

Uncertainty-Aware Diffusion Models for Multimodal Medical Image Segmentation and Cross-Modal Feature Alignment

Eimer MacBride

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.

eimermacbride946@uab.edu

Abstract

Multimodal medical image segmentation has become a central computational task in clinical workflows, where magnetic resonance, computed tomography, positron emission tomography, ultrasound, and pathology-derived representations often must be fused to support diagnosis, treatment planning, and longitudinal monitoring. Although deep discriminative models have achieved strong average performance, they frequently struggle when modalities are missing, noisy, misaligned, or distributed unevenly across institutions. Diffusion models offer a generative alternative because they learn iterative denoising processes that can represent complex anatomical distributions and expose uncertainty through sampling behavior. This paper examines uncertainty-aware diffusion models for multimodal medical image segmentation and cross-modal feature alignment from a system-level perspective. It considers architecture design, latent representation learning, modality fusion, uncertainty calibration, robustness under distribution shift, deployment infrastructure, sustainability, fairness, and governance. The analysis argues that segmentation quality cannot be separated from the socio-technical conditions of data collection, model maintenance, clinical oversight, and regulatory accountability. Cross-modal alignment is treated not merely as an optimization objective but as an infrastructural commitment involving shared ontologies, metadata standards, quality control, and long-term model stewardship. Trade-offs appear between expressive generative modeling and clinical latency, between uncertainty fidelity and computational burden, between centralized fusion and privacy-preserving federation, and between broad generalization and local adaptation. The paper concludes that uncertainty-aware diffusion systems should be evaluated as governed clinical infrastructures rather than isolated segmentation algorithms. Such evaluation requires multidimensional metrics, transparent documentation, human-in-the-loop escalation pathways, environmental accounting, and policy mechanisms that distribute responsibility across developers, providers, regulators, and institutions.

Keywords

uncertainty-aware diffusion models; multimodal medical image segmentation; cross-modal feature alignment; clinical deployment; robustness; fairness; governance.

1. Introduction

Medical image segmentation is not simply a pattern recognition problem. It is a clinical infrastructure problem in which anatomical boundaries, acquisition protocols, scanner variability, patient motion, contrast agents, annotation conventions, and institutional workflows all interact. In oncology, cardiology, neurology, and surgical planning,

segmentation masks influence radiation dosage, resection margins, catheter navigation, disease staging, and treatment response assessment. When segmentation models fail silently, the consequences can propagate through downstream decisions long after the initial inference. This systemic risk explains why uncertainty estimation has moved from a specialist concern to a core requirement for medical artificial intelligence. A segmentation model that reports a confident but incorrect boundary can be more dangerous than one that abstains or requests human review. Diffusion models have recently entered this space because they can model complex conditional distributions over anatomical structures and because their iterative generation process offers multiple pathways for uncertainty expression. Early diffusion formulations established the foundations of denoising probabilistic generation [1], while score-based formulations connected diffusion to stochastic differential equations and enabled flexible sampling [2]. Subsequent work demonstrated that diffusion models can outperform generative adversarial networks on image synthesis [3], and latent diffusion reduced computational cost by operating in compressed perceptual spaces [4]. For medical segmentation, diffusion-driven semantic segmentation with multimodal image fusion has been proposed as a way to enhance perception by combining generative refinement with cross-modal information [5]. These developments suggest a promising direction, but they also introduce new system-level tensions. Generative segmentation may improve boundary fidelity and uncertainty representation, yet it can increase inference latency, complicate validation, and produce plausible but anatomically invalid outputs. Cross-modal alignment may improve robustness when one modality is degraded, yet it requires shared representations that are difficult to maintain across scanners, hospitals, and patient populations. Uncertainty awareness may support deferral and triage, yet it demands calibration methods that remain meaningful under domain shift and class imbalance.

This paper adopts a systems-oriented view of uncertainty-aware diffusion models for multimodal medical image segmentation and cross-modal feature alignment. The central argument is that the technical promise of diffusion segmentation depends on architectural choices, data governance, deployment environments, sustainability constraints, fairness protections, and policy frameworks that are often treated as external to model performance. A diffusion model that performs well in a curated benchmark may fail in a hospital where modalities arrive asynchronously, metadata are incomplete, and scanner protocols change every few months. A cross-modal alignment module that learns a shared latent space may reduce missing-modality errors but can also encode site-specific correlations that do not generalize. An uncertainty estimator that produces useful confidence scores on internal validation data may become overconfident when deployed to a new population. Therefore, the design of such systems should be understood as a socio-technical architecture in which model components, human workflows, computational infrastructure, and regulatory obligations are interdependent.

The paper proceeds through several analytical layers. It first reviews the foundations of diffusion models and their relevance to medical segmentation. It then examines cross-modal feature alignment and architectural trade-offs. Next, it analyzes uncertainty quantification, robustness, and missing-modality behavior. It subsequently considers deployment, governance, sustainability, fairness, and policy implications. Finally, it outlines evaluation strategies and forward-looking perspectives for responsible clinical integration.

2. Foundations of Diffusion Models in Medical Segmentation

Diffusion models learn to reverse a gradual noise process, transforming random noise into structured data through iterative denoising. This generative formulation differs fundamentally from discriminative segmentation models that directly map an input image to a mask. The denoising process can be conditioned on imaging modalities, anatomical priors, or textual prompts, and it can produce multiple plausible outputs rather than a single deterministic prediction. The original denoising diffusion probabilistic framework demonstrated that high-quality samples could be generated through learned reverse transitions [1]. Score-based generative modeling generalized this idea by connecting diffusion to continuous stochastic processes, enabling more flexible noise schedules and samplers [2]. These methodological advances are relevant to medical segmentation because anatomical structures are not arbitrary images. They possess strong shape priors, topological constraints, and inter-patient variability that can be modeled probabilistically. A diffusion model can, in principle, refine coarse segmentation proposals by denoising boundary regions while respecting global anatomical plausibility. However, the same generative flexibility that improves boundary realism can also introduce hallucinated structures, especially when the conditioning signal is weak or ambiguous.

The computational cost of diffusion models has been a major barrier to clinical adoption. Iterative sampling requires multiple network evaluations per image, which can be prohibitive for three-dimensional volumes, real-time surgical guidance, or emergency radiology. Latent diffusion models address this by performing the generative process in a lower-dimensional latent space learned from the imaging data [4]. This reduces memory and compute requirements while preserving perceptual quality. In medical imaging, latent diffusion is attractive because volumetric data are large and because many anatomical structures can be represented compactly. Nevertheless, latent compression can discard fine-grained details that matter for small lesions, thin vessels, or subtle tumor margins. The trade-off between computational efficiency and clinical fidelity is therefore not incidental. It determines whether a model can be deployed at the point of care, whether it can run on edge devices, and whether it can be retrained frequently enough to remain current. Diffusion models that outperform generative adversarial networks in image synthesis [3] may still be unsuitable for time-critical tasks unless their sampling process is distilled, cached, or otherwise optimized.

Multimodal fusion adds another layer of complexity. Clinical imaging often combines complementary modalities: magnetic resonance provides soft tissue contrast, computed tomography provides dense anatomical detail, positron emission tomography provides metabolic activity, and ultrasound provides real-time acquisition. A segmentation system that fuses these sources can be more robust than a single-modality model, but fusion requires alignment at spatial, semantic, and temporal levels. Diffusion-driven semantic segmentation with multimodal image fusion has been proposed as a way to enhance perception by integrating generative refinement with cross-modal information [5]. The approach illustrates both the promise and the difficulty of the paradigm. On one hand, diffusion can model uncertainty over fused representations and refine ambiguous boundaries. On the other hand, the fusion process must handle missing modalities, imperfect registration, inconsistent resolution, and modality-specific artifacts. A system that assumes all modalities are available and perfectly aligned is unlikely to survive real clinical deployment. Thus, the foundation of diffusion segmentation should not be understood only as a generative modeling choice. It is also a decision about how multimodal evidence is assembled, validated, and maintained across heterogeneous infrastructures.

3. Architectural and Cross-Modal Alignment Considerations

The architectural lineage of medical segmentation provides important context for diffusion-based systems. Convolutional encoder-decoder networks such as U-Net established the skip-connection pattern that preserves spatial detail while capturing contextual features [6]. nnU-Net demonstrated that careful preprocessing, configuration, and training heuristics could achieve strong performance across many biomedical segmentation tasks without task-specific architectural novelty [7]. Transformer-based models then introduced global attention mechanisms. TransUNet combined transformers with convolutional decoders for medical image segmentation [8], while UNETR adapted transformers to three-dimensional medical volumes [9]. nnFormer explored interleaved transformer designs for volumetric segmentation [10], and Swin-Unet applied hierarchical transformers in a U-Net-like structure [11]. Medical Transformer introduced gated axial attention to reduce computational burden while retaining long-range dependencies [12]. These architectures matter for uncertainty-aware diffusion because they can serve as denoising backbones, condition encoders, or refinement modules. A diffusion model may use a U-Net-like denoiser, a transformer-based context encoder, or a hybrid design that combines convolutional locality with global attention. The choice affects memory consumption, receptive field, training stability, and the types of anatomical relationships the model can represent.

Cross-modal feature alignment is the process of mapping representations from different imaging modalities into a shared space where segmentation can occur. This can be achieved through early fusion, where modalities are concatenated at the input; intermediate fusion, where features are combined in hidden layers; or late fusion, where predictions are ensembled. Each strategy has distinct trade-offs. Early fusion preserves low-level complementary information but is sensitive to misregistration and missing channels. Intermediate fusion can learn modality-specific and shared features but requires careful regularization to prevent one modality from dominating. Late fusion is robust to missing modalities because each branch can operate independently, but it may fail to capture cross-modal interactions that improve boundary delineation. Diffusion models complicate these choices because the denoising process is iterative. A model may condition each denoising step on aligned multimodal features, or it may generate modality-specific intermediate representations and align them progressively. Progressive alignment can improve robustness but increases architectural complexity and the number of parameters that must be validated.

From a system perspective, cross-modal alignment is not only a neural network design problem. It depends on shared metadata standards, acquisition harmonization, and quality control. If one hospital uses a different slice thickness, contrast protocol, or reconstruction kernel, the learned alignment may degrade. If modalities arrive at different times, the system must decide whether to wait for all inputs, proceed with available data, or generate missing features through imputation. Diffusion models can be used for modality imputation, but synthetic modalities can introduce hallucinated evidence into clinical decision-making. This creates a governance challenge: who is accountable when a generated modality influences a segmentation that informs treatment? The answer cannot be reduced to model accuracy. It requires documentation of input provenance, uncertainty propagation, and human review. Therefore, cross-modal alignment should be designed as part of a traceable clinical pipeline rather than as an isolated representation learning module. The architecture must expose which modalities were used, how they were aligned, and how confident the system is in each fused representation.

4. Uncertainty Quantification, Robustness, and Missing Modalities

Uncertainty in medical segmentation arises from multiple sources. Aleatoric uncertainty reflects noise and ambiguity inherent in the data, such as partial volume effects, motion artifacts, or disagreement among expert annotators. Epistemic uncertainty reflects model ignorance, including limited training data, domain shift, and underrepresented anatomical variants. In clinical practice, these uncertainties interact. A model may be confident about a large tumor but uncertain about infiltrative margins. It may perform well on one scanner and poorly on another. It may be robust to missing modalities in one patient cohort but fail when a modality is missing for a different reason. Bayesian approximation methods such as dropout-based inference provide one way to estimate epistemic uncertainty by sampling from approximate posterior distributions [17]. Kendall and Gal distinguished aleatoric and epistemic uncertainty in computer vision and argued that both must be modeled for safe decision-making [18]. Deep ensembles offer another approach by training multiple models and measuring disagreement, often yielding strong calibration at higher computational cost [19]. For diffusion models, uncertainty can also be derived from multiple denoising trajectories, from variance across generated samples, or from learned noise schedules. However, these forms of uncertainty are not automatically clinically meaningful. A model may produce high sample variance for anatomically irrelevant textures while remaining confidently wrong about a critical boundary.

Robustness under missing modalities is particularly important because clinical data are rarely complete. A patient may have magnetic resonance but no computed tomography, or a research dataset may lack positron emission tomography for a subset of cases. Multimodal benchmarks such as BRATS have advanced brain tumor segmentation by providing standardized multimodal magnetic resonance data [14], and subsequent work expanded the available annotations and radiomic features [15]. Cardiac segmentation challenges have similarly driven progress in multi-structure segmentation and diagnosis [16]. Earlier multi-scale convolutional approaches demonstrated the value of contextual features and conditional random fields for brain lesion segmentation [13]. These efforts show that multimodal data improve average performance, but they also reveal that missing modalities can cause severe degradation if the model has learned to rely on a single dominant channel. A diffusion-based segmentation system can address this by learning a joint distribution over modalities and masks, allowing it to marginalize over missing inputs. Yet marginalization is computationally expensive and may produce plausible but unsupported predictions. The system must therefore distinguish between uncertainty caused by missing evidence and uncertainty caused by ambiguous anatomy.

A robust deployment architecture should include explicit fallback pathways. If a modality is missing, the system might switch to a modality-specific model, request additional imaging, flag the case for human review, or use a diffusion imputation module with clearly displayed uncertainty. The choice depends on clinical urgency, available infrastructure, and regulatory constraints. In emergency settings, waiting for a missing modality may be impossible, so the system must operate with degraded inputs and communicate that degradation. In research settings, strict inclusion criteria may be preferable to imputation. The key point is that uncertainty-aware diffusion models should not be evaluated only by their best-case segmentation accuracy. They should be evaluated by their behavior under missing, noisy, misaligned, and out-of-distribution inputs. This requires stress testing, prospective validation, and continuous monitoring rather than one-time benchmark reporting.

5. Deployment, Governance, Sustainability, and Fairness

Deploying uncertainty-aware diffusion models in clinical environments involves infrastructure choices that shape performance, equity, and sustainability. A centralized cloud deployment can provide substantial compute for iterative sampling and model updates, but it raises privacy, bandwidth, and jurisdictional concerns. An edge deployment can reduce latency and keep data local, but it may require model compression, quantization, or distillation that affects uncertainty calibration. A federated architecture can train across institutions without sharing raw data, but it introduces coordination costs, heterogeneous data distributions, and governance complexity. These are not merely engineering preferences. They determine who can access the model, which patients benefit, and which institutions bear the cost of maintenance. A hospital with advanced computational infrastructure may run a full diffusion ensemble, while a rural clinic may rely on a distilled model with limited uncertainty fidelity. If uncertainty estimates are used to decide when to defer to specialists, then inequitable access to high-quality uncertainty modeling can translate into inequitable care.

Fairness is a central concern because medical imaging datasets often underrepresent certain populations, rare diseases, and atypical presentations. An algorithm used to manage population health was shown to exhibit racial bias because it used health costs as a proxy for health needs [20]. This example demonstrates that technical performance metrics can mask structural inequities. In medical segmentation, bias may appear as lower accuracy for patients with different body habitus, different scanners, or different disease subtypes. Diffusion models can amplify bias if they learn to generate anatomically plausible outputs that reflect the majority distribution. Cross-modal alignment can also encode site-specific correlations that disadvantage patients from underrepresented sites. Fairness auditing should therefore examine segmentation accuracy, uncertainty calibration, and deferral rates across demographic and institutional subgroups. It should also consider intersectional effects, such as the combination of age, sex, comorbidity, and scanner type. Without such auditing, an uncertainty-aware system may appear safe on average while systematically failing vulnerable groups.

Governance and policy implications are equally significant. Implementing machine learning in health care raises ethical challenges related to transparency, consent, liability, and clinical integration [21]. High-performance medicine promises improved efficiency and precision, but it also depends on data infrastructure, workforce training, and regulatory adaptation [22]. An uncertainty-aware diffusion model is not a static device. It may be updated as new data arrive, fine-tuned for local populations, or combined with other clinical decision support tools. Each update can alter behavior, uncertainty thresholds, and failure modes. Governance must therefore address version control, post-market surveillance, incident reporting, and clear accountability when a model contributes to harm. If a segmentation is generated by a diffusion process that produces multiple plausible masks, the clinical record should indicate how the final mask was selected, what uncertainty was present, and whether a human adjudicated the output. Policy frameworks should require documentation of training data provenance, modality coverage, calibration procedures, and known limitations. They should also define when a model must abstain or escalate rather than produce a prediction.

Sustainability is an often-overlooked dimension of clinical artificial intelligence. Iterative diffusion sampling and ensemble-based uncertainty estimation require substantial energy and compute. While individual inference may be modest compared with training, large-scale deployment across hospitals, repeated imaging, and continuous monitoring can accumulate

significant environmental costs. Sustainability planning should consider model efficiency, hardware lifecycle, data storage, and the carbon intensity of cloud regions. It should also weigh the environmental benefits of improved diagnosis against the resource intensity of generative modeling. A system that improves early detection but requires excessive computation may be difficult to justify in resource-constrained settings. Therefore, uncertainty-aware diffusion models should be designed with efficiency budgets, not only accuracy targets. This may involve using latent diffusion, adaptive sampling, early exit strategies, or selective uncertainty estimation for high-risk cases. The goal is not to minimize computation at the expense of safety, but to allocate computational resources where they most improve clinical decisions.

6. Evaluation, Policy, and Forward-Looking Perspectives

Evaluation of uncertainty-aware diffusion segmentation must move beyond Dice scores and pixel accuracy. A clinically meaningful evaluation should assess calibration, deferral behavior, robustness, fairness, and workflow impact. Calibration metrics can reveal whether predicted uncertainty matches observed error, but calibration must be measured across subgroups and acquisition conditions. Deferral behavior should be evaluated by asking whether the system routes uncertain cases to human review in a timely and equitable manner. Robustness testing should include missing modalities, noise, motion, scanner variation, and adversarial perturbations. Fairness testing should examine whether uncertainty is systematically higher or lower for particular populations. Workflow evaluation should consider how clinicians interpret diffusion-generated masks, how they respond to uncertainty visualizations, and whether the system increases or decreases cognitive burden. These dimensions cannot be assessed through a single benchmark. They require prospective studies, simulation environments, and continuous monitoring after deployment.

Policy development should support this multidimensional evaluation. Regulatory frameworks for medical artificial intelligence should not treat uncertainty as an optional feature. They should require evidence that uncertainty estimates are valid, actionable, and communicated in ways that clinicians can understand. Procurement policies should demand transparency about training data, model versioning, and performance across sites. Reimbursement models should account for the human labor required to review uncertain cases and maintain data pipelines. Liability frameworks should clarify responsibility when a diffusion model generates a plausible but incorrect mask that influences treatment. International coordination may be necessary because medical imaging data, software vendors, and cloud infrastructure cross borders. At the same time, local adaptation is essential because clinical practices, disease prevalence, and regulatory norms vary. A balance must be struck between harmonized standards and flexible implementation.

Forward-looking research should explore uncertainty-aware diffusion models that can adapt to missing modalities, align features across heterogeneous scanners, and explain their uncertainty in clinically meaningful terms. It should investigate how diffusion trajectories can be audited, how cross-modal alignment can be monitored for drift, and how generative outputs can be constrained by anatomical priors. It should also examine human factors, including how radiologists, oncologists, and surgeons interpret probabilistic masks. Technical advances alone will not resolve these challenges. The most promising systems will be those that integrate generative modeling with data governance, workflow design, and policy accountability. They will treat uncertainty as a communication tool rather than a numerical afterthought. They will recognize that cross-modal alignment is a long-term infrastructural

commitment, not a one-time training objective. They will measure success by whether they improve patient outcomes equitably and sustainably, not merely by whether they achieve state-of-the-art segmentation scores.

Conclusion

Uncertainty-aware diffusion models represent a significant opportunity for multimodal medical image segmentation and cross-modal feature alignment. Their generative nature allows them to model complex anatomical distributions, refine ambiguous boundaries, and express uncertainty through sampling. Their multimodal capacity allows them to integrate complementary imaging evidence and potentially compensate for missing or degraded inputs. However, this paper has argued that the value of these systems depends on system-level design choices that extend far beyond the model architecture. Cross-modal alignment requires shared standards, quality control, and traceability. Uncertainty estimation requires calibration, subgroup analysis, and clinical interpretability. Deployment requires attention to latency, privacy, compute equity, and environmental sustainability. Governance requires clear accountability, post-market surveillance, and policy mechanisms that protect patients while enabling innovation.

The central conclusion is that uncertainty-aware diffusion segmentation should be evaluated as a governed clinical infrastructure rather than a standalone algorithm. Such evaluation must consider robustness under missing modalities, fairness across populations, transparency of generative outputs, and the human workflows that surround inference. It must also recognize that trade-offs between accuracy, latency, uncertainty fidelity, and resource consumption are inevitable. The most responsible path forward is not to maximize one dimension in isolation but to design systems that make these trade-offs explicit, auditable, and adjustable in response to local clinical needs. By combining technical rigor with governance, fairness, and sustainability, uncertainty-aware diffusion models can move from promising research prototypes toward trustworthy components of multimodal medical imaging practice.

References

1. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.
2. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*.
3. Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34, 8780–8794.
4. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.
5. Wang Z, He D, Zhao L, et al. Difusionseg: Diffusion-driven semantic segmentation with multi-modal image fusion for enhanced perception[J]. *Knowledge-Based Systems*, 2025: 114481.
6. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*, 234–241.

7. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203–211.
8. 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*.
9. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., & Xu, D. (2022). UNETR: Transformers for 3D medical image segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 574–584.
10. Zhou, H.-Y., Guo, J., Zhang, Y., Yu, L., Wang, L., & Yu, Y. (2021). nnFormer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*.
11. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2023). Swin-Unet: Unet-like pure transformer for medical image segmentation. *Computer Vision – ECCV 2022 Workshops*, 205–218.
12. Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*, 36–46.
13. Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., & Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36, 61–78.
14. Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10), 1993–2024.
15. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4, 170117.
16. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., et al. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11), 2514–2525.
17. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, 1050–1059.
18. 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.
19. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 6402–6413.
20. 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.

21. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—Addressing ethical challenges. *New England Journal of Medicine*, 378(11), 981–983.
22. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56.