

Mamba-Based Multi-Modal Fusion for Personalized Diagnosis in Precision Medicine

Songyuan Wang

Department of Computer Science, University of Houston, Houston, TX, USA.
songyuan1975@uh.edu

Damien Wells

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
wells2003@buffalo.edu

Hugo Marshall

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

Abstract

Precision medicine increasingly requires diagnostic systems that can integrate imaging, genomic, laboratory, physiological, and longitudinal clinical data into stable, interpretable, and patient-specific representations. This paper presents a system-level analysis of Mamba-based multi-modal fusion architectures for personalized diagnosis. Unlike conventional attention-centric models, selective state space architectures offer a compelling trade-off between long-range dependency modeling and computational efficiency, but their introduction into clinical infrastructures raises structural questions that extend beyond model accuracy. The paper examines modality alignment, missing data, temporal irregularity, cross-modal representation bottlenecks, and the architectural implications of linear-complexity sequence modeling in hospital settings. It further analyzes governance, deployment, robustness, fairness, and sustainability as first-order design constraints rather than post-hoc considerations. Drawing on cross-domain evidence from medical imaging, oncology, critical care, and federated learning, the discussion reframes personalized diagnosis as an infrastructure problem in which model architecture, data governance, and clinical accountability must co-evolve. The paper argues that Mamba-based fusion can support precision medicine only when embedded within a broader socio-technical framework that addresses data heterogeneity, institutional variation, regulatory oversight, and long-term operational resilience.

Keywords

Mamba; multi-modal fusion; precision medicine; personalized diagnosis; state space models; federated learning; algorithmic fairness; clinical infrastructure.

1. Introduction

Precision medicine has gradually moved away from single-marker reasoning and toward systems-level diagnostic modeling that must coordinate molecular, imaging, physiological, and longitudinal clinical evidence. The computational foundations for many contemporary deep learning systems in medicine rest upon self-attention mechanisms introduced in sequence modeling [1]. Attention-based transformer architectures have enabled flexible

representation learning across language, vision, and structured time series, but their quadratic cost relative to sequence length creates serious barriers in hospital-scale deployments. Recent state space models, particularly the Mamba architecture, reconfigure sequence modeling through selective state transitions that preserve long-range context while operating with linear complexity [2]. A further line of work formalizes structured state space duality and positions these models as efficient alternatives to attention under carefully constrained expressivity [3]. Deep learning in healthcare has been extensively reviewed as a mechanism to derive clinically relevant patterns from raw data [4], and high-performance medicine has been framed as the convergence of human judgment with algorithmic assistance [5]. These developments provide the technical background for a diagnostic infrastructure that must be simultaneously expressive, efficient, and governed by clinical safety constraints.

The central problem in precision medicine is not merely classification accuracy. It is the construction of a coherent patient representation from modalities that differ in scale, noise structure, temporal resolution, and clinical semantics. A genomic variant, a radiological finding, and a laboratory trend do not naturally inhabit the same computational space. Multi-modal fusion architectures therefore perform a critical translational function: they transform heterogeneous observations into a joint representation that can support diagnosis, risk stratification, or treatment selection. Mamba-based models introduce a different set of structural possibilities for this fusion because their linear complexity allows longer sequences and denser temporal aggregation without the memory costs associated with attention. However, this efficiency does not automatically resolve the deeper problems of missing modalities, irregular sampling, institutional coding differences, or clinical interpretability.

From a systems perspective, personalized diagnosis should be understood as an infrastructure problem. The model architecture is one component within a larger arrangement of data pipelines, clinical workflows, regulatory requirements, privacy constraints, and governance mechanisms. Fusing modalities with a Mamba-based backbone may reduce certain computational bottlenecks, but it also creates new requirements for alignment, calibration, uncertainty estimation, and auditability. This paper therefore does not propose a single algorithmic solution. Instead, it examines the structural trade-offs, deployment conditions, and policy implications of Mamba-based multi-modal fusion in precision medicine.

2. Precision Medicine as a Systems Challenge

Precision medicine is often described as the tailoring of diagnosis and treatment to individual patient characteristics, but operationally it is a heterogeneous data integration problem. Publicly available critical care datasets such as MIMIC-III demonstrate the scale and irregularity of multimodal clinical data, including laboratory values, medication records, physiological waveforms, and unstructured notes [6]. Imaging-only models have achieved radiologist-level performance on chest radiograph interpretation [7], but such models typically ignore non-imaging context such as prior admissions, renal function, or medication exposure. Multimodal biomedical AI has therefore been advanced as a framework for combining imaging, text, omics, and structured electronic health record data to produce more clinically comprehensive predictions [8]. In oncology, this integrative logic is especially pronounced because treatment decisions depend on the joint interpretation of histopathology, genomics, and clinical history [9]. These examples illustrate that the value of multi-modal fusion arises not from any single modality but from the consistency and complementarity of the combined signals.

The systems challenge is compounded by the fact that clinical data are not collected according to a uniform experimental design. They are produced as a byproduct of care, often under time pressure, with local coding conventions and variable completeness. A diagnostic model must therefore tolerate absence, misalignment, and distribution shifts across institutions. Data harmonization and ontology alignment become as important as the choice of neural architecture. In this context, a fusion model must be understood as a socio-technical mediator: it standardizes heterogeneous observations into a representation that downstream clinical reasoning can use, but it also inherits the biases and omissions of the underlying records. Any architectural innovation that improves computational efficiency must be evaluated against its ability to preserve clinically meaningful distinctions while remaining robust to operational noise.

Cross-domain comparison with oncology and critical care reveals a recurring tension. In oncology, multi-modal fusion can identify prognostic patterns that are only weakly visible in any single modality, but the resulting models are often highly sensitive to missing genomic assays or staining differences. In critical care, temporal irregularity dominates because laboratory measurements and vital signs are sampled at different frequencies. A Mamba-based fusion architecture may be particularly attractive in these settings because its state space formulation can treat irregular time series as sequences without forcing uniform discretization. Yet this flexibility must be engineered carefully to avoid absorbing clinically irrelevant temporal artifacts.

3. Mamba-Based Fusion Architectures and Structural Trade-offs

A Mamba-based fusion architecture for personalized diagnosis must address three structural questions: how modalities are tokenized, how cross-modal interactions are performed, and how the joint representation is propagated through time. Recent work on cross-fusion Mamba for multi-view medical image classification demonstrates the feasibility of using selective state space layers to combine complementary views without quadratic attention bottlenecks [10]. The approach suggests that Mamba modules can serve as both intra-modality encoders and cross-modality mixers, allowing information to flow between imaging views while maintaining linear scaling with respect to sequence length. For personalized diagnosis, this logic can be extended to imaging, structured clinical variables, and longitudinal events, although the extension introduces challenges that are not present in multi-view imaging alone.

The primary architectural trade-off concerns expressivity. Attention mechanisms can model arbitrary pairwise interactions, which is useful when a rare molecular finding must be linked to a subtle imaging pattern. Selective state space models compress history into a finite state, which improves efficiency but may limit the retention of rare long-range dependencies. The clinical impact of this trade-off depends on the diagnostic task. For tasks dominated by slowly evolving physiological trends, state space compression may be sufficient. For tasks that require abrupt association between distant events, such as an adverse drug reaction following a remote exposure, the compressed state may be insufficient. Hybrid architectures that interleave attention layers with Mamba layers offer one resolution, but they also reintroduce computational cost and complicate deployment.

Another structural consideration is modality alignment. Medical modalities are not naturally synchronized. A genomic test may be performed once, while imaging may be repeated at irregular intervals and laboratory values may arrive continuously. Mamba-based fusion systems need a temporal alignment strategy that does not discard clinically meaningful irregularity. One approach is to construct event sequences ordered by clinical time and rely on

the state space model to implicitly learn temporal decay. Another approach is to create modality-specific branches that are fused at multiple resolutions. The latter is more modular and easier to govern, but it may fail to capture fine-grained cross-modal dependencies. The choice has consequences for interpretability because clinicians generally expect to understand which modality contributed to a particular prediction.

System-level design also depends on the deployment target. In a centralized academic medical center, a large Mamba-based fusion model may be feasible if data are already aggregated. In a distributed network of community hospitals, such centralization is often impossible for legal, privacy, and operational reasons. The architecture must therefore be compatible with federated training or with split learning strategies that keep raw patient data local. Mamba models may offer advantages in this regard because their state space formulation can be updated with smaller communication budgets than full attention matrices, but this remains an active area of systems research rather than a settled result.

4. Data Infrastructure and Federated Governance

Personalized diagnosis depends on high-quality data infrastructure, yet many healthcare institutions operate with fragmented data warehouses, inconsistent identifiers, and limited interoperability. Federated learning has been proposed as a governance mechanism that allows multiple institutions to train models without sharing raw patient data [11]. Medical federated learning systems have demonstrated that multi-institutional collaboration can improve model robustness while preserving local control over sensitive records [12]. Mamba-based fusion models can be incorporated into this federated paradigm, but the heterogeneous modality distributions across institutions create additional synchronization challenges because not all sites collect the same genomic panels or imaging protocols.

Machine learning in medicine has repeatedly shown that deployment failure is often rooted in data pipeline fragility rather than model inadequacy [13]. A fusion model that performs well on curated research datasets may fail when laboratory codes change, when imaging equipment is replaced, or when missingness patterns shift. Governance structures must therefore be established before large-scale deployment. These structures include data use agreements, model versioning, audit logging, and procedures for detecting distributional drift. Federated learning can support governance by reducing the need for centralized data accumulation, but it does not eliminate the need for shared ontologies and common evaluation frameworks. In the absence of shared definitions, federated Mamba-based fusion models may learn institution-specific representations that do not generalize across sites.

From a sustainability perspective, federated governance also changes computational planning. Training a large multimodal model across many hospitals requires coordination of compute, network bandwidth, and energy resources. Mamba architectures may reduce the energy footprint of training compared with attention-based models because of their linear complexity, but the benefit can be offset by the need for longer sequences, larger state dimensions, or more frequent synchronization. A full infrastructure assessment must include not only model accuracy but also the carbon cost of training, the operational burden on local information technology teams, and the long-term maintenance of model artifacts.

5. Fairness, Robustness, and Clinical Safety

Responsible machine learning for healthcare requires attention to harm, equity, and transparency before clinical deployment [14]. Mamba-based multi-modal fusion models are not exempt from these requirements. Bias can enter through underrepresentation in training

data, through modality-specific disparities, or through the fusion process itself. For example, if a model learns to rely on imaging features that are only reliably acquired for patients with better access to advanced diagnostic facilities, it may produce systematically different performance across populations. Fairness in machine learning is a broad and contested area, but the central insight is that algorithmic performance must be evaluated at the level of groups and individuals, not merely through aggregate metrics [15]. Multi-modal models create additional fairness challenges because performance can differ across combinations of available modalities, and a patient with missing genomic data may be assigned a prediction with inappropriate confidence.

Robustness in precision medicine is closely tied to modality integration. Pan-cancer integrative analysis has shown that histology-genomic fusion can improve prognostic performance, but it also reveals that the contribution of each modality varies by cancer type and data quality [16]. Artificial intelligence for multimodal data integration in oncology has similarly emphasized that model behavior must be understood in terms of modality availability and clinical workflow [17]. Fusion of medical imaging and electronic health records has been systematically reviewed, and the results indicate that many reported gains depend on careful preprocessing and feature alignment rather than on any single architectural innovation [18]. These findings imply that Mamba-based systems should be evaluated under realistic missingness and deployment conditions, not only on curated benchmark datasets.

Clinical safety requires uncertainty quantification and abstention mechanisms. A fusion model may produce a plausible diagnosis even when key modalities are absent or conflicting. In such cases, the system should indicate low confidence or defer to human judgment. Deep learning has long emphasized the importance of distributed representations and learned feature hierarchies [19], but clinical translation requires these representations to be linked to calibrated probabilities and actionable explanations. Mamba-based architectures, because they compress temporal history into a latent state, may be more difficult to inspect than attention-based models that explicitly expose pairwise relevance. This does not make them inherently unsafe, but it does require additional interpretability infrastructure, such as perturbation analysis, surrogate models, or state probing methods.

6. Sustainability, Deployment, and Policy Implications

Sustainability in medical artificial intelligence includes environmental, operational, and institutional dimensions. Although convolutional encoders such as U-Net remain widely used for biomedical segmentation [20], newer sequence models must justify their computational cost relative to these established tools. Mamba-based fusion models may reduce memory requirements during training and inference, but their deployment still requires specialized hardware, continuous monitoring, and periodic retraining. A sustainable deployment strategy must consider whether the model can be maintained by the institution over years, whether updates can be performed without disrupting clinical workflows, and whether the computational savings translate into measurable operational benefits.

Policy implications extend beyond model approval. Regulatory frameworks for medical software increasingly demand evidence of generalizability, post-market surveillance, and algorithmic transparency. A Mamba-based diagnostic system that fuses genomics, imaging, and electronic health records must be treated as a medical device in many jurisdictions, and its developers must demonstrate that the intended use is clinically meaningful and that risks are managed. This requires documentation of training data provenance, model limitations, failure modes, and subgroup performance. Because multi-modal fusion models can exhibit

unpredictable interactions between modalities, regulatory review should include stress testing under missing data and domain shift conditions.

The broader policy challenge is to align algorithmic innovation with health system priorities. Precision medicine promises to reduce ineffective treatment and improve diagnostic timeliness, but these benefits will not materialize if models are deployed only in well-resourced academic centers. Mamba-based multi-modal fusion should be evaluated in community settings, rural hospitals, and low-resource environments where data availability and infrastructure differ substantially. Policy incentives may be needed to support shared infrastructure, open benchmarking, and workforce training. In the long term, the goal is not merely to build a more efficient sequence model but to create a diagnostic ecosystem in which technical efficiency, clinical accountability, and equitable access are jointly achieved.

7. Conclusion

Mamba-based multi-modal fusion represents a promising architectural direction for personalized diagnosis in precision medicine. Its linear-complexity sequence modeling can reduce the computational burden of integrating long temporal records, multiple imaging views, genomic features, and structured clinical data. However, the architectural efficiency of Mamba does not automatically resolve the harder problems of clinical data heterogeneity, missing modalities, governance, fairness, and deployment sustainability. These problems are structural and socio-technical rather than purely algorithmic.

A systems-level perspective suggests that future research should focus on the co-design of Mamba architectures, federated governance, and clinical workflows. The value of multi-modal fusion lies not in maximizing benchmark accuracy but in producing robust, interpretable, and equitable diagnostic support under real operational constraints. Precision medicine will benefit from Mamba-based architectures only when they are embedded in infrastructures that respect the complexity of healthcare institutions and the needs of diverse patient populations.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998-6008). Curran Associates.
2. Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
3. Dao, T., & Gu, A. (2024). Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning* (pp. 10041-10071). PMLR.
4. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Culberson, J., Corrado, G., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24-29.
5. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
6. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035.

7. 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.
8. Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. *Nature Medicine*, 28(9), 1773-1784.
9. Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J., & Shah, S. P. (2022). Harnessing multimodal data integration to improve precision oncology. *Nature Reviews Cancer*, 22(2), 114-126.
10. 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.
11. 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.
12. Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10, 12598.
13. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358.
14. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337-1340.
15. 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.
16. Chen, R. J., Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Shaban, M., Shady, M., Williams, M., Joo, B., Noor, Z., Mou, H., Mahmood, F., et al. (2022). Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8), 865-878.
17. Lipkova, J., Chen, R. J., Chen, B., Lu, M. Y., Barbieri, M., Shao, D., Vaidya, A. J., Chen, C., Zhuang, L., Williamson, D. F. K., Shaban, M., Chen, T. Y., & Mahmood, F. (2022). Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*, 40(10), 1095-1110.
18. Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digital Medicine*, 3, 136.
19. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

20. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 234-241). Springer.