Lightweight Mamba Networks for Edge-Based Medical Image Classification and Diagnosis

Mateo M. Ortiz
Department of Computer Science, University of North Texas, Denton, TX, USA.
mateoo@unt.edu

Abstract

Medical image classification has advanced rapidly through large-scale deep learning, yet translating these models into point-of-care diagnostics remains difficult because of privacy, latency, connectivity, and computational constraints. Edge-based deployment offers a possible solution by moving inference and selective learning closer to the clinical site, but conventional convolutional and attention-based architectures often carry assumptions that are poorly aligned with constrained hardware, heterogeneous imaging environments, and regulated healthcare workflows. Selective state space models, particularly lightweight Mamba networks, provide a promising alternative because they combine linear-time sequence processing with input-dependent state transitions that can capture long-range dependencies without the quadratic cost of self-attention. This paper examines the system-level implications of adopting lightweight Mamba networks for edge-based medical image classification and diagnosis. It addresses architectural trade-offs in tokenization, state dimensioning, scan direction, and hybrid feature extraction. It further analyzes the integration of such networks into federated learning, differential privacy, robustness engineering, and clinical governance. Rather than focusing narrowly on accuracy, the discussion emphasizes structural trade-offs across energy budgets, hardware heterogeneity, adversarial exposure, fairness, auditability, and lifecycle management. The paper argues that lightweight Mamba architectures should be understood as socio-technical components embedded within distributed diagnostic infrastructures. Their clinical value depends not only on algorithmic performance but also on how effectively they are governed, monitored, and aligned with delivery systems, regulatory expectations, and sustainability goals.

Keywords

edge computing; medical image classification; state space models; Mamba networks; federated learning; clinical governance; diagnostic infrastructure; lightweight architectures.

1. Introduction

Deep learning has substantially reshaped medical image analysis by enabling automated pattern recognition across radiology, pathology, dermatology, and ophthalmology [1]. Large neural networks have achieved clinician-level classification in selected dermatological tasks [2] and have demonstrated strong diagnostic capacity for thoracic pathologies using chest radiographs [3]. The availability of large labeled datasets has further accelerated these developments [4]. Nevertheless, the translation of high-capacity models into routine healthcare delivery remains constrained by data governance, latency, connectivity, and computational limitations at the point of care. This gap has motivated the search for lightweight yet expressive sequence models that can be embedded in bedside devices, handheld ultrasound systems, and remote diagnostic terminals. Selective state space models, exemplified by Mamba, offer a promising direction because they provide linear-time

sequence processing while retaining the ability to model long-range dependencies through input-dependent state transitions [5]. A system-level approach is needed to understand not only the algorithmic accuracy of such models but also their structural trade-offs across edge hardware, clinical workflows, privacy, regulation, and sustainability. This paper examines lightweight Mamba networks for edge-based medical image classification from an interdisciplinary systems perspective.

2. From Deep Learning to State Space Models in Medical Imaging

The dominant medical imaging models initially relied on convolutional architectures in which residual connections stabilized the training of very deep networks [6]. More recently, vision transformers recast image classification as a sequence modeling problem by decomposing images into patches and applying self-attention [7]. Although transformers can capture global contextual relationships, their attention mechanisms impose quadratic complexity with respect to sequence length, which creates significant pressure for edge deployment. Efficient convolutional scaling methods and neural architecture search have yielded compressed models such as EfficientNet [8] and MobileNets [9] that reduce parameter counts and multiply-accumulate operations. Nevertheless, these models often still enforce a trade-off between local feature extraction and long-range reasoning. Selective state space models provide an alternative computational path by scanning image regions as sequences while maintaining hidden states that selectively propagate relevant information. This structural property makes it possible to combine stable training dynamics with subquadratic inference cost, an important requirement for portable diagnostic systems. The shift from convolutional and attentional processing toward state space dynamics is therefore not merely an algorithmic fashion; it represents a rethinking of how medical images can be represented and processed under constrained computational budgets.

3. Edge Constraints and the Case for Selective State Space Models

The deployment of medical image classification models at the edge introduces constraints that differ from centralized server environments. Mobile-class architectures such as MobileNetV2 have addressed memory and latency constraints through inverted residuals and linear bottlenecks [10]. Edge intelligence research emphasizes the need to move inference closer to data sources to reduce communication cost and improve responsiveness [11]. Yet edge nodes in hospitals, ambulances, and rural clinics often exhibit heterogeneous processor capabilities, intermittent connectivity, and strict energy budgets, which demand careful matching between model structure and hardware capabilities [12]. Federated learning has emerged as a complementary organizational pattern that allows multiple clinical sites to collaborate without centralizing raw imaging data [13]. Within this design space, selective state space models indicate potential because their recurrent scanning structure can be executed on low-power accelerators and can adapt to multimodal or multiview inputs. Cross-fusion state space approaches have been proposed for multiview medical image classification to combine complementary views under a unified state representation [14]. This line of work illustrates how state space formulations can support fusion across imaging modalities and clinical views, but fundamental questions remain about reliability, calibration, and governance when such models are placed in real-world diagnostic pipelines.

4. Lightweight Mamba Architectures: Structural Trade-offs

Designing lightweight Mamba networks for medical image classification involves a series of structural trade-offs that cannot be reduced to parameter count alone. The first trade-off

concerns image tokenization. Medical images contain fine anatomical structures, subtle texture changes, and local contrast differences that are clinically meaningful. A Mamba-based architecture may represent an image as a sequence of patches, similar to vision transformer approaches, but the sequence length and patch size determine both the receptive field and the computational burden. Smaller patches preserve fine-grained morphology but produce longer sequences, whereas larger patches reduce computation but risk losing localized signs such as microcalcifications or small lesions. The selective state mechanism can partially mitigate this tension because it dynamically decides which positions contribute to the hidden state, but the architectural choice still influences whether diagnostically relevant regions are retained.

A second trade-off involves the depth and width of the state space layers relative to the feature extractor. Lightweight designs often couple a compact convolutional stem with a state space backbone to leverage the inductive bias of convolutions in early layers while using selective state transitions in deeper layers for global context. This hybrid strategy can reduce the need for many transformer blocks, but it introduces interface mismatches between feature maps and sequence scans. A third trade-off concerns scan direction. Medical images do not possess a natural left-to-right order, and unidirectional scanning can create asymmetrical feature propagation. Bidirectional scanning improves spatial consistency but roughly doubles the state computation and memory footprint. For edge devices with limited cache, this overhead can be significant. A fourth trade-off concerns state expansion. Increasing the state dimension improves representational capacity but also increases the amount of hidden state carried across patches. On embedded accelerators, memory bandwidth and on-chip storage may become the binding constraint before arithmetic throughput. Consequently, system designers must evaluate architectures not only by accuracy on curated benchmarks but also by latency distribution across batches, worst-case memory usage, and sensitivity to quantization. The deployment of Mamba-based diagnostic models therefore requires a co-design process involving algorithm developers, hardware architects, and clinical engineers.

5. Federated and Privacy-Preserving Diagnosis Workflows

Edge-based Mamba networks align naturally with distributed learning paradigms because inference and incremental training can occur near the patient without transferring raw images to central servers. The federated averaging approach originally proposed for decentralized data demonstrates that multiple participants can collaboratively update a shared model by exchanging model parameters rather than local data [15]. In medical imaging, differential privacy provides a formal framework for limiting the disclosure risk associated with model updates or inferences [16]. Deep learning with differential privacy further adapts this framework to gradient-based optimization by clipping per-example gradients and adding calibrated noise [17]. These mechanisms are especially relevant when Mamba networks are deployed across hospitals, because the hidden states and selective update parameters may encode patient-specific patterns.

However, privacy preservation is not free. Noise addition and clipping can reduce diagnostic sensitivity, particularly for rare findings with limited representation. There is also a governance question regarding where the state space model is trained, who controls the aggregation server, and how model updates are audited. A lightweight Mamba network may reduce the communication burden in federated settings because smaller state representations and fewer parameters decrease the size of transmitted updates, but this advantage depends on whether updates occur in the frequency domain, gradient space, or parameter space. System designers should therefore treat privacy as a structural requirement rather than an add-on,

integrating differential privacy budgets, federated topologies, and device capabilities into the same architectural envelope. In addition, clinical institutions must determine how much local personalization is allowable before a model becomes too site-specific and difficult to validate across the broader population.

6. Robustness, Security, and Adversarial Considerations

Medical image classifiers deployed at the edge are exposed to adversarial manipulation, sensor noise, and distribution shifts that are less controlled than in centralized evaluation settings. Adversarial examples have shown that imperceptible input perturbations can cause high-confidence misclassification in deep networks [18]. For diagnostic devices, the consequences of such failures extend beyond performance metrics into patient harm and institutional liability. Lightweight Mamba networks may exhibit different adversarial vulnerabilities than attention-based models because their state transition functions are input-selective and their hidden representations are sequentially compressed. This does not imply inherent robustness; rather, the attack surface shifts. An adversary could target the tokenization stage, the scanning order, or the state update policy, and edge devices may lack the computational reserve for adversarial detection or ensemble defenses.

Security concerns also intersect with privacy protections. The use of differential privacy during training or inference can obscure some gradients but may not protect against model inversion at the aggregation server if the number of model parameters is small enough to permit reconstruction attacks. Robustness therefore requires continuous monitoring, input sanity checks, uncertainty estimation, and secure update mechanisms. In practice, a lightweight Mamba diagnostic system must balance model compactness against the need for redundancy, fail-safe thresholds, and human review for low-confidence outputs. Regulatory frameworks increasingly expect medical artificial intelligence systems to include postmarket surveillance and incident reporting, which means that edge-based Mamba networks should log relevant operational parameters in ways that support audit and forensic analysis without creating excessive storage burdens.

7. Governance, Fairness, and Clinical Accountability

The deployment of edge-based diagnostic models raises fairness and accountability concerns that must be addressed before widespread clinical adoption. Research on healthcare algorithms has demonstrated that predictive models can encode and propagate racial and socioeconomic biases when training data are unrepresentative or when labels reflect unequal access to care [19]. Lightweight Mamba networks are not immune to these dynamics. The sequential state representation may amplify dataset-specific shortcuts if the selective mechanism learns to suppress image regions that are correlated with underrepresentation. Governance frameworks based solely on ethical principles are insufficient because high-level commitments do not automatically translate into operational constraints, audit mechanisms, or accountability for adverse outcomes [20]. High-performance medicine has emphasized the importance of combining machine learning with human judgment, yet the integration must be carefully managed to avoid alert fatigue and automation bias [21]. Fairness in machine learning is a multidimensional concept that includes demographic parity, equal opportunity, and calibration, and these definitions can conflict with one another in clinical settings [22].

For edge-based Mamba systems, accountability is complicated by the large number of distributed devices, the possibility of local model drift, and the opacity of selective state transitions. Clinical institutions need clear protocols for documenting model versions, training

data provenance, performance across subgroups, and the thresholds used for triage or diagnosis. Regulatory bodies may also require evidence that a lightweight model preserves performance across imaging equipment from different vendors, patient populations, and acquisition protocols. In this context, governance is not a bureaucratic afterthought but a necessary design dimension that interacts with architecture selection, deployment topology, and monitoring infrastructure.

8. Infrastructure, Sustainability, and Deployment Policy

The practical sustainability of edge-based Mamba networks depends on how well they integrate into existing health information infrastructure. Edge intelligence frameworks have emphasized that moving computation closer to data sources can reduce network congestion and enable real-time feedback [11]. However, medical facilities often operate with legacy imaging systems, limited IT support, and strict cybersecurity requirements, which means that additional edge nodes must be provisioned, maintained, and updated without disrupting clinical operations [12]. Lightweight state space architectures can contribute to sustainability by lowering energy consumption per inference relative to large transformer models, but the full environmental footprint includes model training, periodic retraining, hardware manufacturing, and cooling. A narrow focus on inference latency may overlook the energy and carbon costs associated with federated training cycles, simulation testing, and continuous monitoring.

Deployment policy should therefore establish minimum hardware specifications, remote update protocols, and decommissioning procedures. It should also clarify how edge devices behave during network outages. In some settings, diagnostic devices must operate autonomously with local fallback models, while in others they may be required to defer to cloud services when connectivity permits. The choice between autonomous operation and gated escalation has important implications for patient safety and legal responsibility. A systems perspective suggests that Mamba-based edge diagnostic tools should be designed with modular components that can be updated independently, inspected for data lineage, and switched to conservative modes when environmental conditions exceed validated operating ranges. Ultimately, the success of these technologies will depend on their ability to operate reliably within the organizational rhythms of hospitals, clinics, and remote care programs rather than on isolated benchmark performance.

9. Conclusion

Lightweight Mamba networks offer a conceptually attractive pathway toward edge-based medical image classification and diagnosis. Their selective state space formulation can align the computational demands of modern diagnostic inference with the practical constraints of distributed clinical infrastructure. However, the transition from algorithmic promise to responsible deployment requires attention to a much broader set of system-level factors. Architectural decisions about tokenization, state dimensioning, scan direction, and hybrid backbones shape not only accuracy but also memory usage, robustness, and compatibility with low-power hardware. Equally important are the governance arrangements that determine how such models are trained, updated, audited, and held accountable across heterogeneous institutional environments. Federated learning and differential privacy provide valuable mechanisms for protecting patient data, but they introduce their own trade-offs in sensitivity, communication cost, and regulatory oversight. Robustness and security cannot be assumed from model compactness, and fairness concerns demand careful consideration of training data provenance, subgroup performance, and clinical workflow integration. Sustainability and

infrastructure policy further influence whether edge-based diagnostic models can be maintained over time without excessive energy consumption or operational disruption. A truly system-oriented approach therefore treats lightweight Mamba networks not as standalone innovations but as components of distributed, regulated, and socially embedded diagnostic ecosystems. Future work should continue to explore cross-fusion strategies, hardware-aware co-design, and lifecycle monitoring while maintaining a clear focus on clinical safety, equity, and accountability.

References

1. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88.

2. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

3. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.

4. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R. L., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 590–597.

5. Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

6. 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.

7. 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. International Conference on Learning Representations.

8. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning, 6105–6114.

9. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

10. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4510–4520.

11. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE, 107(8), 1738–1762.

12. Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. Proceedings of the IEEE, 107(8), 1655–1674.

13. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119.

14. 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.

15. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273–1282.

16. Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Theory of Cryptography Conference, 265–284.

17. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318.

18. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations.

19. 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.

20. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507.

21. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

22. 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.