Uncertainty-Aware Mamba Networks for Reliable Computer-Aided Medical Diagnosis

Anirudh Menon
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
menon458@unh.edu

Yaoshan Tan
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
yaoshant@binghamton.edu

Abstract

Computer-aided medical diagnosis has made substantial progress through deep learning, yet clinical translation remains constrained by overconfidence, poor calibration, and sensitivity to distributional variation. Transformer architectures improved long-range context modeling but introduced significant computational costs, while state space models such as Mamba offer linear-time sequence modeling with selective state propagation. However, diagnostic reliability requires more than predictive accuracy. This paper presents a system-level examination of uncertainty-aware Mamba networks for medical diagnosis, integrating architectural design, uncertainty quantification, data governance, fairness, deployment infrastructure, and regulatory policy. Rather than focusing narrowly on model structure, the discussion emphasizes the socio-technical conditions under which such models can be safely embedded in clinical workflows. A conceptual diagnostic system is analyzed in which a Mamba backbone is coupled with separate uncertainty estimators for aleatoric and epistemic variability, enabling calibrated rejection, clinician review, and continuous monitoring. The paper further examines trade-offs in centralized and federated deployment, the role of interpretability, and the governance mechanisms required for sustainable adaptation. By situating uncertainty-aware Mamba models within broader institutional and regulatory frameworks, the paper argues that reliability in medical diagnosis emerges from the interaction of architectural capacity, uncertainty calibration, data infrastructure, and human oversight rather than from model architecture alone.

Keywords

uncertainty-aware deep learning; Mamba networks; computer-aided diagnosis; reliability; clinical deployment; data governance.

1. Introduction

The rapid expansion of deep learning in medical image analysis has produced systems with specialist-level performance across dermatology, radiology, ophthalmology, and pathology. Early successes demonstrated that convolutional neural networks could identify skin lesions and detect pneumonia from chest radiographs with high sensitivity, but these systems often lacked calibrated confidence estimates and could silently fail when deployed outside their training distributions. The subsequent introduction of attention mechanisms enabled models to capture long-range dependencies more effectively [1]. Transformer-based architectures became the dominant approach for many image and language tasks, but the quadratic computational burden of self-attention introduced significant barriers for high-resolution

volumetric medical data and real-time clinical decision support. In response, state space models with selective state updates have emerged as an efficient alternative that preserves long-range context while maintaining linear-time sequence modeling properties [2]. Visual adaptations of Mamba have extended this principle to two-dimensional and three-dimensional image understanding [3], and preliminary work in biomedical image segmentation has shown that Mamba-based networks can be applied effectively to medical imaging tasks [4].

Despite these architectural advances, the clinical adoption of diagnostic models remains limited by uncertainty-related failures. A diagnostic model that assigns a high probability to an incorrect class without any indication of doubt is less useful and potentially more dangerous than a model that expresses low confidence and triggers expert review. The distinction between aleatoric uncertainty, which arises from inherent noise in the data, and epistemic uncertainty, which arises from limitations in the model or its training distribution, is central to the design of trustworthy medical systems [5]. Surveys of uncertainty estimation in deep neural networks have catalogued a wide range of Bayesian and frequentist methods, including Monte Carlo dropout, deep ensembles, evidential learning, and conformal prediction [6]. These methods differ in computational cost, calibration quality, and suitability for high-dimensional imaging tasks, and their integration with state space architectures is still at an early stage. A broader review of uncertainty quantification in deep learning has highlighted the need to couple uncertainty estimates with actionable clinical decision rules [7]. The present paper therefore examines how uncertainty-aware Mamba networks can be designed not simply as classifiers, but as components of a reliable diagnostic infrastructure that includes calibration, rejection, governance, and human oversight.

2. Background and Related Work

The history of deep learning in medical diagnosis has been shaped by a series of influential demonstrations. Dermatologist-level classification of skin cancer using deep neural networks showed that models trained on large image datasets could match expert performance in a narrow and well-defined task [8]. A subsequent radiology study reported radiologist-level pneumonia detection on chest radiographs, fueling considerable interest in automated diagnostic tools [9]. In ophthalmology, a clinically applicable deep learning system for retinal disease referral demonstrated the importance of embedding uncertainty and triage pathways into the architecture rather than treating diagnosis as a simple categorical decision [10]. These examples collectively established that machine learning could contribute to high-performance medicine, but they also revealed that clinical integration requires more than a receiver operating characteristic curve [11]. A comprehensive survey of deep learning in medical image analysis described both the promise and the persistent challenges associated with annotation scarcity, data heterogeneity, and interpretability [12].

More recently, the research community has begun to address the distributed and privacy-sensitive nature of medical data through federated learning, which allows multiple institutions to collaboratively train models without sharing raw patient records [13]. Federated approaches are particularly relevant for uncertainty-aware Mamba networks because uncertainty estimates often depend on data diversity, and isolated training environments may fail to represent rare conditions, demographic subgroups, or device-specific imaging artifacts. At the same time, fairness research has demonstrated that diagnostic models can inherit and even amplify social biases embedded in clinical data [14]. Studies of pneumonia detection models illustrated that strong performance in one hospital system does not guarantee generalization to another, largely because of differences in equipment, patient populations, and prevalence rates

[15]. These findings underscore the necessity of uncertainty modeling that is sensitive to distributional shifts rather than assuming that a high test set accuracy will transfer across clinical sites.

Interpretability and explanation have also become central concerns in medical artificial intelligence. Clinicians often reject recommendations that cannot be located within an understandable clinical reasoning process, and explainability frameworks have been proposed to reconcile model outputs with epistemic transparency [16]. A recent cross-fusion Mamba architecture for multi-view medical image classification demonstrated that state space models can be adapted to complex fusion problems involving multiple imaging views [17]. However, such architectures remain primarily accuracy-oriented, and their uncertainty profiles have not yet been systematically integrated into the clinical decision pathway. Parallel work on structured health data has shown that scalable deep learning across electronic health records introduces additional challenges in calibration, missingness, and temporal drift, emphasizing that diagnostic models must be designed for institutional data environments rather than for static benchmark datasets [18]. Together, these lines of research indicate that the next generation of computer-aided diagnosis will require architectures that are efficient, uncertainty-aware, and embedded in robust governance structures.

3. Architectural Integration and Uncertainty-Aware System Design

A reliable computer-aided diagnosis system based on Mamba networks can be conceptualized as a modular pipeline in which representation learning, uncertainty estimation, and decision governance are treated as distinct but interacting layers. The representation layer uses a selective state space backbone to encode long-range dependencies across spatial regions, slices, or views without incurring the quadratic costs associated with dense self-attention. Unlike conventional convolutional backbones that may privilege local texture and require deep stacking for global context, Mamba-style state propagation can model interactions across distant anatomical structures while preserving computational feasibility for three-dimensional volumes. In a diagnostic setting, this efficiency is valuable not only because of hardware constraints but also because clinical workflows often demand rapid triage and repeated inference across large imaging archives.

The uncertainty layer builds on top of the representation backbone and is responsible for producing calibrated estimates of both data-level and model-level variability. Aleatoric uncertainty can be modeled through output distributions that allow the network to express intrinsic ambiguity, such as borderline lesions or artifacts that obscure anatomical boundaries. Epistemic uncertainty can be estimated through ensembles or stochastic state perturbations, where multiple forward passes with different dropout patterns or state-space regularizations yield a distribution over predictions. The separation of these two uncertainty types is not merely a conceptual refinement; it has direct clinical consequences. High aleatoric uncertainty may prompt repeat imaging or additional views, while high epistemic uncertainty may indicate that the case falls outside the model's reliable operating envelope and requires specialist review. The integration of uncertainty estimates with Mamba backbones therefore requires careful architectural design so that state propagation does not inadvertently collapse the very diversity that uncertainty estimation seeks to preserve.

From a system-level perspective, uncertainty should not be treated as a post hoc calibration step applied after training. Instead, it should be part of the training objective, the inference protocol, and the user interface. Training objectives can be designed to penalize overconfident errors while rewarding appropriate abstention, although this must be done without introducing

complex symbolic formulations that obscure clinical interpretation. At inference time, the system can implement rejection rules and triage thresholds that compare predicted uncertainty against predefined risk tolerances. In multi-view diagnosis, the fusion of evidence from different imaging modalities must also account for agreement and disagreement among views. A cross-fusion Mamba approach for multi-view classification illustrates that state space models can effectively combine heterogeneous views [17], but reliability demands that uncertainty be propagated through the fusion process rather than computed only at the final decision layer. If one view is highly ambiguous and another is diagnostically clear, the system should reflect this asymmetry in its confidence estimate and, where appropriate, request additional views or clinician adjudication.

The surrounding decision governance layer converts model outputs into clinical actions. This layer includes thresholding, escalation pathways, audit logging, and periodic drift monitoring. By treating the model as a socio-technical component rather than an autonomous oracle, the system can accommodate institutional variation in risk tolerance, specialty standards, and legal responsibilities. Fairness considerations become particularly important here because uncertainty thresholds may have differential impacts across patient subgroups [14]. For example, a rejection rule that refers uncertain cases to specialists may inadvertently disadvantage rural clinics with limited access to subspecialty care if those clinics rely more heavily on automated screening. Generalization failures observed in pneumonia detection across different hospital systems demonstrate the practical consequences of ignoring such institutional variation [15]. Consequently, the design of uncertainty-aware Mamba networks must include explicit mechanisms for auditing subgroup performance, calibration across demographic categories, and adjustment of decision thresholds to local resource constraints.

4. Data Governance, Fairness, and Infrastructure Deployment

The dependable deployment of uncertainty-aware Mamba networks depends heavily on data governance. Medical imaging datasets are often fragmented across institutions, governed by heterogeneous privacy regulations, and characterized by inconsistent labeling practices. Federated learning has been proposed as a privacy-preserving way to leverage distributed data [13], and it is especially compatible with state space models because the selective state mechanism can be trained with relatively compact parameter updates. Yet federated training introduces its own uncertainties. Heterogeneous hardware, imaging protocols, and patient demographics can produce divergent local model updates, and the global model may fail to capture clinically meaningful subgroup-specific variation. Uncertainty-aware federated systems must therefore aggregate not only model weights but also calibration statistics, enabling each participating site to understand how global uncertainty behaves on its local population. Without such governance, a model may appear well calibrated across aggregate test data while being severely miscalibrated for rare diseases, underrepresented demographic groups, or specific scanner models.

Fairness is equally critical. Bias in medical artificial intelligence can arise from training data that reflect historical disparities in diagnosis, treatment, and access to care. Even when the model architecture is neutral, the choice of labels, exclusion criteria, and outcome measures can encode systemic inequities. Uncertainty estimation intersects with fairness because models often express higher epistemic uncertainty for underrepresented groups, which can lead to more frequent rejection or deferral of their cases. If deferral is framed as a safety mechanism, it may also become a mechanism of exclusion if the deferred patients do not receive timely alternative evaluation. A governance framework should therefore require that

uncertainty-based rejection rates be monitored across demographic groups, clinical sites, and disease subtypes. When significant disparities are observed, the appropriate response may involve targeted data collection, recalibration, or redesign of the triage pathway rather than simply retraining the backbone model.

Infrastructural deployment raises additional trade-offs between centralized and edge-based inference. Centralized deployment simplifies model versioning, auditing, and continuous monitoring but may introduce latency and network dependency that are unacceptable in emergency or point-of-care settings. Edge deployment on hospital servers or specialized accelerators reduces data movement and improves availability, but it complicates the distribution of model updates and the collection of calibration feedback. Mamba-style architectures are attractive for edge deployment because their linear-time state propagation can lower memory and compute requirements compared with large transformer models. However, uncertainty estimation methods such as deep ensembles or repeated stochastic inference multiply computational cost, potentially offsetting the efficiency gains of the backbone. System designers must therefore evaluate whether uncertainty can be approximated through lightweight post hoc calibration, distilled ensembles, or selective state perturbation without undermining reliability. The choice among these approaches is not purely technical; it depends on the clinical risk profile, available infrastructure, and organizational capacity for oversight.

Sustainability also deserves attention. Medical artificial intelligence systems often degrade over time as imaging equipment changes, patient populations shift, and practice patterns evolve. An uncertainty-aware Mamba network can support sustainability by continuously signaling when its confidence distribution diverges from expected operational baselines. This enables monitoring teams to detect drift before performance failures accumulate. The institutional capacity to respond to such signals, however, requires sustained investment in data curation, model auditing, and clinical feedback loops. Without such capacity, uncertainty estimates may simply document decline without enabling corrective action. The integration of uncertainty-aware state space models into health systems should therefore be accompanied by governance arrangements that assign responsibility for monitoring, escalation, and periodic revalidation. The World Health Organization has emphasized that artificial intelligence for health must be evaluated not only for accuracy but also for accountability, equity, transparency, and human oversight [19]. This guidance reinforces the need to treat uncertainty as a governance object rather than a numerical byproduct.

5. Robustness, Policy, and Sociotechnical Reliability

Robustness under distributional shift remains one of the most persistent challenges for medical diagnostic systems. Models trained on curated research datasets often encounter clinical images with different artifacts, contrast protocols, and patient positioning. Uncertainty-aware Mamba networks can mitigate some of these risks by expressing lower confidence when input features diverge from the training distribution, but the models themselves must be evaluated under realistic perturbation and corruption settings. Robustness benchmarks have shown that even high-performing classifiers can degrade substantially under common image corruptions and perturbations [20]. For medical imaging, the relevant corruptions may include motion artifacts, metal implants, low-dose noise, or inconsistent field of view. A reliability-oriented deployment should therefore include site-specific stress testing, in which uncertainty estimates are examined under perturbations that reflect local operational conditions. When uncertainty increases appropriately under corruption, the system can trigger

quality-control protocols or request reacquisition. When uncertainty remains inappropriately low despite corrupted inputs, the system is not robust and should not be trusted.

Policy implications extend beyond individual model evaluation. Regulatory frameworks for medical software increasingly emphasize postmarket surveillance, algorithmic change management, and transparency about intended use. Uncertainty estimates can serve as part of the evidence package for regulatory review, but regulators and clinicians must be able to interpret them without requiring specialized statistical knowledge. This implies that uncertainty outputs need to be translated into clear clinical categories such as routine review, additional imaging, or specialist referral. Such translations should be informed by clinical societies, patient representatives, and institutional safety officers rather than left solely to model developers. The interpretation of uncertainty is not a purely quantitative task; it involves judgments about acceptable risk, resource availability, and the consequences of false reassurance. A governance framework that treats uncertainty thresholds as adjustable institutional parameters can accommodate these differences while preserving a common technical substrate.

The interaction between uncertainty-aware Mamba networks and electronic health records further illustrates the sociotechnical character of reliable diagnosis. Diagnostic models rarely operate in isolation; they consume imaging data alongside laboratory results, clinical notes, and prior history. Scalable learning from electronic health records has shown that multimodal and longitudinal data can improve predictive performance, but also that missingness, coding variability, and temporal irregularity introduce new sources of uncertainty [18]. When a Mamba-based imaging model is combined with health record data, the uncertainty from each source must be propagated and communicated in a coherent manner. Otherwise, the overall system may appear confident because one modality is confident, even when the other modality is highly ambiguous. The design of multimodal diagnostic systems therefore requires explicit protocols for uncertainty aggregation, conflict detection, and clinician override.

Reliability, from this perspective, is an emergent property of a well-governed system rather than a fixed attribute of a trained model. The architecture of the Mamba backbone can improve computational scalability and long-range feature extraction, and uncertainty layers can provide useful signals about ambiguity and epistemic limitation. However, these technical components become clinically meaningful only when they are embedded in workflows that allow for human judgment, institutional audit, and adaptive improvement. The human clinician remains the ultimate decision-maker in most regulatory regimes, and uncertainty outputs are best understood as decision support information that helps clinicians decide when to trust, question, or override a model recommendation. This requires user interfaces that display uncertainty in context, training programs that help clinicians understand model limitations, and legal frameworks that clarify responsibility when automated and human judgments conflict. Without these sociotechnical supports, even the most sophisticated uncertainty-aware Mamba network may fail to improve patient outcomes.

6. Conclusion

Uncertainty-aware Mamba networks represent a promising direction for reliable computer-aided medical diagnosis because they combine efficient long-range sequence modeling with the capacity to express calibrated doubt. This paper has examined the design of such systems at a system level, emphasizing that architectural innovation alone is insufficient for clinical trust. The selective state space backbone provides practical benefits in computational

efficiency and scalability, while uncertainty layers can help distinguish ambiguous cases from model-level ignorance and support better triage decisions. However, these technical advantages must be integrated with data governance, fairness monitoring, deployment infrastructure, robustness testing, and regulatory compliance. Federated learning offers a pathway for privacy-preserving collaboration, but it also demands careful aggregation of calibration signals across heterogeneous sites. Fairness considerations require continuous audit of rejection rates and subgroup performance to ensure that uncertainty-based deferrals do not reproduce existing inequities. Robustness evaluation under realistic clinical perturbations is necessary to prevent silent failures. Ultimately, reliability in medical diagnosis emerges from the interaction between model architecture, uncertainty estimation, human oversight, and institutional accountability. A Mamba-based diagnostic system that expresses uncertainty is not automatically trustworthy. It becomes trustworthy when its uncertainty signals are interpretable, actionable, contextually embedded, and governed by clear policies. Future research should therefore move beyond narrow comparisons of classification accuracy and advance integrated architectures, governance protocols, and empirical evaluations that jointly address the technical and organizational dimensions of diagnostic reliability.

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. Advances in Neural Information Processing Systems, 30, 5998-6008.

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

3. 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.

4. Ma, J., Li, F., & Wang, B. (2024). U-Mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722.

5. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574-5584.

6. Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., & Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. Artificial Intelligence Review, 56(Suppl 1), 1513-1589.

7. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenko, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76, 243-297.

8. 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.

9. 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.

10. 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., Strang, F., Beattie, R. K., Madge, R., Peterson, S., Sadeghi, K., Magni, L., ... & Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine, 24(9), 1342-1350.

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

12. 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., & Sanchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88.

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. 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.

15. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine, 15(11), e1002683.

16. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20, 310.

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. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., ... & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1, 18.

19. World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization.

20. Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261.