware pooling head, and g is a linear classifier. Masked instances contribute neither to normalization, pooling, nor to gradient. Scan order is randomly permuted during training.

Encoders and Pooling Heads. We compare ResNet-50 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2021), both ImageNet-pretrained, with four mask-aware pooling heads: mean, max, ABMIL, and gated attention.

Cross-validation and Metrics. Patient-grouped stratified five-fold cross-validation; each fold served as the test set once and the validation set once. Primary metrics: AUROC and AUPRC (threshold-free) metrics. Secondary: F1 at a threshold selected by Youden's J on out-of-fold validation predictions. Point estimates are fold-level means and 95% CIs are percentile bootstraps across folds (descriptive fold spread).

3. Results

Table 1 summarizes the infected-only cohort comparison. Every ViT row exceeds every ResNet-50 row in AUROC. Within the ViT block, pooling choice has little effect: all four pooling heads perform similarly, with a narrow AUROC range (0.952–0.962) far narrower than the ViT–ResNet gap.

Table 1: Test metrics (150 eyes, 24 perforated) for four pooling heads on ResNet-50 and ViT-B/16. Both encoders are ImageNet-pretrained, fully trainable.

Backbone	Pooling head	AUROC	AUPRC	F1 score
ViT-B/16	Mean	0.954 (0.919-0.985)	0.791 (0.643-0.940)	0.739 (0.636-0.829)
ViT-B/16	Max	0.962 (0.931-0.983)	0.856 (0.776-0.926)	0.674 (0.562-0.787)
ViT-B/16	ABMIL	0.952 (0.912-0.980)	0.806 (0.643-0.918)	0.689 (0.535-0.817)
ViT-B/16	Gated	0.957 (0.915-0.982)	0.770 (0.575-0.943)	0.758 (0.598-0.875)
ResNet-50	Mean	0.918 (0.902-0.933)	0.632 (0.529-0.734)	0.592 (0.493-0.675)
ResNet-50	Max	0.901 (0.864-0.945)	0.645 (0.475-0.826)	0.506 (0.345-0.667)
ResNet-50	ABMIL	0.945 (0.908-0.983)	0.763 (0.581-0.930)	0.775 (0.673-0.876)
ResNet-50	Gated	0.832 (0.744-0.920)	0.574 (0.486-0.684)	0.505 (0.391-0.619)

4. Discussion and Conclusion

In this small, fixed-bag ASOCT setting (N=6), encoder quality dominates pooling head choice: the ViT–ResNet gap exceeds the within-backbone gap across pooling heads, and mean pooling performs comparably to the more expressive learned pooling heads. This contrasts with the large-bag regime where expressive pooling is typically beneficial.

This study is limited by a modest, class-imbalanced, single-center cohort. Nonetheless, to our knowledge, this is one of the largest ASOCT cohorts for perforation assessment in microbial keratitis, and the dataset is clinically relevant to the intended low-resource setting.

The practical implication is that mean pooling is a strong baseline in small, fixed-bag biomedical MIL, and effort is better spent on encoder quality. A promising next step is stronger ASOCT-specific pretraining, such as masked autoencoder self-supervised learning for ViTs.

Acknowledgments

The authors gratefully acknowledge support from the National Institutes of Health (K23EY032988 and R33EY034343 to N.S.S.), KeraLink International, and the Stephen F. Raab and Mariellen Brickley-Raab Rising Professorship in Ophthalmology. The authors also thank the staff at SNC Chitrakoot for their assistance with patient recruitment and data collection.

References

Ahmed A. Abdelghany, Jorge L. Alio, and Heba Radi AttaAllah. Role of Anterior Segment Optical Coherence Tomography in Staging and Evaluation of Treatment Response in Infectious Keratitis. *Cornea*, 43(10):1216–1222, October 2024. ISSN 1536-4798. doi: 10.1097/ICO.0000000000003466.

Yu Jeat Chong, Matthew Azzopardi, Gulmeena Hussain, Alberto Recchioni, Jaishree Gandhewar, Constantinos Loizou, Ioannis Giachos, Ankur Barua, and Darren S. J. Ting. Clinical Applications of Anterior Segment Optical Coherence Tomography: An Updated Review. *Diagnostics*, 14(2):122, January 2024. ISSN 2075-4418. doi: 10.3390/diagnostics14020122.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Seth R. Flaxman, Rupert R. A. Bourne, Serge Resnikoff, Peter Ackland, Tasanee Braithwaite, Maria V. Cicinelli, Aditi Das, Jost B. Jonas, Jill Keeffe, John H. Kempen, Janet Leasher, Hans Limburg, Kovin Naidoo, Konrad Pesudovs, Alex Silvester, Gretchen A. Stevens, Nina Tahhan, Tien Y. Wong, Hugh R. Taylor, Rupert Bourne, Peter Ackland, Aries Arditi, Yaniv Barkana, Banu Bozkurt, Tasanee Braithwaite, Alain Bron, Donald Budenz, Feng Cai, Robert Casson, Usha Chakravarthy, Jaewan Choi, Maria Vittoria Cicinelli, Nathan Congdon, Reza Dana, Rakhi Dandona, Lalit Dandona, Aditi Das, Iva Dekaris, Monte Del Monte, Jenny Deva, Laura Dreer, Leon Ellwein, Marcela Frazier, Kevin Frick, David Friedman, Joao Furtado, Hua Gao, Gus Gazzard, Ronnie George, Stephen Gichuhi, Victor Gonzalez, Billy Hammond, Mary Elizabeth Hartnett, Minguang He, James Hejtmancik, Flavio Hirai, John Huang, April Ingram, Jonathan Javitt, Jost Jonas, Charlotte Joslin, Jill Keeffe, John Kempen, Moncef Khairallah, Rohit Khanna, Judy Kim, George Lambrou, Van Charles Lansingh, Paolo Lanzetta, Janet Leasher, Jennifer Lim, Hans Limburg, Kaweh Mansouri, Anu Mathew, Alan Morse, Beatriz Munoz, David Musch, Kovin Naidoo, Vinay Nangia, Maria Palaiou, Maurizio Battaglia Parodi, Fernando Yaacov Pena, Konrad Pesudovs, Tunde Peto, Harry Quigley, Murugesan Raju, Pradeep Ramulu, Zane Rankin, Serge Resnikoff, Dana Reza, Alan Robin, Luca Rossetti, Jinan Saaddine, Mya Sandar, Janet Serle, Tueng Shen, Rajesh Shetty, Pamela Sieving, Juan Carlos Silva, Alex Silvester, Rita S. Sitorus, Dwight Stambolian, Gretchen Stevens, Hugh Taylor, Jaime Tejedor, James Tielsch, Miltiadis Tsilimbaris, Jan van

Meurs, Rohit Varma, Gianni Virgili, Ya Xing Wang, Ning-Li Wang, Sheila West, Peter Wiedemann, Tien Wong, Richard Wormald, and Yingfeng Zheng. Global causes of blindness and distance vision impairment 1990–2020: A systematic review and meta-analysis. *The Lancet Global Health*, 5(12):e1221–e1234, December 2017. ISSN 2214-109X. doi: 10.1016/S2214-109X(17)30393-5.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Folahan Ibukun, Kamini Reddy, Subeesh Kuyyadiyil, Elesh Jain, Gautam Parmar, and Nakul S. Shekhawat. Detection of Infectious Corneal Perforation Using Anterior Segment Optical Coherence Tomography. *medRxiv: The Preprint Server for Health Sciences*, page 2026.01.28.26345085, January 2026. doi: 10.64898/2026.01.28.26345085.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2127–2136. PMLR, July 2018.

Vishal Jhanji, Alvin L. Young, Jod S. Mehta, Namrata Sharma, Tushar Agarwal, and Rasik B. Vajpayee. Management of corneal perforation. *Survey of Ophthalmology*, 56(6): 522–538, 2011. ISSN 1879-3304. doi: 10.1016/j.survophthal.2011.06.003.

Aristides Konstantopoulos, Ghasem Yadegarfar, Marina Fievez, David F. Anderson, and Parwez Hossain. In vivo quantification of bacterial keratitis with optical coherence tomography. *Investigative Ophthalmology & Visual Science*, 52(2):1093–1097, February 2011. ISSN 1552-5783. doi: 10.1167/iovs.10-6067.

Ayumi Koyama, Dai Miyazaki, Yuji Nakagawa, Yuji Ayatsuka, Hitomi Miyake, Fumie Ehara, Shin-Ichi Sasaki, Yumiko Shimizu, and Yoshitsugu Inoue. Determination of probability of causative pathogen in infectious keratitis using deep learning algorithm of slit-lamp images. *Scientific Reports*, 11(1):22642, November 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-02138-w.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, June 2021. ISSN 2157-846X. doi: 10.1038/s41551-020-00682-w.

M. P. Upadhyay, P. C. Karmacharya, S. Koirala, N. R. Tuladhar, L. E. Bryan, G. Smolin, and J. P. Whitcher. Epidemiologic characteristics, predisposing factors, and etiologic diagnosis of corneal ulceration in Nepal. *American Journal of Ophthalmology*, 111(1): 92–99, January 1991. ISSN 0002-9394. doi: 10.1016/s0002-9394(14)76903-x.

J. P. Whitcher, M. Srinivasan, and M. P. Upadhyay. Corneal blindness: A global perspective. *Bulletin of the World Health Organization*, 79(3):214–221, 2001. ISSN 0042-9686.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Appendix A. Supplementary Material

A.1. Preprocessing and Augmentation.

Scans are z-score normalized using per-angle (length-N) training means and standard deviations computed on each training fold to avoid cross-fold leakage. Training-time bag-level augmentations are: bilinear resize to 224×224, random horizontal flip (p=0.5), random rotation in $[-10°, 10°]$, brightness/contrast jitter, randomly applied blacking out of the inferior portion of the image (to suppress OCT instrument artifacts), and random *instance* dropout up to two scans masked out via $\mathbf{m}$. Scan order within a bag is randomly permuted during training so the pooling head cannot exploit angular position as a shortcut. Validation and test apply only resize and per-angle normalization.

A.2. Training and Evaluation.

Class-weighted binary cross-entropy with AdamW (Loshchilov and Hutter, 2019), cosine schedule, early stopping on validation AUPRC, and a single seed (42) were used. StratifiedGroupKFold was applied on the bag label on a patient level, cycling each fold through validation and test roles exactly once. AUROC and AUPRC are the primary threshold-free metrics with F1 as a secondary threshold-based metric; unless otherwise stated, CIs are percentile bootstraps over the five fold-level estimates, read as descriptive fold spread.

A.3. Full Sweep.

The complete 2 backbones, 2 inits, 3 freezing regimes, 4 pooling heads = 48 grid, the scan-level non-MIL controls (eye-mean and eye-max aggregation of a supervised single-scan classifier), the full-cohort summary, and the angle-aware pooling stretch extension are retained as supplementary tables. The primary conclusions of the main paper rely only on the infected-only rows summarized in Table 1.

Table 2: Contralateral eyes of enrolled patients without evidence of corneal infection, scarring, or perforation were used as healthy controls. Full-cohort results for the same four paper-claim configurations as Table 1. The manuscript body reports only the infected-only cohort. The full-cohort produces stronger performance, yet conclusions are the same as in the infected-only cohort.

Backbone	Pool	AUROC	AUPRC	F1
ViT-B/16	Mean	0.969 (0.957-0.985)	0.873 (0.808-0.948)	0.689 (0.592-0.786)
ViT-B/16	ABMIL	0.965 (0.950-0.978)	0.856 (0.803-0.906)	0.748 (0.683-0.814)
ViT-B/16	Gated	0.962 (0.932-0.992)	0.835 (0.702-0.962)	0.752 (0.687-0.819)
ResNet-50	Mean	0.924 (0.889-0.960)	0.745 (0.587-0.875)	0.601 (0.500-0.700)

Table 3: MIL-mean and MIL-attention (ViT-B/16, ImageNet, fully trainable) vs. scan-level ViT-B/16 classifiers aggregated by eye-mean and eye-max. MIL rows use fold-level means. Scan rows use pooled out-of-fold metrics. Table shows that bag-level MIL has slightly better performance than just training a scan-level classifier and aggregating its six scan predictions afterward.

Cohort	Backbone	Pool	AUROC	AUPRC	F1
Infected-only	ViT-B/16	Mean	0.954 (0.920-0.985)	0.791 (0.648-0.940)	0.739 (0.636-0.829)
Infected-only	ViT-B/16	ABMIL	0.952 (0.912-0.980)	0.806 (0.648-0.918)	0.689 (0.535-0.817)
Infected-only	ViT-B/16 (scan)	eye-mean	0.943 (0.899-0.975)	0.747 (0.573-0.895)	0.632 (0.491-0.756)
Infected-only	ViT-B/16 (scan)	eye-max	0.939 (0.895-0.973)	0.695 (0.517-0.879)	0.667 (0.511-0.792)
Full cohort	ViT-B/16	Mean	0.969 (0.957-0.985)	0.873 (0.808-0.948)	0.689 (0.592-0.786)
Full cohort	ViT-B/16	ABMIL	0.965 (0.950-0.978)	0.856 (0.803-0.906)	0.748 (0.683-0.814)
Full cohort	ViT-B/16 (scan)	eye-mean	0.957 (0.918-0.985)	0.815 (0.666-0.930)	0.657 (0.513-0.781)
Full cohort	ViT-B/16 (scan)	eye-max	0.946 (0.903-0.978)	0.769 (0.592-0.908)	0.629 (0.492-0.750)

Table 4: Angle-aware pooling (learned and sinusoidal angle encodings) vs. baseline pooling heads on the infected-only cohort.

Encoding	Pool	AUROC	AUPRC	F1
Learned	Mean	0.960 (0.938-0.984)	0.820 (0.673-0.950)	0.599 (0.480-0.691)
Sinusoidal	Mean	0.953 (0.933-0.968)	0.802 (0.700-0.884)	0.708 (0.629-0.793)