

Vision Mamba for Automated Classification of Retinal Diseases from Fundus Images

Taoqing Fu

Department of Computer Science, University of New Hampshire, Durham, NH, USA.
taoqingfu6294@unh.edu

Nicolas L. Chambers

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
nicolas.chambers869@missouri.edu

Pedro Greene

Department of Computer Science, University of Central Florida, Orlando, FL, USA.
pedrogreene03@ucf.edu

Abstract

Automated classification of retinal diseases from fundus images has become a critical public health application of machine learning because it can expand screening capacity for diabetic retinopathy, glaucoma, age-related macular degeneration, and other sight-threatening conditions. Recent advances in selective state space modeling, especially Vision Mamba, offer a promising alternative to convolutional and attention-based architectures by combining linear-time sequence processing with strong long-range context modeling. This paper presents a system-level analysis of Vision Mamba for retinal disease classification. The discussion moves beyond benchmark accuracy to examine structural trade-offs among local feature extraction, global context integration, scanning path sensitivity, memory footprint, and inference throughput. It further analyzes data governance and infrastructure requirements, including dataset diversity, annotation quality, integration with clinical workflow systems, and versioned model management. The paper addresses robustness and fairness challenges such as hidden stratification, underdiagnosis bias, domain shift across fundus cameras, and the need for subgroup performance auditing. Deployment considerations include cloud and edge operation, continuous monitoring, model updating, regulatory compliance, and environmental sustainability. The analysis positions Vision Mamba as a promising component of ophthalmic screening pipelines but emphasizes that its clinical value depends on governance, transparency, fairness, and human oversight rather than architectural innovation alone. Future directions include federated learning, multimodal foundation models, and policy frameworks for equitable and sustainable ophthalmic artificial intelligence.

Keywords

Vision Mamba, state space models, fundus imaging, diabetic retinopathy, medical image classification, health systems, robustness, fairness.

1. Introduction

Recent progress in selective state space modeling has introduced Mamba [1], a linear-time sequence model that addresses some computational limitations of attention-based architectures. Optimized attention kernels such as FlashAttention [2] reduce memory

overhead, but vision transformers [3] and efficient transformer variants [4] still introduce substantial latency and capacity demands. Convolutional networks, including residual architectures [5], remain effective for local pattern extraction but often lack global context. The intersection of these architectural families has become especially relevant for medical imaging because clinical screening requires both diagnostic accuracy and operational tractability. Deep learning systems have shown strong performance in detecting diabetic retinopathy from fundus photographs [6], including validation across multiethnic populations [7], and have been extended to a wider range of retinal conditions [8].

Despite these successes, many screening programs still face infrastructure constraints related to camera variability, image quality, bandwidth, specialist availability, and computational capacity. A clinically viable automated system must therefore be evaluated not only by receiver operating characteristic metrics but also by robustness, fairness, interoperability, governance, and environmental footprint. This paper examines Vision Mamba for automated classification of retinal diseases from fundus images through a systems lens. The goal is not to claim architectural superiority on a single benchmark but to analyze how selective state space models alter the structural trade-offs that govern real-world ophthalmic screening. Subsequent sections discuss related work, architectural considerations, data governance, robustness, deployment, sustainability, and policy implications.

2. Background and Related Work

The application of deep learning to medical image analysis has progressed from task-specific convolutional pipelines to increasingly general representations [9]. More recently, foundation models for generalist medical artificial intelligence have emphasized the value of reusable visual encoders across institutions and tasks [10]. Within this progression, the choice of backbone architecture determines how information flows across local lesions, the optic disc, vessels, and peripheral retina. Convolutional models excel at local texture but can struggle with spatially distributed manifestations that require holistic retinal context. Attention-based models address long-range interactions but may be difficult to deploy at population scale due to quadratic complexity. This gap has motivated visual state space architectures.

Vision Mamba [11] and VMamba [12] adapt selective state space models to two-dimensional images by scanning patches along multiple directions, preserving global receptive fields while reducing complexity. Hybrid designs such as MambaVision [13] combine Mamba layers with transformer blocks to balance throughput and representational capacity. In medical imaging, SegMamba [14] and U-Mamba [15] have shown that long-range sequence modeling can improve segmentation performance in three-dimensional and biomedical contexts. MambaMIL [16] reformulates multiple instance learning for whole slide images using sequence reordering, and VM-UNet [17] brings state space modeling into medical image segmentation. A cross-fusion Mamba approach has also been proposed for multi-view medical image classification [18].

For retinal disease classification specifically, fundus images contain multiple scale-dependent features. Microaneurysms, hemorrhages, exudates, and neovascularization can appear as small punctate findings, while vascular arcades and disc morphology provide global anatomical context. State space models may offer a useful representational trade-off because their linear complexity permits processing high-resolution fundus images without aggressive downsampling. However, the clinical value of such models depends on the surrounding data infrastructure, evaluation protocols, and deployment safeguards. Thus, the present work

extends prior research by analyzing Vision Mamba as an element of a broader ophthalmic screening system rather than as an isolated classifier.

3. System Architecture and Structural Trade-offs

Vision Mamba architectures typically convert a fundus image into a sequence of patches, process the sequence through selective state space blocks, and aggregate the resulting representations for classification. The selective mechanism allows the model to modulate information flow based on input content, which can be useful for irregular lesions. Because the sequence transformation has linear complexity, the architecture can accommodate larger patch grids or higher-resolution inputs than attention-based models with comparable memory budgets. This property is clinically significant because subtle retinal abnormalities may be lost when images are heavily downsampled. Retaining finer spatial detail can improve the detection of microaneurysms, small hemorrhages, and early exudative changes.

However, scanning path is an important structural concern. Unlike convolution, which has translation-equivariant local filters, or global attention, which is permutation-equivariant under positional encodings, Mamba-style processing depends on the order of patches. Different scanning directions can change how contextual information propagates across the retina. Bidirectional and multi-path scanning reduce this sensitivity but increase parameter count and memory traffic. Consequently, system designers must choose between isotropy and efficiency. A model that is highly accurate on a curated dataset may become unstable when fundus orientation, field of view, or cropping behavior changes in routine clinical practice.

Hybrid architectures that interleave Mamba layers with attention or convolutional layers can restore local spatial priors and improve robustness to scanning order. These designs, however, create additional scheduling complexity on edge accelerators. State space layers rely on recurrent scans that may be less friendly to highly parallel chips than convolution or attention, despite their asymptotic complexity. Thus, clinical deployment requires co-design of model architecture and hardware capabilities. Inference throughput, peak memory, and batch scheduling must be analyzed under realistic screening workloads, not only in isolated benchmark environments.

Beyond backbone design, the classification head and disease ontology shape system behavior. Retinal conditions may co-occur, and a single label per image may not reflect diabetic retinopathy with macular edema, glaucoma, and cataract simultaneously. Multi-label, ordinal, or multi-task heads can better align with clinical needs but also increase calibration and monitoring complexity. Vision Mamba can support hierarchical labels, but governance around label definitions is essential. If the ontology is too coarse, the system may fail to triage urgent complications. If it is too fine, annotation reliability may decline and intergrader disagreement may introduce noise into training and evaluation.

4. Data Governance and Infrastructure

Retinal screening datasets are often collected from specific clinical settings and may not represent the distribution of disease severity, camera manufacturer, patient demographics, and imaging protocol encountered in community screening. A Vision Mamba model trained on such data may learn shortcuts such as camera artifacts, disc coloration, or dust spots rather than pathological features. Therefore, data governance must include provenance tracking, image quality criteria, and demographic disaggregation. Annotation should follow standardized grading protocols and include adjudication for ambiguous cases. Without this

infrastructure, even a well-designed state space model can reproduce and amplify existing inequities.

Infrastructure for automated retinal classification must integrate with existing picture archiving and communication systems, electronic health records, and teleophthalmology platforms. DICOM metadata, patient identifiers, and image acquisition parameters must be parsed without breaking privacy safeguards. The model should not operate as a standalone algorithm but as a clinical decision support component within a workflow that includes human verification for positive or uncertain cases. Integration requirements influence architectural choices because the model may need to run on heterogeneous hardware, from hospital servers to portable screening kits.

Dataset versioning is critical because fundus imaging protocols change over time. Hardware upgrades, field-of-view changes, and image compression settings can shift input distributions. A robust governance framework should record model version, training data, exclusion criteria, and performance across subpopulations. Audit trails enable retrospective analysis of systematic errors and support post-market surveillance. Version control also connects model outputs to specific software configurations, which is necessary for patient safety investigations and regulatory inspection.

5. Robustness, Fairness, and Clinical Safety

Evaluation measures such as area under the receiver operating characteristic curve can conceal clinically meaningful failures in subgroups or rare findings [19]. Hidden stratification is especially relevant to retinal screening because vision-threatening conditions may be rare, subtle, or concentrated in patients with darker fundi, advanced media opacity, or atypical anatomy. A Vision Mamba system should be assessed using subgroup analyses for disease severity, image quality, and patient demographics. Reporting only aggregate accuracy can create a false sense of safety and delay the detection of harmful model behavior.

Studies of medical imaging algorithms have shown underdiagnosis bias in underserved populations, where models may systematically miss disease in certain demographic groups because of data imbalance or label noise [20]. For fundus imaging, the risk extends to differences in retinal pigmentation, vascular patterns, camera types, and comorbidity distributions. Fairness cannot be addressed solely by collecting more images; it requires explicit performance audits, participatory data governance, and metric selection that balances overall accuracy with group-specific sensitivity. A state space model is not inherently fair merely because it is computationally efficient.

Responsible machine learning frameworks emphasize that clinical AI should be evaluated according to its impact on health outcomes rather than technical metrics alone [21]. In ophthalmic screening, a false negative may delay treatment for a preventable condition, while excessive false positives may overload referral centers. System-level safety therefore involves calibration, abstention mechanisms for low-quality images, and escalation pathways for uncertain predictions. Vision Mamba can support uncertainty estimation, but this requires deliberate design and validation rather than treating softmax confidence as a calibrated probability.

Translation frameworks further highlight the gap between laboratory validation and real-world clinical delivery [22]. Clinical impact depends on workflow integration, clinician trust, and organizational capacity to respond to model outputs [23]. Ethical oversight must therefore be embedded in the design of Vision Mamba screening systems through institutional

review, prospective validation, and transparent reporting of model limitations. The next section examines deployment considerations that emerge when state space backbones are moved from research environments into routine clinical operations.

6. Deployment and Operational Considerations

Deployment of a Vision Mamba classifier in a screening network requires careful attention to input standardization. Fundus images may arrive with different resolutions, color spaces, field orientations, and compression formats. A state space model trained on normalized images can experience performance degradation when preprocessing differs. Therefore, clinical systems should include a standardized preprocessing layer that performs field-of-view detection, vessel-based orientation, color normalization, and quality assessment before inference. This layer must be versioned and treated as part of the medical device because changes in preprocessing can alter downstream predictions even if model weights remain fixed.

Inference scheduling is another operational concern. Mamba’s recurrent scanning may be executed on central processing units, graphics processing units, or specialized accelerators. Throughput depends on batch size, sequence length, and memory bandwidth. In a mobile screening van with intermittent connectivity, the system may need to process images locally and upload only positive cases. Lightweight Mamba variants with reduced hidden dimensions can improve edge deployment but may sacrifice sensitivity for rare findings. Operational decisions should therefore be grounded in a formal risk assessment that weighs missed disease against referral overflow and device energy consumption.

Cloud deployment offers centralized model updating and monitoring but introduces privacy risks and network dependency. Hybrid architectures in which locally extracted embeddings are transmitted rather than full images can reduce bandwidth and protect patient privacy, yet may complicate interpretability. Clinicians often need to examine the original fundus image alongside model heatmaps to trust and verify predictions. Therefore, a deployment system that supports only abstract embeddings may hinder clinical review. User interfaces should present model outputs alongside the source image and allow clinicians to override decisions, with all interactions logged for audit.

Continuous monitoring is necessary because input distributions drift as cameras age, software updates change color rendering, and patient populations shift. Monitoring should include image quality metrics, prediction distributions, and alert rates. Statistically significant deviations should trigger model recalibration or retraining. However, retraining must follow a controlled change management process that includes locked datasets, predefined acceptance criteria, and regulatory review. State space models may be updated more frequently than attention-based models because their training cost is lower, but more frequent updates can increase governance burden and complicate longitudinal performance comparison.

7. Sustainability and Resource Efficiency

The computational efficiency of Vision Mamba offers potential sustainability benefits for large-scale screening. Attention-based models require substantial memory during training and inference, which translates into higher energy consumption in data centers. Linear-time sequence processing can reduce the carbon footprint of model training, especially when combined with mixed-precision arithmetic and efficient scanning kernels. Nevertheless, the sustainability of a machine learning system must be evaluated over its entire life cycle, including data collection, annotation, storage, training, deployment, and eventual

decommissioning. Efficiency gains in one component may be offset by higher-resolution input requirements or more frequent updates.

The environmental impact of large-scale artificial intelligence has drawn increasing scrutiny. While many studies focus on natural language processing, medical imaging models also consume considerable energy [24]. A Vision Mamba classifier trained on thousands of fundus images may require less energy than a transformer of comparable accuracy, but the difference can be small if training is repeated for hyperparameter search. System architects should therefore adopt reproducibility standards that report training energy, hardware type, and carbon intensity. These disclosures allow health systems to compare models not only on accuracy but also on environmental burden. For resource-constrained regions, lower energy consumption can improve feasibility of local fine-tuning and maintenance.

Sustainability also has social and economic dimensions. Automated screening systems must be maintainable by local teams without dependence on expensive proprietary cloud services. Open-source implementations of Vision Mamba and community-supported evaluation benchmarks can reduce vendor lock-in and encourage technology transfer. However, open code alone does not guarantee operational sustainability. Local institutions need access to computing resources, annotation expertise, and regulatory guidance. Capacity building through partnerships between academic centers and public health agencies can strengthen long-term ownership. Federated learning may further support this goal by enabling collaborative model improvement without centralizing sensitive data [25].

8. Policy and Governance Implications

Policies for ophthalmic artificial intelligence should distinguish between diagnostic assistance, screening triage, and autonomous diagnosis. A Vision Mamba system used for triage may operate with lower physician oversight than one that directly initiates treatment. Regulators should require evidence that the model performs well in the intended deployment population, including subgroup analyses for diabetes status, age, sex, and retinal pigmentation. Post-market surveillance should include mechanisms for reporting missed diagnoses and false positives, with timely investigation by the manufacturer and health authority. Algorithmic audit frameworks can guide this process by specifying how performance metrics are selected, how subgroups are defined, and how errors are adjudicated.

Fairness governance requires more than performance parity. A model that achieves equal sensitivity across groups may still produce unequal clinical benefit if referral pathways or follow-up access differ. Health system leaders should therefore integrate algorithmic fairness audits with process evaluations of care delivery. If a screening program refers more patients from underserved communities, those patients must have access to ophthalmologists, treatment, and transportation support. Otherwise, the algorithm may simply identify disease without improving outcomes. Policy instruments such as reimbursement codes for teleophthalmology can align financial incentives with equitable access.

Transparency around model development and training data is essential for trust and accountability. Developers should disclose the datasets used, inclusion and exclusion criteria, and known limitations. However, full public release of fundus images may conflict with patient privacy, especially when rare conditions make re-identification possible. Data governance models such as structured access, federated evaluation, and synthetic validation can balance transparency with privacy. Regulators should encourage the use of external

validation datasets that are representative of the deployment context, rather than relying solely on internal test splits.

International coordination can help harmonize evaluation protocols and reduce duplication. Different countries have varying regulatory requirements for medical software, which can slow innovation and create barriers for smaller institutions. A common framework for ophthalmic algorithm evaluation, analogous to reference standards in diagnostic imaging, would allow model developers to generate evidence once and apply it across jurisdictions. Vision Mamba and other state space models should be assessed under such frameworks to ensure that architectural novelty does not obscure clinical validity.

9. Conclusion

This paper has examined Vision Mamba as a system-level technology for automated classification of retinal diseases from fundus images. The selective state space backbone offers meaningful advantages in long-range context modeling and computational efficiency, which are relevant to population-scale screening. However, architectural performance alone does not determine clinical value. The system must be embedded within robust data governance, standardized preprocessing, subgroup-aware evaluation, and transparent deployment workflows. Structural trade-offs among scanning path sensitivity, hardware compatibility, and input resolution require careful co-design. Fairness and safety depend on continuous monitoring, clinician override, and equitable care pathways. Policy frameworks should promote algorithmic audits, federated learning, and environmental reporting. Vision Mamba can contribute to accessible and sustainable ophthalmic artificial intelligence if it is developed not merely as an accurate classifier but as a governed, auditable, and human-centered decision support system.

References

1. Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.
2. Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention: Fast and memory-efficient exact attention with IO-awareness. *Advances in Neural Information Processing Systems*, 35, 16344–16359.
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
4. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.
5. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
6. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning

algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402–2410.

7. Ting, D. S. W., Cheung, C. Y.-L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I. Y., Lee, S. Y., Wong, E. Y. M., Sabanayagam, C., Baskaran, M., Ibrahim, F., Tan, N. C., Finkelstein, E. A., Lamoureux, E. L., Wong, I. Y., Bressler, N. M., ... Wong, T. Y. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318(22), 2211–2223.
8. De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., ... Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9), 1342–1350.
9. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88.
10. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956), 259–265.
11. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision Mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.
12. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., & Liu, Y. (2024). VMamba: Visual state space model. *arXiv preprint arXiv:2401.10166*.
13. Hatamizadeh, A., & Kautz, J. (2024). MambaVision: A hybrid Mamba-Transformer vision backbone. *arXiv preprint arXiv:2407.08083*.
14. Xing, Z., Ye, T., Yang, Y., Liu, G., & Zhu, L. (2024). SegMamba: Long-range sequential modeling Mamba for 3D medical image segmentation. *arXiv preprint arXiv:2401.13560*.
15. Ma, J., Li, F., & Wang, B. (2024). U-Mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*.
16. Yang, S., Wang, Y., & Chen, H. (2024). MambaMIL: Enhancing long sequence modeling with sequence reordering in multiple instance learning for whole slide image analysis. *arXiv preprint*.
17. Ruan, J., & Xiang, S. (2024). VM-UNet: Vision Mamba UNet for medical image segmentation. *arXiv preprint arXiv:2402.02491*.
18. Zheng, X., Chen, X., Gong, S., Griffin, X., & Slabaugh, G. (2025, September). Xfmamba: Cross-fusion mamba for multi-view medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 672–682). Cham: Springer Nature Switzerland.
19. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Ré, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 151–159.

20. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12), 2176–2182.
21. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Goldenberg, A., Keim-Malpass, J., ... Jung, K. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337–1340.
22. Liu, X., Glocker, B., McCradden, M. M., Ghassemi, M., Denniston, A. K., & Oakden-Rayner, L. (2022). The medical algorithmic audit. *The Lancet Digital Health*, 4(5), e384–e397.
23. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31–38.
24. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650.
25. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H., Albarqouni, S., ... Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ Digital Medicine*, 3, 119.