

Multi-Scale State Space Modeling for Microscopic Image Classification in Digital Pathology

Vimal Chakraborty

Department of Computer Science, University of Central Florida, Orlando, FL, USA.
vc23@ucf.edu

Leonard R. Wilson

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

Andreas Thornton

Department of Computer Science, Binghamton University, Binghamton, NY, USA.
andreasthornton856@binghamton.edu

Emmett Cox

Department of Computer Science, University of New Hampshire, Durham, NH, USA.
emmett.work@unh.edu

Abstract

The adoption of deep learning in digital pathology has produced substantial gains in microscopic image classification, yet clinical deployment remains constrained by architectural complexity, computational cost, and weak alignment with the hierarchical tissue representation used by pathologists. This paper presents a systems-oriented analysis of multi-scale state space modeling for microscopic image classification. We argue that state space models offer a structurally favorable balance between long-range context integration and computational efficiency when compared to attention-based vision transformers, especially for gigapixel whole slide images. A multi-scale formulation is developed in conceptual terms, describing how image pyramids, patch encoders, and selective state space layers can be combined to preserve cellular detail while capturing tissue architecture. The paper examines structural trade-offs in sequence length, receptive field, memory, and throughput, and situates these trade-offs within digital pathology infrastructure. Further, we analyze governance, robustness, fairness, and sustainability concerns that arise when such models are integrated into hospital information systems and laboratory workflows. We provide cross-domain comparisons with radiology and genomics, discuss deployment models including edge inference and cloud processing, and outline policy implications for regulatory approval, auditability, and institutional oversight. The paper does not introduce a new algorithm; rather, it develops a systems-level foundation for the design, evaluation, and responsible use of multi-scale state space models in clinical pathology. The analysis indicates that selective state space architectures, when coupled with multi-scale tokenization and clinically grounded validation, may reduce operational barriers while raising new requirements for interpretability, data governance, and lifecycle monitoring.

Keywords

digital pathology; microscopic image classification; state space models; multi-scale representation; clinical artificial intelligence; governance.

1. Introduction

The classification of microscopic images in digital pathology is a large-scale systems problem as much as a modeling problem. Whole slide images routinely exceed gigapixel dimensions and contain heterogeneous tissue regions with diagnostic information distributed across cellular, glandular, stromal, and architectural scales. Deep learning approaches based on convolutional networks have provided strong local feature extraction, while attention-based vision transformers have expanded receptive fields but introduced quadratic memory costs for long sequences. More recently, structured state space models and selective state space mechanisms have been proposed as an alternative for long-range sequence modeling with linear scaling [1][2]. When adapted to visual data, these models can process flattened image patches without the full pairwise attention matrix that dominates transformer computation [3][4]. Residual convolutional backbones remain relevant for local texture and boundary detail [5]. However, the pathology domain imposes distinctive constraints: tissue regions must be analyzed across many optical magnifications, annotations are often weak or slide-level, and computational infrastructure in clinical laboratories is heterogeneous. Therefore, architectural choices cannot be evaluated solely by accuracy on curated benchmarks; they must be assessed through structural trade-offs involving memory, throughput, recurring inference cost, interoperability, and regulatory traceability. This paper examines how multi-scale state space modeling can be positioned within these broader system requirements. We conceptualize a multi-scale state space architecture for microscopic image classification and discuss its design dimensions from a systems engineering perspective.

The field has produced several influential paradigms for computational pathology. Weakly supervised multiple instance learning enables slide-level classification using large numbers of unannotated patches [6], and clinical-grade deployments have demonstrated that carefully validated deep learning pipelines can identify cancer with high sensitivity in real-world workflows [7]. More recently, pathology foundation models have shifted emphasis toward self-supervised pretraining over large institutional slide collections, producing representations that transfer across diagnostic tasks [8]. These developments make the computational properties of the downstream sequence model increasingly important. If a foundation encoder produces millions of patch embeddings for a single slide, the aggregating model must integrate them without excessive memory or latency. State space models offer a promising structural response, but their application to multi-scale pathology data requires careful consideration of patch ordering, scale mixing, local inductive bias, and clinical validation. The remainder of this paper develops a system-level analysis of these considerations.

2. Background and Related Work

Digital pathology image analysis has historically relied on patch-based processing because whole slide images cannot be directly handled by standard convolution or attention models at full resolution. The resulting pipelines often use a tile server to extract smaller image regions, followed by a convolutional feature extractor and a slide-level aggregation method. Weakly supervised learning has become the dominant paradigm because detailed pixel-level annotations are expensive and difficult to obtain in clinical practice [6]. In this setting, the central challenge is to combine thousands or millions of patch-level representations into a coherent slide-level decision. Attention-based aggregation is effective but becomes expensive as the number of patches grows. Structured state space models provide an alternative by maintaining a recurrent state that summarizes information over long input sequences without explicit pairwise interactions [1][2]. When extended to vision, these models can be arranged

in bidirectional or multi-directional scanning patterns to recover spatial context that would otherwise be lost in a simple raster ordering [9][10].

The move toward state space models in digital pathology is not merely an architectural novelty. Clinical translation of computational pathology systems has shown that technical performance is necessary but insufficient without alignment to workflow, quality assurance, and interpretability requirements [11]. Multi-view medical image classification further illustrates that state space modules can be integrated across complementary image representations to improve representation consistency [12]. These developments suggest that state space models may be particularly suitable for pathology because they combine long-range context with moderate computational requirements. However, the precise way in which multi-scale information should enter the state space sequence remains an open systems question. Pathology images contain meaningful structures at low magnification, such as tumor architecture and invasion patterns, as well as at high magnification, such as nuclear atypia and mitotic figures. A well-designed multi-scale model must preserve this interpretive hierarchy rather than collapsing it into a single resolution.

3. Multi-Scale State Space Architecture for Microscopic Image Data

A multi-scale state space architecture for digital pathology can be understood as a pipeline that transforms a whole slide image into several sequences of tokens, processes those sequences through selective state space layers, and then fuses the resulting representations for classification. The first stage is image pyramid construction, in which the slide is represented at multiple magnifications or physical resolutions. Low-resolution views capture tissue architecture, while high-resolution views capture nuclear and cytoplasmic detail. Open-source platforms for digital pathology have established practical standards for working with such pyramids, including tile export and region annotation [13]. In a multi-scale state space design, each magnification level can be tokenized independently using a patch size that is appropriate for the biological structures of interest. These tokens are then arranged into sequences that respect the spatial layout of the tissue, often using a bidirectional scan along rows and columns.

The state space stage processes the resulting sequences using selective state transitions that depend on the input at each position. Unlike a transformer, which compares every pair of tokens, a state space model maintains a compact internal state that is updated as the sequence is traversed. This structural property makes it possible to process very long sequences with memory requirements that scale more gracefully than attention. In a multi-scale setting, it is natural to process each resolution level with a separate state space branch and then introduce cross-scale fusion layers. These fusion layers can be implemented by aligning feature maps at different resolutions and allowing lower-scale contextual features to modulate higher-scale cellular features. Such cross-scale interaction is important because many diagnostic judgments require simultaneous evidence from multiple biological scales. A region that appears architecturally suspicious at low magnification may need to be confirmed by high-magnification nuclear features, while a region with atypical nuclei may need architectural context to determine whether it represents invasive carcinoma or a benign reactive process.

Training a multi-scale state space model for slide-level classification generally requires weak supervision. The model receives a slide-level label and must learn which patches and scales contribute most to the overall decision. This can be achieved by pooling the state space outputs across the sequence and applying a classification objective. The selective nature of the state space layer supports this process because the state update is influenced by the relevance

of each input token. However, care must be taken to avoid overfitting to spurious correlations such as staining intensity, tissue folding, or scanner-specific artifacts. Multi-scale architectures can help by encouraging the model to reconcile evidence across resolutions, but they also introduce additional degrees of freedom that may require stronger regularization and larger validation cohorts. From a systems perspective, the key design trade-off is between scale richness and computational burden. Each additional magnification level increases the number of tokens and the number of state space branches, which raises training cost, inference latency, and the complexity of model debugging.

4. Structural Trade-Offs and Architectural Comparisons

The choice of a state space backbone for microscopic image classification involves several structural trade-offs. The first is the trade-off between sequence length and diagnostic coverage. Large patch sizes reduce the number of tokens and the memory required for processing, but they may obscure small diagnostic structures. Small patch sizes preserve fine detail but increase sequence length and the risk that semantically related regions are separated by long gaps in the sequence. State space models are more tolerant of long sequences than attention-based models, but their recurrent structure can still suffer from information decay if the ordering is poorly chosen. Therefore, sequence ordering is a critical design decision. Bidirectional and multi-scanning strategies can mitigate the loss of context but increase computational complexity and complicate distributed training.

A second trade-off concerns the granularity of the internal state. A larger state dimension gives the model more capacity to store contextual information, but it also increases the cost of each state update and the total memory footprint. This is especially important for inference on whole slide images, where many patches must be processed sequentially. In cloud deployments, large state dimensions may be acceptable if the infrastructure supports high-throughput accelerators. In edge deployments, where the model may run on a workstation inside a pathology laboratory, the state size must be constrained to meet latency and memory limits. Thus, the architecture cannot be designed independently of the expected deployment environment.

A third trade-off involves the balance between local convolutional processing and global state space processing. Convolutional encoders are highly efficient for local texture extraction and can be pretrained on natural images or histopathology data. State space layers are better suited to integrating information across long spatial ranges. A hybrid design may use a convolutional stem to produce compact patch features and then apply state space layers to aggregate those features across the slide. This division of labor is structurally appealing because it separates local perception from global reasoning. However, it creates an interface challenge: the patch features must be sufficiently informative for the state space layer, and the state space layer must be able to recognize which local features are diagnostically relevant across different tissue regions. Multi-scale designs intensify this challenge because the interface must also handle multiple feature granularities.

Compared with hierarchical vision transformers, which build multi-scale representations through shifted window attention, state space models offer the advantage of linear sequence scaling but may require more careful inductive bias for spatial structure [21]. Hierarchical attention can capture local dependencies within windows and global dependencies across windows, but it still relies on attention matrices that grow with window size. State space models avoid explicit pairwise attention, yet their recurrent formulation can be less transparent to interpret. This structural opacity is a concern in clinical contexts where model

decisions must be communicated to pathologists and regulatory bodies. Therefore, the architectural comparison cannot be reduced to accuracy or throughput alone; it must include auditability, failure mode clarity, and the ability to inspect which regions contributed to a decision.

5. Infrastructure, Data Governance, and Clinical Workflow Integration

Deploying multi-scale state space models in digital pathology requires infrastructure that extends beyond model training. The system must interface with whole slide scanners, image management systems, laboratory information systems, and electronic health records. Tissue slides are digitized at multiple magnifications and stored in formats that vary by vendor. The model must therefore be embedded in a processing pipeline that performs slide ingestion, quality assessment, tissue detection, patch extraction, inference, and report generation. The growing literature on artificial intelligence in histopathology emphasizes that these tools can improve cancer research and clinical oncology, but their value depends on robust integration into existing laboratory workflows [14]. Digital pathology artificial intelligence systems are increasingly positioned for diagnosis and precision oncology, yet their operational success requires careful attention to data format standardization and interoperability [15].

Validation is another critical infrastructure concern. Landmark multi-institutional studies have evaluated deep learning systems for detecting lymph node metastases in breast cancer and have shown that algorithmic performance can be strong under controlled protocols [16]. However, clinical deployment introduces variability in slide preparation, staining, scanning equipment, and patient populations. A multi-scale state space model must be validated across these sources of variation. This requires data governance structures that specify how slides are collected, deidentified, stored, and used for model development. Without such structures, it is difficult to know whether a model will generalize from one institution to another. The use of interpretability methods such as gradient-based saliency maps can provide some insight into model decisions [17]. In pathology, saliency maps can highlight regions of the slide that influenced the classification, allowing pathologists to verify that the model is attending to biologically plausible tissue structures. However, saliency methods are not sufficient for clinical assurance on their own.

Fairness and bias introduce further governance challenges. Algorithmic bias has been documented in clinical machine learning contexts where models trained on skewed data may produce systematically different performance across demographic groups [18]. Surveys of fairness in machine learning emphasize that bias can arise at many stages, including data collection, label generation, and model training [19]. In digital pathology, bias may be related to population differences in disease presentation, differences in tissue processing across institutions, or underrepresentation of certain patient groups in training cohorts. Multi-scale models may be especially sensitive to such biases because they can exploit features at any resolution, including features that are not diagnostically meaningful but correlate with site or population. For example, a model might learn to rely on stromal architecture or staining intensity that differs across laboratories. Governance frameworks must therefore require subgroup evaluation, site-stratified analysis, and continuous monitoring after deployment. Health artificial intelligence systems need prospective validation and oversight rather than one-time retrospective performance claims [20].

6. Deployment and Sustainability

The deployment of multi-scale state space models can follow several patterns depending on institutional resources and data privacy constraints. In cloud-based deployment, slides are transmitted to a remote server where inference is performed on high-throughput accelerators. This model benefits from centralized infrastructure management and can support large state space architectures. However, it raises data governance concerns because whole slide images contain sensitive patient information. Network bandwidth is also a constraint because gigapixel images are large and may not be suitable for real-time transmission in all settings. In edge deployment, the model runs on local workstations or dedicated hardware within the pathology laboratory. This reduces data movement and supports privacy, but it requires careful compression or distillation of the model to meet local memory and compute limits. A multi-scale state space architecture can be designed with edge deployment in mind by limiting the number of resolution branches and using compact patch encoders.

Sustainability is an increasingly important consideration for large-scale artificial intelligence systems. The computational cost of training pathology models over large slide collections can be substantial, and recurrent inference across millions of patches can consume significant energy in routine clinical use. State space models may reduce energy costs relative to attention-based models because they avoid the quadratic computation associated with long sequences. However, multi-scale processing can offset these gains if each magnification level is processed independently. Systems designers should consider strategies such as shared patch encoders across scales, mixed-resolution inference, and caching of static slide features. Local segmentation models based on convolutional encoder-decoder architectures remain useful for quality control tasks such as tissue detection and artifact rejection [22]. These lightweight components can prevent the full state space model from processing irrelevant regions, thereby reducing computational waste.

Sustainability also includes institutional capacity. Many pathology laboratories lack dedicated machine learning engineering staff, and the long-term maintenance of an artificial intelligence system requires software updates, model monitoring, and periodic retraining. A multi-scale state space model that is highly customized to a single scanner or staining protocol may become difficult to maintain as laboratory equipment changes. From a systems engineering perspective, it is preferable to define stable interfaces between slide ingestion, patch extraction, state space inference, and reporting. This modularity allows individual components to be updated without replacing the entire system. It also supports regulatory documentation because each component can be evaluated separately and tracked through version control. Cross-domain comparisons with radiology and genomics suggest that modular, interoperable pipelines are more sustainable than monolithic systems, particularly in hospital environments where procurement and governance procedures are complex.

7. Policy Implications and Future Directions

The policy implications of multi-scale state space modeling in digital pathology are substantial. Regulatory bodies increasingly view clinical artificial intelligence as software as a medical device, and they expect manufacturers to provide evidence of safety, effectiveness, and ongoing monitoring. Multi-scale architectures complicate this process because their behavior emerges from interactions across resolution levels and state space layers. It is not always straightforward to identify which scale contributed most to a decision, and the recurrent state may encode information that is difficult to visualize. Policy frameworks should therefore encourage the development of architectural documentation, validation protocols, and postmarket surveillance plans that are specific to multi-scale models. Auditors should be

able to inspect how image pyramids are constructed, how tokens are ordered, and how state space branches are fused.

Fairness policy must also evolve. Clinical artificial intelligence systems should be evaluated not only on aggregate performance but on performance across relevant patient subgroups, tissue types, and acquisition sites. For pathology models, this may require collaboration across institutions to assemble diverse slide collections. Data sharing is difficult because of privacy regulations, but federated learning and other privacy-preserving approaches can support model development without centralizing patient data. Multi-scale state space models can be adapted to federated settings, although their recurrent structure may require careful aggregation of state parameters. Policymakers should consider incentives for institutions to participate in federated evaluation networks, because single-institution validation is often insufficient to establish generalizability.

Future research should examine adaptive scale selection, in which the model learns to allocate computational resources to diagnostically ambiguous regions rather than processing every patch at every magnification uniformly. Such adaptive approaches could improve both efficiency and interpretability by producing explicit records of where the model looked more closely. Another promising direction is the integration of state space models with foundation representations trained on broad histopathology corpora. If foundation encoders can produce compact token sequences that already encode rich tissue semantics, the downstream state space aggregator may become simpler and more robust. This would shift the burden of representation learning to the pretraining stage while preserving the efficiency of state space aggregation. However, the governance of foundation models remains unsettled, and their use in clinical decision support must be accompanied by clear documentation of training data and failure modes.

The field also needs studies that compare multi-scale state space models with attention-based and convolutional baselines under realistic deployment constraints. Such studies should report not only classification metrics but also memory consumption, inference latency, energy use, and model update cost. These systems-level metrics are essential for hospital procurement decisions and for regulatory assessment. In addition, human factors research should examine how pathologists interact with multi-scale model explanations. A model that highlights regions at multiple resolutions may be more useful than a single-resolution saliency map, but it may also increase cognitive burden if explanations are not presented coherently. The design of user interfaces for multi-scale models is therefore not a peripheral concern; it is central to the safe and effective use of the technology.

8. Conclusion

Multi-scale state space modeling offers a promising architectural direction for microscopic image classification in digital pathology. It addresses the fundamental need to combine long-range tissue context with high-resolution cellular detail while avoiding some of the memory and throughput limitations of attention-based models. However, the value of this approach depends on more than model architecture. The design of image pyramids, token ordering, cross-scale fusion, and state space processing must be aligned with clinical workflows, computational infrastructure, data governance, and regulatory requirements. Structural trade-offs involving sequence length, state capacity, local versus global processing, and deployment location must be evaluated explicitly rather than treated as implementation details. At the same time, fairness, interpretability, and sustainability must be considered from the earliest stages of system design. A multi-scale model that is accurate but opaque, energy-intensive, or

poorly integrated into the laboratory will not deliver meaningful benefit in practice. Future work should therefore treat the model and its surrounding infrastructure as a single socio-technical system, developing not only better state space architectures but also better governance, validation, and monitoring frameworks for their responsible deployment.

References

1. Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.
2. Gu, A., Goel, K., & Ré, C. (2022). Efficiently modeling long sequences with structured state spaces. In International Conference on Learning Representations.
3. 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. In International Conference on Learning Representations.
4. 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.
5. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).
6. Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6), 555-570.
7. Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8), 1301-1309.
8. Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L. L., Wang, J. J., Vaidya, A., Le, L. P., Gerber, G. K., Sahai, S., Williams, W., & Mahmood, F. (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3), 850-862.
9. 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.
10. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., & Liu, Y. (2024). VMamba: Visual state space model. arXiv preprint arXiv:2401.10166.
11. van der Laak, J., Litjens, G., & Ciompi, F. (2021). Deep learning in histopathology: The path to the clinic. *Nature Medicine*, 27(5), 775-784.
12. 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.

13. Bankhead, P., Loughrey, M. B., Fernandez, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., McQuaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., James, J. A., Salto-Tellez, M., & Hamilton, P. W. (2017). QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1), 16878.
14. Shmatko, A., Ghaffari Laleh, N., Gerstung, M., & Kather, J. N. (2022). Artificial intelligence in histopathology: Enhancing cancer research and clinical oncology. *Nature Cancer*, 3(9), 1026-1038.
15. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V., & Madabhushi, A. (2019). Artificial intelligence in digital pathology: New tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*, 16(11), 703-715.
16. Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., Hermesen, M., Manson, Q. F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M. C., Bult, P., Beca, F., Beck, A. H., Wang, D., Khosla, A., Gargeya, R., ... Venancio, R. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22), 2199-2210.
17. 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. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618-626).
18. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.
19. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1-35.
20. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31-38.
21. 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012-10022).
22. 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.