Attention-Mamba Fusion for Explainable Cardiovascular Disease Detection from Medical Images

Kabir L. Ganguly
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
kabir696@binghamton.edu

Massimo Rose
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
massimo.rose1978@unr.edu

Damien Wagner
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
contactdamien@ucf.edu

Brooks Horton
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
brooks359@missouri.edu

Abstract

Recent advances in deep learning have enabled high-accuracy medical image classification, but cardiovascular disease detection remains difficult to translate into routine clinical practice because of interpretability gaps, heterogeneous imaging workflows, and governance barriers. This paper provides a system-level examination of attention-Mamba fusion, a hybrid architecture that combines the global relational capacity of attention mechanisms with the linear-complexity sequence modeling properties of selective state-space models. The analysis focuses on structural trade-offs rather than algorithmic detail, considering how tokenization strategies, hybrid encoder stages, cross-view fusion, and explanation modules interact under clinical deployment constraints. We discuss explainability infrastructure, including spatial attribution, concept-based auditing, and uncertainty communication, and examine how these mechanisms can support regulatory review and clinician trust. The paper further evaluates data governance, fairness, robustness, sustainability, and operational resilience as first-order design requirements, not post hoc considerations. Cross-domain comparisons with segmentation systems and foundation models are used to highlight architectural choices that influence interpretability and maintainability. We argue that attention-Mamba fusion offers a viable pathway toward explainable cardiovascular disease detection when embedded within a broader socio-technical infrastructure that includes longitudinal data governance, prospective validation, audit trails, and human oversight. The conclusion identifies future research directions related to multi-modal fusion, federated deployment, model drift monitoring, and policy alignment.

Keywords

attention mechanisms; state-space models; medical image classification; explainability; cardiovascular disease detection; data governance.

1. Introduction

Deep learning has substantially expanded the analytical capacity of medical image interpretation, but its translation into cardiovascular care remains constrained by the need for transparent, resilient, and operationally deployable systems [1]. Earlier work on chest radiograph classification demonstrated that large convolutional models can approach clinician-level performance for selected findings, establishing strong baselines for automated image triage [2]. The availability of large de-identified chest radiograph repositories with structured labels has further accelerated reproducible model development and external evaluation [3]. These advances have created both opportunity and obligation: the opportunity to detect subtle cardiovascular abnormalities earlier, and the obligation to ensure that such systems behave safely across heterogeneous institutions.

The attention mechanism reframed sequence modeling by allowing models to capture long-range dependencies without sequential recurrence [4]. In vision, attention-only architectures such as vision transformers have shown that global self-attention can be competitive with convolutional inductive biases when sufficient data are available [5]. More recently, selective state-space sequence models have been introduced to address the quadratic computational cost of attention while preserving context propagation over long sequences [6]. Subsequent work has formalized structured state-space duality, connecting the efficiency of state-space models to the expressiveness of attention mechanisms [7]. Hierarchical attention architectures have also been developed to balance global context, local feature granularity, and computational cost in visual recognition [8]. These lines of work indicate that no single architectural paradigm dominates every medical imaging scenario; instead, system design requires careful management of trade-offs among representational capacity, computational cost, interpretability, and robustness.

Despite these advances, cardiovascular disease detection introduces unique system-level demands that are not fully resolved by accuracy alone. The clinical task involves heterogeneous image sources, subtle structural and functional abnormalities, variable acquisition protocols, and high-stakes decision thresholds. A promising direction is the fusion of attention-based global context with efficient state-space sequence modeling to create medical image classifiers that are both interpretable and computationally tractable. This paper examines the architectural trade-offs, explainability infrastructure, governance requirements, deployment constraints, robustness expectations, and policy implications of attention-Mamba fusion for cardiovascular disease detection. The contribution is not a single benchmark result but a synthesis of design principles for building explainable and governable medical image classification systems.

2. Background and Clinical Motivation

In cardiovascular imaging, disease evidence may appear as localized structural abnormalities, diffuse textural changes, or temporal motion irregularities. Deep learning systems must therefore reason across scales, from fine anatomical boundaries to global cardiac silhouette and vascular patterns. Post hoc explainability methods such as gradient-weighted class activation mapping have become standard for localizing image regions that influence classifier decisions [9]. Model-agnostic attribution frameworks such as Shapley additive explanations provide a theoretically grounded way to assign importance to input features in

high-dimensional image spaces [10]. Local interpretable model-agnostic explanations offer complementary insights by constructing locally faithful surrogate models around individual predictions [11]. These methods differ in computational cost, stability, and clinical actionability, and their integration into clinical workflows requires careful attention to the types of explanation that cardiologists can actually use during review.

Foundational segmentation architectures have long used encoder-decoder structures with skip connections to preserve spatial detail in biomedical images [12]. In parallel, large chest radiograph datasets with uncertainty-aware labels have shown that label ambiguity must be explicitly modeled to avoid overconfident learning from noisy clinical annotations [13]. Cardiovascular disease labels often reflect heterogeneous diagnostic criteria, follow-up periods, and inter-reader variability. A system designed only to optimize area under the receiver operating characteristic curve may therefore fail in operational settings where uncertainty communication and auditability matter as much as aggregate performance. This clinical motivation supports a shift from monolithic classifiers to modular systems that separate representation learning, fusion, explanation, and decision support.

3. Attention-Mamba Fusion Architecture and System Design

Attention-Mamba fusion systems integrate the global relational reasoning of attention with the scale-efficient sequence modeling of state-space layers. A cross-fusion Mamba approach for multi-view medical image classification has been proposed to handle multiple image views in a cohesive representation space [14]. In such a design, the encoder does not simply process each view independently; it learns to exchange information across views while maintaining a compact state representation that can scale to long sequences. The present discussion abstracts away from any single implementation and instead examines the structural properties that make attention-Mamba fusion suitable for cardiovascular disease detection.

Hybrid architectures that embed transformer blocks within U-shaped medical segmentation frameworks have demonstrated strong performance in multi-scale structural analysis [15]. Similar fusion strategies that combine transformer representations with convolutional feature maps have also been used to recover precise spatial boundaries in medical images [16]. These systems illustrate a broader design principle: architectural components should be selected according to the clinical granularity required at each stage of analysis. In cardiovascular disease detection, early stages may require fine local detail to capture small vascular changes, while later stages may require global relational reasoning to compare different cardiac structures.

In a typical attention-Mamba fusion design, the image is first divided into patches or tokens. The tokenizer determines the granularity of reasoning. Fine tokens preserve subtle vascular changes but increase sequence length and computational load; coarse tokens reduce computation but may lose small abnormalities. Attention layers then model global interactions among tokens, which is useful for relating distant anatomical regions, while selective state-space layers maintain a hidden state that propagates information over the token sequence with lower computational cost. This division of labor allows the system to retain global context without incurring the full quadratic cost of attention over very large images. The fusion can occur at multiple levels: early fusion concatenates raw views, intermediate fusion exchanges features across views, and late fusion combines classification probabilities. Cross-view fusion is particularly relevant for cardiovascular applications where multiple projections or modalities may reveal complementary evidence.

From a systems perspective, the architecture must support reproducible training, versioned checkpoints, and modular replacement of components when new evidence emerges. Attention-based components may require more memory for high-resolution feature maps, while state-space components may reduce memory but require careful initialization and normalization. These trade-offs influence not only accuracy but also maintainability, energyThese trade-offs influence not only accuracy but also maintainability, energy consumption, and the capacity to produce faithful explanations. In high-volume cardiovascular screening, a purely attention-based architecture can become costly because pairwise token interactions grow rapidly with image resolution and sequence length. Selective state-space layers compress historical token information into a fixed-dimensional hidden state, which substantially reduces memory and energy per token while preserving long-range context under controlled assumptions. However, state-space components may be less transparent with respect to spatial attribution than attention maps, since the hidden state is a distributed representation that does not align directly with image pixels. A hybrid system can therefore be designed to apply attention only at selected high-level stages where global clinical reasoning is required, while state-space layers process lower-level token sequences efficiently. This staged design allows the system to expose interpretable attention weights for key decision nodes without imposing excessive computational demands on the entire network.

From a systems perspective, the architecture must support reproducible training, versioned checkpoints, and modular replacement of components when new evidence emerges. Attention-based components may require more memory for high-resolution feature maps, while state-space components may reduce memory but require careful initialization and normalization. Maintenance teams need the ability to isolate a failing module without retraining the entire system, and regulators need evidence that changes to one module do not invalidate earlier validation studies. In cardiovascular imaging, the acquisition pipeline introduces substantial variation, including differences in patient position, radiation dose, contrast timing, and image reconstruction algorithms. The encoder should therefore be paired with robust preprocessing modules that normalize intensity distributions, align anatomical landmarks, and preserve metadata about the imaging protocol. These metadata are essential not only for training but also for auditing and failure analysis.

The cross-view fusion dimension is especially important for cardiovascular disease detection because no single projection fully captures the heart and great vessels. Chest radiographs, computed tomography slices, and echocardiographic views each provide complementary information, but they also introduce alignment challenges. A cross-fusion architecture can exchange feature maps across views at intermediate stages, allowing each view to inform the representation of the others without losing view-specific detail. This fusion strategy reduces the risk that a classifier relies on an incidental feature in one view while ignoring stronger evidence in another. From an explainability standpoint, cross-view fusion creates the need for attribution methods that can indicate which view and which region contributed most to a decision, which is more clinically meaningful than a single heatmap over one image.

4. Explainability Infrastructure and Clinical Interpretability

Explainability in medical image classification should not be treated as a retrospective add-on but as an infrastructure requirement that shapes data collection, model architecture, and deployment interfaces. The attention-Mamba hybrid offers two distinct explanation pathways. Attention maps can reveal pairwise token relationships and highlight image regions that the model compares when making a decision. State-space layers, by contrast, produce hidden

states that are harder to visualize directly, but their sequential nature can be probed by measuring how perturbations in earlier tokens influence later states. Both pathways must be validated for stability, because small changes in intensity or cropping can produce large changes in attribution maps even when the final prediction remains unchanged. Clinicians need explanations that are robust enough to support decision review without creating false confidence in local features that may be artifacts.

Existing post hoc explanation methods provide useful baselines for interpreting medical image classifiers. Gradient-weighted class activation mapping localizes discriminative regions by combining gradients with feature maps, and has been widely used in radiology applications. Shapley additive explanations offer a theoretically principled attribution framework by decomposing a prediction into contributions from individual input features, while local interpretable model-agnostic explanations generate locally faithful surrogate models around a specific prediction. These methods differ in computational cost, sensitivity to model architecture, and ease of clinical interpretation. In attention-Mamba fusion systems, the explanation module should be integrated at multiple abstraction levels, including pixel-level heatmaps, region-level concept scores, and view-level contribution summaries. This layered approach allows a cardiologist to move from a broad question about which view mattered most to a finer question about which anatomical structure contributed to the decision.

A central challenge is that explanations must align with clinical reasoning rather than merely highlighting statistically salient pixels. Concept-based explanations can help by mapping internal activations to clinically meaningful attributes such as cardiomegaly, vascular congestion, pleural effusion, or abnormal cardiac silhouette. These concepts can be derived from structured radiology reports or annotated atlases, and the model can be trained to produce concept scores alongside the final disease prediction. The state-space component can serve as a memory mechanism that aggregates concept-level evidence across the image sequence, while the attention component can explicitly compare concepts that are spatially distant. This design supports a form of concept-based auditing in which a clinician can inspect whether the model activated the correct clinical concepts for a given case. It also facilitates post-market monitoring because deviations in concept distributions can signal distribution shift before aggregate accuracy drops.

Uncertainty communication is a further essential element of explainability in cardiovascular care. Many cardiovascular labels are inherently uncertain because diagnostic criteria vary and follow-up findings may be incomplete. Large chest radiograph datasets with uncertainty-aware labels have demonstrated that treating annotation ambiguity as a first-class signal improves model calibration and prevents overconfident learning from noisy labels. In deployment, the system should report not only a probability estimate but also a confidence interval or an uncertainty category derived from ensembles, test-time augmentation, or Bayesian approximations. Uncertainty estimates are especially important when the model encounters atypical anatomy, poor image quality, or missing views. An explanation that includes uncertainty can reduce the risk of automation bias by explicitly signaling when the model is not confident and when human review is necessary.

The interpretability of medical imaging systems is also shaped by governance artifacts such as model cards and algorithmic audit trails. Model cards provide structured documentation of intended use, performance characteristics, training data composition, and known limitations. When paired with audit logs that record every model version, input preprocessing step, and human override, they create a durable evidence base for regulatory review. For attention-

Mamba fusion systems, model cards should describe the tokenization strategy, the positions of attention and state-space layers, the cross-view fusion scheme, and the explanation outputs. This documentation is not only a transparency mechanism but also an engineering discipline that forces designers to make explicit choices that might otherwise remain implicit in code.

5. Governance, Fairness, and Data Stewardship

The governance of cardiovascular disease detection systems must address fairness not as a statistical afterthought but as a structural property of the data pipeline and model lifecycle. Deep learning models trained on historical medical data can inherit demographic, socioeconomic, and site-specific biases that affect underdiagnosis or overdiagnosis across patient groups. The labels used for cardiovascular disease are often derived from administrative codes, billing records, or retrospective annotations, each of which carries measurement error and selection bias. A hybrid architecture with strong representational capacity can fit these biases more easily if data curation and subgroup auditing are not enforced. Fairness analyses should therefore be conducted across age, sex, race and ethnicity, body habitus, imaging site, and equipment vendor, with pre-specified tolerance bounds for performance disparities.

Data stewardship is equally critical. Cardiovascular imaging datasets are large, sensitive, and often distributed across institutions that operate under different legal frameworks. Federated learning offers one pathway for training robust models without centralizing raw images, but it introduces new governance challenges related to participant contribution, model aggregation, and adversarial leakage. Secure aggregation protocols and differential privacy mechanisms can reduce the risk of re-identification, yet they may interact with the architecture in ways that affect explanations. For instance, privatized state-space updates may obscure the attribution signals that clinicians rely on, requiring explicit compatibility testing between privacy mechanisms and explanation outputs. The governance framework should specify which data elements can be used for training, how long checkpoints are retained, and how patient consent is documented for secondary research use.

Algorithmic auditing in medical AI increasingly requires both internal and external review mechanisms. Internal audits can be integrated into the development pipeline through automated subgroup performance reports, stress tests for domain shift, and red-team evaluations that probe known failure modes. External audits may be conducted by regulatory bodies, health system quality committees, or independent researchers. Attention-Mamba fusion systems should expose standardized interfaces for audit queries, including the ability to retrieve model outputs, confidence scores, and explanation artifacts for any given input. Audit logs must be tamper-evident and associated with model version identifiers. If a model is updated, the audit system should preserve previous versions so that historical predictions remain reproducible and contestable. This is especially important in medical contexts where a delayed diagnosis may become the subject of legal or ethical review years after the original prediction.

Regulatory oversight for medical device software adds another layer of constraints. In many jurisdictions, a machine learning system intended to support diagnosis is classified as software as a medical device and must undergo premarket review, post-market surveillance, and change management. The choice of architecture influences the regulatory pathway because some components may be locked while others are updated through continuous learning. Attention-Mamba fusion systems with modular attention and state-space stages may allow selective locking of the explanation module or the input preprocessing pipeline while

permitting retraining of the encoder on new data under controlled validation. However, regulatory reviewers will require evidence that the locked components do not interact with updated components in ways that degrade safety. This argues for architecture-level change control, with explicit contracts between modules and versioned data schemas.

6. Deployment, Robustness, and Sustainability

Deployment of an explainable cardiovascular disease detection system involves more than model export and containerization. The system must integrate with existing picture archiving and communication systems, radiology information systems, and electronic health record workflows. In practical terms, the model may need to accept DICOM inputs, respect patient identity workflows, and return results in formats that can be embedded into clinician dashboards. The attention-Mamba fusion model should be structured as a service with clear input and output contracts, configurable thresholds for alerts, and support for asynchronous batch processing during off-peak hours. Latency requirements vary by clinical context; emergency settings may demand near-real-time inference, while screening programs can tolerate batch processing. The architecture can be tuned accordingly by adjusting token resolution, number of attention layers, and state-space state size.

Robustness to distribution shift is a persistent challenge for medical image classifiers trained on limited institutions and then deployed broadly. Changes in scanner calibration, reconstruction kernel, patient population, and label definition can all cause silently degraded performance. The system should therefore include out-of-distribution detectors that monitor the input image and intermediate representations against the training distribution. The state-space module can serve as a useful anomaly signal because the hidden state trajectory encodes long-range statistical regularities that may be disrupted by unusual image patterns. Attention maps can further reveal whether the model is attending to clinically implausible regions, which may indicate spurious correlations. Robustness evaluation should include adversarial stress tests, including small perturbations that are imperceptible to clinicians but may alter predictions. Explainability outputs should be evaluated under these perturbations to ensure that attribution maps remain stable and continue to support human review.

Sustainability considerations increasingly shape the design of medical AI systems. Training large attention-based models on high-resolution cardiovascular images can consume substantial computational energy and produce significant carbon emissions. State-space layers offer one route to reduce the per-token energy cost, but the overall efficiency depends on tokenization choices, fusion frequency, and how often attention is applied. Model compression techniques such as pruning, quantization, and knowledge distillation can reduce deployment footprint, but they may alter explanation behavior. Compression should therefore be followed by a re-validation of attribution quality, uncertainty calibration, and subgroup fairness. The sustainability framework should also account for the ongoing energy cost of inference across thousands or millions of patients, not just the one-time training cost. Federated learning can reduce centralized data storage, but it may increase communication and coordination energy across sites.

Continuous monitoring and model updating are essential for long-term safe operation. A deployed model should be monitored for drift in input characteristics, prediction distributions, and clinician override rates. If radiologists begin overriding an increasing proportion of positive predictions, this may indicate that the model is either miscalibrated or poorly aligned with current clinical practice. Monitoring should trigger a structured review that examines explanation artifacts, recent cases, and subgroup performance before any retraining decision

is made. Retraining itself should follow pre-registered protocols that specify data inclusion criteria, validation endpoints, and comparison against the incumbent model. Attention-Mamba fusion can support this lifecycle by maintaining modular checkpoints that allow the state-space encoder to be retrained while attention-based explanation heads remain fixed, or vice versa, subject to regulatory approval. This modular lifecycle management reduces the cost of adaptation while preserving explainability.

7. Policy Implications and Future Directions

The translation of attention-Mamba fusion systems into cardiovascular practice will be shaped by policy choices at the intersection of health care, technology regulation, and professional accountability. Reimbursement policies must recognize that explainability, monitoring, and auditability are not optional overhead but core components of safe deployment. Payment models that reward only raw predictive accuracy may discourage investment in explanation infrastructure and subgroup fairness. Policy incentives should instead encourage the adoption of systems that demonstrate robust performance across heterogeneous sites, transparent documentation, and a clear human review pathway. Liability frameworks also need to clarify the responsibilities of model developers, health care institutions, and clinicians when a model contributes to a missed or delayed diagnosis. Explainability artifacts can support such determinations by reconstructing what the model attended to and how confident it was at the time of the prediction.

Cross-domain comparisons provide useful guidance for future development. In medical segmentation, U-shaped architectures with skip connections have persisted because they preserve spatial detail, while hybrid transformer-convolutional designs have improved contextual understanding. Attention-Mamba fusion can adopt similar U-shaped principles by balancing fine-grained local state propagation with global attention at bottleneck layers. In foundation models for medical imaging, self-supervised pretraining on large unlabeled datasets has demonstrated transferable representations, but foundation models also raise new questions about data provenance, consent, and downstream accountability. Attention-Mamba fusion models intended for cardiovascular disease detection may benefit from pretraining on broad chest imaging corpora, provided that the pretraining data are documented and the downstream fine-tuning process includes subgroup fairness checks.

Future research should explore several directions that extend the current system-level understanding. First, multi-modal fusion combining static images, echocardiographic video, electronic health record data, and laboratory values can provide a more comprehensive view of cardiovascular risk, but it will require architecture-level mechanisms to align asynchronous modalities. Second, concept-level supervision can be strengthened by training models to predict clinically meaningful concepts before final disease predictions, thereby making explanation outputs more aligned with medical reasoning. Third, causal representation learning may help separate anatomical variation from disease-related change, reducing spurious correlations that arise from imaging site or body habitus. Fourth, longitudinal modeling of patient trajectories can enable earlier detection of cardiovascular deterioration, but it demands careful governance of sequential data and privacy-preserving learning. Fifth, federated cross-institutional validation studies are needed to establish whether explainability benefits observed in single sites persist under real-world diversity.

From a policy perspective, the development of explainable cardiovascular AI should be aligned with emerging standards for algorithm auditing, bias evaluation, and post-market surveillance. Transnational regulatory harmonization remains difficult because data protection,

medical device, and professional accountability rules differ across regions. However, common technical documentation standards can lower the cost of compliance while preserving local oversight. Attention-Mamba fusion systems have an opportunity to lead in this area because their modular structure supports documentation at the component level and their dual explanation pathways can be mapped to different stakeholder needs. Regulators may inspect attention maps and uncertainty outputs, cardiologists may inspect concept-based explanations and view contributions, and hospital administrators may inspect subgroup performance and audit logs. This multi-stakeholder alignment is a system-level design objective that should guide both technical and governance decisions.

8. Conclusion

Attention-Mamba fusion represents a promising architectural direction for explainable cardiovascular disease detection from medical images. By combining the global relational capacity of attention mechanisms with the efficient sequential processing of selective state-space models, such systems can balance representational expressiveness, computational cost, and interpretability. This paper has argued that the value of the architecture lies not only in benchmark accuracy but in its capacity to support layered explanations, concept-based auditing, uncertainty communication, and modular lifecycle management. Explainability must be embedded as an infrastructure requirement rather than added after training, and it must be tested under distribution shift, adversarial perturbation, and subgroup analysis.

The broader deployment of these systems depends on governance frameworks that address fairness, data stewardship, regulatory compliance, and continuous monitoring. Attention-Mamba fusion can facilitate these requirements by exposing interpretable components, maintaining versioned checkpoints, and supporting audit queries. At the same time, policy makers and health system leaders must create incentives for transparency, robustness, and equity rather than treating these qualities as subordinate to raw accuracy. Future research should extend the approach to multi-modal and longitudinal data, strengthen concept-level supervision, and validate explainability across federated networks. When developed and governed as a socio-technical infrastructure, attention-Mamba fusion offers a meaningful step toward trustworthy cardiovascular AI that clinicians can understand, audit, and ultimately rely on in patient care.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

2. 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.

3. 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. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2097–2106.

4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & 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., 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.

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

7. Dao, T., & Gu, A. (2024). Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. International Conference on Machine Learning.

8. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012–10022.

9. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626.

10. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

11. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

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

13. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., 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(1), 590–597.

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. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022). Swin-Unet: Unet-like pure transformer for medical image segmentation. European Conference on Computer Vision Workshops, 205–218.

16. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

17. Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F.-M., Tengg-Kobligk, H. von, Summers, R. M., & Wiest, R. (2020). On the interpretability of artificial intelligence in radiology: Challenges and opportunities. Radiology: Artificial Intelligence, 2(3), e190043.

18. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.

19. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229.

20. Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence, 1(1), 20–23.

21. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866–872.

22. Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. JAMA Internal Medicine, 178(11), 1544–1547.

23. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29.

24. 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.

25. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44.