

# Domain-Generalized Mamba Networks for Multi-Center Medical Image Classification

Hector Perry

Department of Computer Science, University of Houston, Houston, TX, USA.  
perry616@uh.edu

## Abstract

Multi-center medical image classification is a central task in clinical artificial intelligence, yet its practical adoption remains constrained by severe domain shift across acquisition equipment, reconstruction pipelines, and patient populations. Selective state-space sequence models, particularly Mamba-style architectures, offer linear-complexity long-range modeling and create new opportunities for domain-generalized diagnostic pipelines. This paper presents a system-level analysis of domain-generalized Mamba networks for multi-center medical image classification, treating the model not as an isolated algorithm but as a component within a broader socio-technical infrastructure. The discussion examines architectural trade-offs between convolutional, attention-based, and state-space designs, with emphasis on memory efficiency, input resolution, and domain invariance. It further addresses multi-center data governance, federated readiness, deployment architecture, drift monitoring, and regulatory alignment. The paper argues that the value of domain-generalized state-space models depends on more than benchmark accuracy; it requires deliberate integration with data provenance systems, computational constraints, fairness audits, and clinical safety governance. Robustness under distribution shift, sustainability across hardware generations, and contestability of model decisions are positioned as first-order design requirements. The analysis synthesizes evidence from medical image analysis, distributed learning, fairness research, and clinical model reporting to provide a forward-looking research agenda for decentralized and domain-generalized medical imaging systems.

## Keywords

domain generalization, Mamba networks, multi-center medical imaging, federated learning, clinical AI governance, state-space models.

## 1. Introduction

Multi-center medical image classification is foundational for diagnostic support, but the gap between single-center validation and cross-institutional deployment remains significant. Images acquired at different hospitals vary due to scanner manufacturers, acquisition parameters, reconstruction kernels, and patient demographics. These variations produce domain shifts that degrade classifiers trained under independent and identically distributed assumptions. In recent years, selective state-space sequence models have emerged as a promising family for modeling long sequences with linear complexity. The Mamba architecture [1] builds on structured state-space models [2] and introduces input-dependent transitions that enable selective compression of context. This design avoids the quadratic attention patterns that motivated fast exact attention kernels [3] within transformer models [4]. The resulting efficiency profile is especially relevant for medical images, where high-resolution two-dimensional views, volumetric slabs, and multi-instance patches force difficult trade-offs between contextual coverage and memory consumption. Vision transformers [5]

introduced global receptive fields but impose substantial compute requirements, whereas convolutional residual architectures [6] and U-Net variants [7] have long been favored for biomedical imaging because of translational inductive biases and hardware maturity. Mamba-based visual encoders occupy an intermediate position by maintaining long-range dependencies while retaining favorable scaling in sequence length. Their potential for multi-center medical image classification therefore warrants a system-level examination.

The central argument of this paper is that domain generalization cannot be understood purely as a property of a network architecture. It emerges from the interaction between model inductive biases, data governance mechanisms, computational infrastructure, evaluation protocols, and clinical oversight. In multi-center settings, each hospital contributes a distinct mixture of acquisition hardware, patient demographics, annotation practices, and workflow constraints. A domain-generalized Mamba network must therefore be assessed not only by its average accuracy on retrospective test sets, but also by its behavior under local drift, its calibration across subgroups, and its compatibility with federated or privacy-preserving training. These considerations place the architecture within a broader systems research agenda. The following sections consider architectural trade-offs, governance arrangements, deployment infrastructure, robustness, fairness, sustainability, and regulatory alignment.

## **2. Background and System Motivation**

Medical image analysis has been reshaped by deep learning over the past decade. Surveys and clinical guides [8], [9] document strong performance across detection, segmentation, and classification tasks, but they also emphasize that translation into healthcare requires more than retrospective accuracy. Multi-center data introduce institution-specific artifacts that are not present in many public research datasets, and models frequently fail when evaluated on unseen scanners or patient cohorts. Early work on cross-institution collaboration demonstrated the value of distributed training without centralized data sharing. Federated learning frameworks in digital health [10] and multi-institutional medical imaging [11] addressed privacy and governance constraints while exposing new challenges in statistical heterogeneity and communication overhead. Public challenges such as CAMELYON17 [12] explicitly framed multi-center histopathology classification as a domain generalization problem, revealing that test performance could decline considerably when acquisition sources differed. Broader benchmarks for distribution shift, including WILDS [13], further underscored the need to evaluate models under realistic variations in data collection. These findings motivated a shift from empirical risk minimization toward invariant representation learning and robust optimization. Methods such as invariant risk minimization [14] and systematic domain generalization evaluations [15] clarified that learning stable features across environments is difficult, particularly when domain labels are incomplete or confounded by patient-level variables.

Within this larger effort, state-space models have gained attention as a scalable alternative to attention. Vision-specific adaptations of Mamba use bidirectional scanning to capture spatial structure and have shown competitive results on image classification [16]. They retain linear-time sequence processing while introducing selective gating that modulates information flow according to input content. For medical imaging, this is potentially valuable because a multi-center classifier must process heterogeneous texture, scale, and anatomical context without assuming a fixed receptive field. Nevertheless, architectural advantages alone do not solve domain generalization. The model's inductive biases must be aligned with acquisition physics, annotation processes, and downstream clinical pathways. A system-level perspective

therefore regards the Mamba encoder as one component in a pipeline that includes data harmonization, provenance tracking, model selection, and post-deployment monitoring.

### **3. Architectural Trade-offs in Domain-Generalized Medical Imaging**

Several architectural families coexist in medical image classification. Convolutional networks provide parameter efficiency and local equivariance, but they may struggle with long-range dependencies unless combined with attention or recurrences. Transformer models provide flexible global context but incur high memory costs for large inputs, limiting their use in volumetric or high-resolution multi-view data. State-space models aim to reconcile these constraints. Recent visual Mamba variants adopt bidirectional expansions and cross-fusion mechanisms to aggregate information from multiple views or image regions [17], illustrating that the design space extends beyond simple patch tokenization. However, adopting such architectures for multi-center classification should be guided by the structural requirements of medical imaging rather than benchmark novelty. Self-configuring segmentation pipelines [18] demonstrate the value of automated adaptation to dataset geometry and intensity distributions, a principle that applies equally to classification when scanner protocols vary. In Mamba-based classifiers, the key architectural choices include the number of selective state dimensions, the expansion ratio in the state-space blocks, the scanning strategy for two-dimensional and three-dimensional inputs, and the fusion strategy across views or longitudinal slices. Each choice creates trade-offs between representational capacity, memory use, inference latency, and sensitivity to domain-specific nuisance factors. A deeper selective state model may capture more context but may also overfit to center-specific texture if not paired with domain augmentation or gradient-level interventions. Conversely, a lightweight model may generalize more gracefully but could miss subtle diagnostic features in rare disease presentations.

At the system level, the choice of Mamba network cannot be separated from the data pipeline. Multi-center medical images often arrive with inconsistent voxel spacing, bit depth, and anatomical coverage. Preprocessing decisions such as resampling, windowing, and intensity normalization influence the effective domain shift before the network is trained. A domain-generalized deployment therefore requires architectural evaluation under multiple preprocessing regimes, not only under fixed benchmark settings. Hardware-aware parallelism is another consideration. Selective state-space blocks can be implemented with efficient scan operations, but their performance depends on memory bandwidth and kernel maturity. In contrast, convolutional networks benefit from highly optimized inference libraries on hospital edge devices, while transformers may require specialized attention kernels. A system designer must assess whether the theoretical linear complexity of Mamba translates into practical throughput on the intended clinical infrastructure. This consideration is particularly acute in low-resource settings where graphics processing hardware is heterogeneous or limited. The architecture should also support calibration and uncertainty estimation. Calibration becomes more difficult with complex context aggregation, and poorly calibrated outputs can undermine clinical trust even when discrimination is high. Thus, architectural selection should include evaluation of expected calibration error and decision stability under controlled perturbations.

### **4. Multi-Center Data Governance and Federated Readiness**

Multi-center imaging initiatives are governed by privacy regulations, institutional review requirements, and data use agreements. Centralizing raw images is often impossible or undesirable. Federated learning offers a governance-compatible route because hospitals can retain data and share only model updates. However, federated training of state-space models

introduces synchronization challenges. Selective state parameters, transition functions, and normalization statistics must be aggregated under heterogeneous class distributions. Naive averaging of client updates can drift toward majority institutions and erode performance on underrepresented sites. Domain-generalized objectives can be integrated into the local training loss, but this requires institutions to expose enough metadata for environment grouping without violating patient privacy. Secure aggregation and differential privacy further constrain gradient information, which may affect the stability of selective state-space updates. These system trade-offs remain underexplored in the Mamba literature.

Governance mechanisms should include data provenance records that capture device type, acquisition protocol, and annotation workflow. Provenance enables domain-aware evaluation and helps separate true generalization from accidental overfitting to a small number of source centers. Without such records, multi-center models cannot be meaningfully audited after deployment. Federated readiness also depends on the maturity of local machine learning operations. Institutions vary widely in their ability to run distributed training, manage model versions, and detect data drift. Domain-generalized architectures should therefore be packaged with configuration tools and evaluation harnesses that reduce local integration burden. In addition, consortium governance structures must define how updates are approved, how disagreements about data contribution are resolved, and how participants can withdraw from a federated network without losing access to previously developed models. These institutional arrangements are as important as the choice of state-space block or fusion strategy.

## **5. Infrastructure and Deployment Considerations**

Deployment of domain-generalized Mamba classifiers requires careful mapping from research training to clinical inference. Runtime environments vary across centralized hospital data centers, edge devices in radiology departments, and cloud-based inference services. The model's memory footprint, batch scheduling, and input preprocessing must be reproducible across these settings. Medical images are often large, and sequence-based state-space models may need to process thousands of tokens from a multi-view study. Inference latency can influence clinician workflow if results are delivered interactively. A deployment architecture should therefore include model optimization steps such as quantization, pruning, or distillation, but these interventions must be revalidated because they can interact with selective state dynamics in unexpected ways. Aggressive quantization may disproportionately affect input-dependent gating and degrade performance on rare disease subtypes. Deployment pipelines should include shadow mode evaluation, where model predictions are generated but not shown to clinicians until stability and calibration are verified.

This is especially important in multi-center settings because patient mix and scanner firmware change over time. Continuous monitoring tools should track input distributions, prediction rates, and manually corrected discrepancies. When drift is detected, the system needs a clear pathway for local fine-tuning or federated update without disrupting clinical operations. Domain generalization does not mean a model is immune to operational drift; it means the system has explicit mechanisms for detecting and responding to distribution changes. The infrastructure must also support audit trails that connect predictions to model version, data provenance, and any preprocessing transformations. Such audit trails enable retrospective review when a model error leads to a patient safety event. They also provide the evidentiary basis for regulatory reporting and institutional oversight. In many health systems, these

operational capabilities are still immature, and their absence can limit the safe adoption of otherwise promising machine learning methods.

## **6. Robustness, Fairness, and Clinical Risk**

Robustness and fairness are central to multi-center deployment. A domain-generalized model may achieve high average accuracy while performing poorly on minority subgroups. Surveys on bias and fairness in machine learning [19] document how training data composition, label noise, and proxy variables can produce systematic disparities. In healthcare, biased risk scores have been shown to underestimate illness severity for Black patients when cost proxies are used [20]. These lessons apply directly to medical image classification. If a Mamba-based multi-center model learns features from predominantly urban academic hospitals, it may fail on rural or community sites whose imaging equipment and patient presentations differ. Fairness cannot be assessed solely through aggregate performance. Model reporting frameworks such as model cards [21] encourage explicit disclosure of intended use, training data, evaluation procedures, and known limitations. For medical imaging, such documentation should include scanner manufacturers, institution types, disease prevalence, and subgroup performance.

Public chest radiograph datasets have driven progress in thoracic disease classification [22], [23], [24], but they also illustrate dataset bias and label uncertainty. Models trained on large public corpora may not transfer cleanly to new centers, particularly when labels are derived from radiology reports rather than confirmed diagnoses [22]. CheXpert introduced uncertainty handling for labels [23], and MIMIC-CXR provided linked reports that enable richer text-image research [24]. These resources, while valuable, are not necessarily representative of global patient populations. Robustness evaluation should therefore include stress tests across sex, age, race, and comorbidity strata, as well as hardware and protocol perturbations. In state-space architectures, fairness audits should examine whether selective gating amplifies site-specific artifacts. If the input-dependent state transition learns to rely on high-frequency texture associated with a particular vendor, the model may become overconfident for that vendor while degrading elsewhere. Mitigations include domain adversarial regularization, environment-aware normalization, subgroup-targeted data collection, and post hoc recalibration. However, these interventions are not purely technical; they require governance structures that prioritize equity in evaluation and permit the exclusion of models with unacceptable subgroup error.

## **7. Sustainability and Lifecycle Management**

Sustainability encompasses environmental, operational, and institutional dimensions. Large-scale model training consumes energy, but the iterative nature of multi-center domain generalization can amplify this cost through repeated federated rounds, hyperparameter sweeps, and ablation studies. Mamba architectures may reduce attention-related computation, but any efficiency gain should be measured across the full lifecycle, including data preprocessing, transfer learning, fine-tuning, and monitoring. A model that is cheap at inference but expensive to retrain across multiple federated sites may be less sustainable than a slightly slower model that can be updated locally with modest resources. Lifecycle management should also account for model versioning, data retention, and the eventual decommissioning of outdated classifiers. In multi-center consortia, different institutions may adopt the same model at different times, creating version fragmentation that complicates safety surveillance. A central registry of model versions, evaluation results, and known failure modes can reduce this risk. Such registries are consistent with uncertainty-aware reporting

frameworks and align with the broader movement toward transparent machine learning artifacts [21]. The computational footprint of state-space models warrants further study. Preliminary evidence suggests that selective scan operations can reduce memory peaks relative to attention, but the overall energy cost depends on implementation maturity and hardware utilization. Future work should report not only accuracy but also carbon-equivalent training costs, inference energy per study, and maintenance overhead. These metrics would enable healthcare systems to compare Mamba variants with convolutional and attention-based baselines in terms of total cost of ownership rather than isolated benchmark performance.

## **8. Regulatory Alignment and Clinical Governance**

The deployment of domain-generalized Mamba classifiers must be situated within evolving regulatory frameworks for artificial intelligence in medicine. The United States Food and Drug Administration has proposed a framework for modifications to machine learning-based software as a medical device that anticipates continuous learning and post-market updates [25]. This framework encourages manufacturers to establish predetermined change control plans, which define how algorithms may evolve without triggering a new regulatory submission. For multi-center imaging models, such plans must explicitly address domain adaptation and data drift, because updates intended to improve performance at one hospital may inadvertently degrade performance at another. The European Union Artificial Intelligence Act similarly classifies many medical diagnostic tools as high-risk systems, imposing requirements for risk management, data governance, technical documentation, and human oversight [26]. A domain-generalized Mamba network that aggregates features across views and centers raises particular governance questions regarding explainability. Selective state-space models are not inherently interpretable, and their input-dependent gating can obscure the source of a particular prediction. Explainable artificial intelligence methods have been adapted for medical imaging, but their fidelity in state-space architectures remains an open problem [27]. Regulatory reviewers may require supplementary evidence that a given explanation method faithfully reflects the model's decision process rather than providing a post hoc rationalization. Clinical governance should therefore treat explainability as a system property, encompassing not only saliency maps but also provenance records, calibration reports, and subgroup performance summaries. Local hospital committees need access to these materials to assess whether a model is acceptable for their patient population. Without such infrastructure, regulatory clearance at a national level may not translate into safe local adoption.

## **9. Outlook and Research Agenda**

The intersection of Mamba architectures and multi-center medical imaging remains an open research area with several promising directions. First, dynamic domain adaptation mechanisms could be integrated with selective state-space backbones to adjust scanning windows or state dimensions based on the estimated acquisition domain. Such adaptation would require a clear separation between domain-invariant anatomical features and domain-specific nuisance factors, which remains challenging in practice. Second, foundation model development for medical imaging could benefit from state-space pre-training objectives that emphasize cross-center invariance. Large-scale self-supervised learning on federated cohorts may produce representations that transfer more gracefully across scanners, but this requires coordinated governance and computational resources. Third, cross-modal fusion, which has been advanced by multi-view state-space approaches [17], can be extended to combine images with structured electronic health record data, laboratory results, and longitudinal

history. Such integration demands careful handling of missing modalities and temporal irregularity, areas where selective state-space models may offer advantages due to their sequence processing capabilities. Fourth, open benchmarks for multi-center domain generalization should be expanded to include prospective evaluation phases, hidden external test sets, and standardized fairness metrics. Existing retrospective challenges provide useful guidance but do not capture the operational realities of vendor updates and patient mix changes. Finally, the research community should invest in governance studies that examine how institutions negotiate federated model updates, resolve disagreements about local performance, and maintain trust over time. These socio-technical questions are inseparable from the algorithmic design of domain-generalized networks.

## Conclusion

Domain-generalized Mamba networks present a compelling architectural option for multi-center medical image classification, offering linear-complexity sequence modeling and flexible long-range context aggregation. However, their clinical value depends on more than segmentation or classification accuracy. This paper has argued that system-level considerations, including data governance, federated readiness, deployment infrastructure, robustness, fairness, lifecycle management, and regulatory alignment, are equally decisive. The selective state-space design introduces new trade-offs that must be evaluated in the context of heterogeneous acquisition hardware, institutional workflows, and patient populations. Multi-center domain generalization is not a property that can be achieved by a single architectural innovation; it requires a coordinated effort across data curation, training protocols, evaluation standards, and governance structures. Future work should integrate architectural research with infrastructure development and policy analysis to ensure that advances in state-space modeling translate into equitable and sustainable clinical impact.

## References

1. Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.
2. Gu, A., Goel, K., & Ré, C. (2021). Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396.
3. 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.
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

7. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015* (pp. 234-241). Springer.
8. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60-88.
9. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24-29.
10. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1), 119.
11. Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., & Bakas, S. (2019). Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (pp. 92-104). Springer.
12. Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., ... & Litjens, G. (2018). From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2), 550-560.
13. Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., ... & Liang, P. (2021). WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning* (pp. 5637-5664). PMLR.
14. Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.
15. Gulrajani, I., & Lopez-Paz, D. (2020). In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.
16. 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*.
17. 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.
18. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203-211.
19. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1-35.
20. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.

21. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).
22. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2097-2106).
23. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence, 33(01), 590-597.
24. Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., ... & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1), 317.
25. U.S. Food and Drug Administration. (2021). Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). FDA.
26. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). European Commission.
27. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793-4813.